Chapter 3
Transport Layer

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

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

Thanks and enjoy! JFK/KWR

All material copyright 1996-2016
© J.F Kurose and K.W. Ross, All Rights Reserved
Chapter 3: Transport Layer

our goals:

- understand principles behind transport layer services:
  - multiplexing, demultiplexing
  - reliable data transfer
  - flow control
  - congestion control

- learn about Internet transport layer protocols:
  - UDP: connectionless transport
  - TCP: connection-oriented reliable transport
  - TCP congestion control
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer
3.5 connection-oriented transport: TCP
  • segment structure
  • reliable data transfer
  • flow control
  • connection management
3.6 principles of congestion control
3.7 TCP congestion control
Transport services and 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**: logical communication between hosts
- **transport layer**: logical communication between processes
  - relies on, enhances, network layer services

**household analogy:**

12 kids in Ann’s house sending letters to 12 kids in Bill’s house:
- hosts = houses
- processes = kids
- app messages = letters in envelopes
- transport protocol = Ann and Bill who demux to in-house siblings
- network-layer protocol = postal service
Internet transport-layer protocols

- reliable, in-order delivery (TCP)
  - congestion control
  - flow control
  - connection setup
- unreliable, unordered delivery: UDP
  - no-frills extension of “best-effort” IP
- services not available:
  - delay guarantees
  - bandwidth guarantees
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer
3.5 connection-oriented transport: TCP
  • segment structure
  • reliable data transfer
  • flow control
  • connection management
3.6 principles of congestion control
3.7 TCP congestion control
**Multiplexing/demultiplexing**

**Multiplexing at sender:**
handle data from multiple sockets, add transport header (later used for demultiplexing)

**Demultiplexing at receiver:**
use header info to deliver received segments to correct socket
How demultiplexing works

- host receives IP datagrams
  - each datagram has source IP address, destination IP address
  - each datagram carries one transport-layer segment
  - each segment has source, destination port number

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

TCP/UDP segment format:
- source port #
- dest port #
- other header fields
- application data (payload)
- 32 bits

Transport Layer 3-9
Connectionless demultiplexing

- **recall:** created socket has host-local port #:
  ```java
  DatagramSocket mySocket1 = new DatagramSocket(12534);
  ```

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

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

IP datagrams with *same dest. port #*, but different source IP addresses and/or source port numbers will be directed to *same socket* at dest.
Connectionless demux: example

DatagramSocket
mySocket2 = new
DatagramSocket
(9157);

DatagramSocket
mySocket1 = new
DatagramSocket
(6428);

source port: 6428
dest port: 9157

source port: 9157
dest port: 6428

source port: ?
dest port: ?

source port: ?
dest port: ?
Connection-oriented demux

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

- demux: receiver 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: example

three segments, all destined to IP address: B, dest port: 80 are demultiplexed to different sockets
Connection-oriented demux: example

Transport Layer 3-14
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer
3.5 connection-oriented transport: TCP
   • segment structure
   • reliable data transfer
   • flow control
   • connection management
3.6 principles of congestion control
3.7 TCP congestion control
UDP: User Datagram Protocol [RFC 768]

- “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

- UDP use:
  - streaming multimedia apps (loss tolerant, rate sensitive)
  - DNS
  - SNMP

- reliable transfer over UDP:
  - add reliability at application layer
  - application-specific error recovery!
UDP: segment header

<table>
<thead>
<tr>
<th>source port #</th>
<th>dest port #</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>checksum</td>
</tr>
</tbody>
</table>

length, in bytes of UDP segment, including header

- 32 bits

application data (payload)

UDP segment format

why is there a UDP?

- no connection establishment (which can add delay)
- simple: no connection state at sender, receiver
- small header size
- no congestion control: UDP can blast away as fast as desired
UDP checksum

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

**sender:**
- treat segment contents, including header fields, as sequence of 16-bit integers
- checksum: addition (one’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.

*But maybe errors nonetheless? More later...*
Internet checksum: example

example: add two 16-bit integers

```
1 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0
1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
```

wraparound

```
1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1
```

sum

```
1 0 1 1 1 0 1 1 1 0 1 1 1 1 1 0 0
```

checksum

```
0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 1
```

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

* Check out the online interactive exercises for more examples: http://gaia.cs.umass.edu/kurose_ross/interactive/
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer
3.5 connection-oriented transport: TCP
   - segment structure
   - reliable data transfer
   - flow control
   - connection management
3.6 principles of congestion control
3.7 TCP congestion control
Principles of reliable data transfer

- important in application, transport, link layers
  - top-10 list of important networking topics!

- characteristics of unreliable channel will determine complexity of reliable data transfer protocol (rdt)
Principles of reliable data transfer

- important in application, transport, link layers
  - top-10 list of important networking topics!

- characteristics of unreliable channel will determine complexity of reliable data transfer protocol (rdt)
Principles of reliable data transfer

- important in application, transport, link layers
  - top-10 list of important networking topics!

- characteristics of unreliable channel will determine complexity of reliable data transfer protocol (rdt)
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
Reliable data transfer: getting started

we’ll:
- incrementally develop sender, receiver sides of **reliable data transfer protocol (rdt)**
- consider only unidirectional data transfer
  - but control info will flow on both directions!
- use finite state machines (FSM) to specify sender, receiver

**state:** when in this “state” next state uniquely determined by next event.
**rdt1.0: 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 reads data from underlying channel

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

Wait for call from below
rdt_rcv(packet)
  extract (packet, data)
  deliver_data(data)
```

**sender**

**receiver**
underlying channel may flip bits in packet
  • checksum to detect bit errors

the question: how to recover from errors:

How do humans recover from “errors” during conversation?
rdt2.0: channel with bit errors

- underlying channel may flip bits in packet
  - checksum to detect bit errors

- *the question*: how to recover from errors:
  - *acknowledgements (ACKs)*: receiver explicitly tells sender that pkt received OK
  - *negative acknowledgements (NAKs)*: receiver explicitly tells sender that pkt had errors
  - sender retransmits pkt on receipt of NAK

- new mechanisms in *rdt2.0* (beyond *rdt1.0*):
  - error detection
  - feedback: control msgs (ACK, NAK) from receiver to sender
**rdt2.0: FSM specification**

sender

- `rdt_send(data)`
- `sndpkt = make_pkt(data, checksum)`
- `udt_send(sndpkt)`

Wait for call from above

- `rdt_rcv(rcvpkt) && isNAK(rcvpkt)`
- `udt_send(sndpkt)`

Receiver

- `rdt_rcv(rcvpkt) && isACK(rcvpkt)`

- `wait for ACK or NAK`

- `wait for call from below`

- `rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)`
- `extract(rcvpkt, data)`
- `deliver_data(data)`
- `udt_send(ACK)`

Sender

- `rdt_send(data)`
- `sndpkt = make_pkt(data, checksum)`
- `udt_send(sndpkt)`

Wait for call from above

- `rdt_rcv(rcvpkt) && isNAK(rcvpkt)`
- `udt_send(sndpkt)`

Wait for ACK or NAK

- `rdt_rcv(rcvpkt) && isACK(rcvpkt)`

- `wait for call from above`
rdt2.0: operation with no errors

- rdt_send(data)
- snkpkt = make_pkt(data, checksum)
- udt_send(sndpkt)
- Wait for call from above

- rdt_rcv(rcvpkt) && isNAK(rcvpkt)
- udt_send(sndpkt)
- Wait for ACK or NAK

- rdt_send(NAK)
- rdt_rcv(rcvpkt) && isACK(rcvpkt)
- Wait for call from above

- rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
- extract(rcvpkt, data)
- deliver_data(data)
- udt_send(ACK)

- Wait for ACK or NAK
- Wait for call from below
rdt2.0: error scenario

- `rdt_send(data)`
- `snkpkt = make_pkt(data, checksum)`
- `udt_send(sndpkt)`
- `rdt_rcv(rcvpkt) && isNAK(rcvpkt)`
- `udt_send(sndpkt)`
- `rdt_rcv(rcvpkt) && isACK(rcvpkt)`
- `udt_send(ACK)`
- `rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)`
- `extract(rcvpkt, data)`
- `deliver_data(data)`
- `udt_send(NAK)`
- `Wait for call from above`
- `Wait for ACK or NAK`
- `Wait for call from below`
rdt2.0 has a fatal flaw!

what happens if ACK/NAK corrupted?

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

handling duplicates:

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

stop and wait
sender sends one packet, then waits for receiver response
rdt2.1: sender, handles garbled ACK/NAKs

rdt_send(data)

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

wait for call 0 from above

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt) && isACK(rcvpkt)

Lambda

wait for ACK or NAK 0

rdt_send(sndpkt)

wait for call 1 from above

Lambda

rdt_rcv(rcvpkt) && notcorrupt(rcvpkt) && isACK(rcvpkt)

rdt_send(data)

sndpkt = make_pkt(1, data, checksum)
udt_send(sndpkt)
rdt2.1: receiver, handles garbled ACK/NAKs

```c
// Wait for 0 from below
rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
&& has_seq0(rcvpkt)
extract(rcvpkt, data)
deliver_data(data)
sndpkt = make_pkt(ACK, chksum)
udt_send(sndpkt)

// Wait for 1 from below
rdt_rcv(rcvpkt) && notcorrupt(rcvpkt)
&& has_seq1(rcvpkt)
extract(rcvpkt, data)
deliver_data(data)
sndpkt = make_pkt(ACK, chksum)
udt_send(sndpkt)
```

```c
// rdt2.1: receiver, handles garbled ACK/NAKs
```

rdt2.1: discussion

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

receiver:
- must check if received packet is duplicate
  - state indicates whether 0 or 1 is expected pkt seq #
- note: receiver can not know if its last ACK/NAK received OK at sender
rdt2.2: a NAK-free protocol

- same functionality as rdt2.1, using ACKs only
- instead of NAK, receiver sends ACK for last pkt received OK
  - receiver must explicitly include seq # of pkt being ACKed
- duplicate ACK at sender results in same action as NAK: retransmit current pkt
**rdt2.2: sender, receiver fragments**

- **Sender FSM**
  - \( rdt\_send(data) \)
  - \( sndpkt = \text{make}_\text{pkt}(0, \text{data}, \text{checksum}) \)
  - \( udt\_send(sndpkt) \)
  - \( \text{Wait for call 0 from above} \)
  - \( \text{Wait for ACK 0} \)
  - \( rdt\_rcv(rcvpkt) \) \&\& \((\text{corrupt(rcvpkt)} \mid\mid \text{isACK(rcvpkt,1)}) \)
  - \( udt\_send(sndpkt) \)

- **Receiver FSM**
  - \( rdt\_rcv(rcvpkt) \) \&\& \( \text{notcorrupt(rcvpkt)} \) \&\& \( \text{isACK(rcvpkt,0}) \)
  - \( \text{Wait for 0 from below} \)
  - \( \text{extract(rcvpkt, data)} \)
  - \( \text{deliver data(data)} \)
  - \( \text{sndpkt} = \text{make}_\text{pkt}(\text{ACK1, checksum}) \)
  - \( udt\_send(sndpkt) \)
rdt3.0: channels with errors and loss

new assumption: underlying channel can also lose packets (data, ACKs)
- checksum, seq. #, ACKs, retransmissions will be of help … but not enough

approach: sender waits “reasonable” amount of time for ACK
- retransmits if no ACK received in this time
- if pkt (or ACK) just delayed (not lost):
  - retransmission will be duplicate, but seq. #’s already handles this
  - receiver must specify seq # of pkt being ACKed
- requires countdown timer
rdt3.0 sender

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

wait for call 0 from above

rdt_rcv(rcvpkt)
\[\Lambda\]

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

wait for call 1 from above

rdt_send(data)

\[\Lambda\]

udt_send(sndpkt)
start_timer

timeout

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

wait for call 1 from above

rdt_rcv(rcvpkt)
\[\Lambda\]

udt_send(sndpkt)
start_timer

timeout

wait for call 0 from above

rdt_rcv(rcvpkt)
&& (corrupt(rcvpkt) || isACK(rcvpkt,1))
```
rdt3.0 in action

(a) no loss

(b) packet loss
**rdt3.0 in action**

**sender**
- Send pkt0
- Rcv pkt0
- Rcv ack0
- Rcv pkt1
- Rcv ack1
- Rcv ack1 (detect duplicate)
- Send pkt0
- Send ack0

**receiver**
- Send pkt0
- Rcv pkt0
- Rcv ack0
- Rcv pkt1
- Rcv ack1
- Send pkt0
- Rcv ack1
- Rcv ack1 (detect duplicate)

(c) ACK loss

(d) premature timeout/ delayed ACK
Performance of rdt3.0

- rdt3.0 is correct, but performance stinks
- e.g.: 1 Gbps link, 15 ms prop. delay, 8000 bit packet:

\[ D_{\text{trans}} = \frac{L}{R} = \frac{8000 \text{ bits}}{10^9 \text{ bits/sec}} = 8 \text{ microseconds} \]

- \( U_{\text{sender}} \): utilization – fraction of time sender busy sending

\[ U_{\text{sender}} = \frac{L / R}{RTT + L / R} = \frac{.008}{30.008} = 0.00027 \]

- if RTT=30 msec, 1KB pkt every 30 msec: 33kB/sec throughput over 1 Gbps link

- network protocol limits use of physical resources!
**rdt3.0: stop-and-wait operation**

- First packet bit transmitted, $t = 0$
- Last packet bit transmitted, $t = L/R$
- First packet bit arrives
- Last packet bit arrives, send ACK
- ACK arrives, send next packet, $t = RTT + L/R$

\[
U_{sender} = \frac{L/R}{RTT + L/R} = \frac{.008}{30.008} = 0.00027
\]
Pipelined protocols

**pipelining**: sender allows multiple, “in-flight”, yet-to-be-acknowledged pkts

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

- two generic forms of pipelined protocols: *go-Back-N*, *selective repeat*
Pipelining: increased utilization

first packet bit transmitted, \( t = 0 \)
last bit transmitted, \( t = \frac{L}{R} \)

first packet bit arrives
last packet bit arrives, send ACK
last bit of 2\(^{nd}\) packet arrives, send ACK
last bit of 3\(^{rd}\) packet arrives, send ACK

ACK arrives, send next packet, \( t = RTT + \frac{L}{R} \)

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

3-packet pipelining increases utilization by a factor of 3!
Pipelined protocols: overview

**Go-back-N:**
- sender can have up to N unacked packets in pipeline
- receiver only sends *cumulative ack*
  - doesn’t ack packet if there’s a gap
- sender has timer for oldest unacked packet
  - when timer expires, retransmit *all* unacked packets

**Selective Repeat:**
- sender can have up to N unack’ed packets in pipeline
- rcvr sends *individual ack* for each packet
- sender maintains timer for each unacked packet
  - when timer expires, retransmit only that unacked packet
**Go-Back-N: sender**

- k-bit seq # in pkt header
- “window” of up to N, consecutive unack’ed pkts allowed

![Diagram of Go-Back-N: sender]

- ACK(n): ACKs all pkts up to, including seq # n - “cumulative ACK”
  - may receive duplicate ACKs (see receiver)
- timer for oldest in-flight pkt
- timeout(n): retransmit packet n and all higher seq # pkts in window
GBN: sender extended FSM

\[ rdt\_send(data) \]

\[ \text{if } (\text{nextseqnum} < \text{base} + \text{N}) \{ \]
\[ \quad \text{sndpkt}[\text{nextseqnum}] = \text{make_pkt}(<\text{nextseqnum}, \text{data}, \text{chksum}) \]
\[ \quad \text{udt\_send(sndpkt[\text{nextseqnum}])} \]
\[ \text{if } (\text{base} == \text{nextseqnum}) \]
\[ \quad \text{start\_timer} \]
\[ \quad \text{nextseqnum}++ \]
\[ \} \]
\[ \text{else} \]
\[ \quad \text{refuse\_data(data)} \]

base = 1

nextseqnum = 1

\[ \text{rdt\_rcv(rcvpkt)} \]
\[ \text{&& corrupt(rcvpkt)} \]

\[ \text{timeout} \]
\[ \text{start\_timer} \]
\[ \text{udt\_send(sndpkt[base])} \]
\[ \text{udt\_send(sndpkt[base+1])} \]
\[ \ldots \]
\[ \text{udt\_send(sndpkt[nextseqnum-1])} \]

\[ \text{rdt\_rcv(rcvpkt)} \]
\[ \text{&& notcorrupt(rcvpkt)} \]

base = getacknum(rcvpkt) + 1

If (base == nextseqnum)
\[ \text{stop\_timer} \]
else
\[ \text{start\_timer} \]
ACK-only: always send ACK for correctly-received pkt with highest \textit{in-order} seq #

- may generate duplicate ACKs
- need only remember \texttt{expectedseqnum}

\begin{itemize}
  \item out-of-order pkt:
    \begin{itemize}
      \item discard (don’t buffer): \textit{no receiver buffering}!
      \item re-ACK pkt with highest in-order seq #
    \end{itemize}
\end{itemize}
GBN in action

**sender window (N=4)**

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

**sender**
- send pkt0
- send pkt1
- send pkt2
- send pkt3 (wait)

**receiver**
- receive pkt0, send ack0
- receive pkt1, send ack1
- receive pkt3, discard, (re)send ack1
- receive pkt4, discard, (re)send ack1
- receive pkt5, discard, (re)send ack1

**pkt 2 timeout**
- send pkt2
- send pkt3
- send pkt4
- send pkt5

**rcv pkt2, deliver, send ack2**
- rcv pkt3, deliver, send ack3
- rcv pkt4, deliver, send ack4
- rcv pkt5, deliver, send ack5

**X loss**
- ignore duplicate ACK

Transport Layer 3-50
Selective repeat

- receiver *individually* acknowledges all correctly received pkts
  - buffers pkts, as needed, for eventual in-order delivery to upper layer
- sender only resends pkts for which ACK not received
  - sender timer for each unACKed pkt
- sender window
  - $N$ consecutive seq #’s
  - limits seq #s of sent, unACKed pkts
Selective repeat: sender, receiver windows

(a) sender view of sequence numbers

(b) receiver view of sequence numbers
Selective repeat

**sender**

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

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

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

**receiver**

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

- **pkt n in [rcvbase-N,rcvbase-1]:**
  - ACK(n)

- otherwise:
  - ignore
Selective repeat in action

sender window \((N=4)\)

send pkt0
send pkt1
send pkt2
send pkt3
(wait)

receiver

receive pkt0, send ack0
receive pkt1, send ack1
receive pkt3, buffer, send ack3
receive pkt4, buffer, send ack4
receive pkt5, buffer, send ack5

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

Q: what happens when ack2 arrives?
Selective repeat: dilemma

example:
- seq #'s: 0, 1, 2, 3
- window size = 3
- receiver sees no difference in two scenarios!
- duplicate data accepted as new in (b)

Q: what relationship between seq # size and window size to avoid problem in (b)?
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer
3.5 connection-oriented transport: TCP
  • segment structure
  • reliable data transfer
  • flow control
  • connection management
3.6 principles of congestion control
3.7 TCP congestion control
TCP: Overview  RFCs: 793,1122,1323, 2018, 2581

- **point-to-point:**
  - one sender, one receiver

- **reliable, in-order byte steam:**
  - no “message boundaries”

- **pipelined:**
  - TCP congestion and flow control set window size

- **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
TCP segment structure

- **Source port #**
- **Destination port #**
- **Sequence number**
- **Acknowledgement number**
- **Receive window**
- **Urg data pointer**
- **Options (variable length)**
- **Application data** (variable length)

**Fields and Descriptions:**

- **URG**: URGent data (generally not used)
- **ACK**: ACK No. valid
- **PSH**: push data now (generally not used)
- **RST, SYN, FIN**: connection estab (setup, teardown commands)
- **Internet checksum** (as in UDP)

**Checksum:**

- Counting by bytes of data (not segments!)
- # bytes rcvr willing to accept
TCP seq. numbers, ACKs

**sequence numbers:**
- byte stream “number” of first byte in segment’s data

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

**Q:** how receiver handles out-of-order segments
- **A:** TCP spec doesn’t say, up to implementor
TCP seq. numbers, ACKs

Host A

User types 'C'

host ACKs receipt of echoed 'C'

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

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

Seq=43, ACK=80

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

Host B

simple telnet scenario
**Q:** how to set TCP timeout value?

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

**Q:** how to estimate RTT?

- **SampleRTT:** measured time from segment transmission until ACK receipt
  - ignore retransmissions
- **SampleRTT** will vary, want estimated RTT “smoother”
  - average several recent measurements, not just current **SampleRTT**
**TCP round trip time, timeout**

\[
\text{EstimatedRTT} = (1 - \alpha) \times \text{EstimatedRTT} + \alpha \times \text{SampleRTT}
\]

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

![Graph showing RTT: gaia.cs.umass.edu to fantasia.eurecom.fr](image)
TCP round trip time, timeout

- **timeout interval:** EstimatedRTT plus “safety margin”
  - large variation in EstimatedRTT → larger safety margin
- estimate SampleRTT deviation from EstimatedRTT:
  \[
  \text{DevRTT} = (1-\beta) \times \text{DevRTT} + \\
  \beta \times |\text{SampleRTT} - \text{EstimatedRTT}|
  \]
  (typically, $\beta = 0.25$)

```
TimeoutInterval = EstimatedRTT + 4 \times \text{DevRTT}
```

* Check out the online interactive exercises for more examples: http://gaia.cs.umass.edu/kuroseRoss/interactive/
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer

3.5 connection-oriented transport: TCP
  • segment structure
  • reliable data transfer
  • flow control
  • connection management

3.6 principles of congestion control
3.7 TCP congestion control
TCP reliable data transfer

- TCP creates rdt service on top of IP’s unreliable service
  - pipelined segments
  - cumulative acks
  - single retransmission timer
- retransmissions triggered by:
  - timeout events
  - duplicate acks

Let’s initially consider simplified TCP sender:
- ignore duplicate acks
- ignore flow control, congestion control
TCP sender events:

**data rcvd from app:**
- create segment with seq #
- seq # is byte-stream number of first data byte in segment
- start timer if not already running
  - think of timer as for oldest unacked segment
  - expiration interval: `TimeOutInterval`

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

**ack rcvd:**
- if ack acknowledges previously unacked segments
  - update what is known to be ACKed
  - start timer if there are still unacked segments
TCP sender (simplified)

wait for event

\[ \lambda \]

NextSeqNum = InitialSeqNum
SendBase = InitialSeqNum

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

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

ACK received, with ACK field value y

timeout
retransmit not-yet-acked segment with smallest seq. #
start timer
TCP: retransmission scenarios

lost ACK scenario

precipitate timeout

Transport Layer 3-68
TCP: retransmission scenarios

Host A

Host B

Seq=92, 8 bytes of data

Seq=100, 20 bytes of data

ACK=100

Cumulative ACK

Seq=120, 15 bytes of data

timeout

Seq=100, 20 bytes of data

ACK=120

Transport Layer 3-69
## TCP ACK generation

### event at receiver

<table>
<thead>
<tr>
<th>Event Description</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>
TCP 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.

TCP fast retransmit
if sender receives 3 ACKs for same data
(“triple duplicate ACKs”), resend unacked segment with smallest seq #
- likely that unacked segment lost, so don’t wait for timeout
TCP fast retransmit

Host A

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

Host B

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

timeout

fast retransmit after sender receipt of triple duplicate ACK

Transport Layer 3-72
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer

3.5 connection-oriented transport: TCP
  • segment structure
  • reliable data transfer
  • flow control
  • connection management

3.6 principles of congestion control
3.7 TCP congestion control
TCP flow control

Flow control: receiver controls sender, so sender won’t overflow receiver’s buffer by transmitting too much, too fast.

Application may remove data from TCP socket buffers.

... slower than TCP receiver is delivering (sender is sending).

TCP flow control

Receiver protocol stack
TCP flow control

- Receiver “advertises” free buffer space by including `rwnd` value in TCP header of receiver-to-sender segments
  - `RcvBuffer` size set via socket options (typical default is 4096 bytes)
  - Many operating systems autoadjust `RcvBuffer`

- Sender limits amount of unacked (“in-flight”) data to receiver’s `rwnd` value

- Guarantees receive buffer will not overflow
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer

3.5 connection-oriented transport: TCP
   • segment structure
   • reliable data transfer
   • flow control
   • connection management

3.6 principles of congestion control
3.7 TCP congestion control
Connection Management

before exchanging data, sender/receiver “handshake”:

- agree to establish connection (each knowing the other willing to establish connection)
- agree on connection parameters

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

Socket connectionSocket = welcomeSocket.accept();
```
**Agreeing to establish a connection**

2-way handshake:

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

- variable delays
- retransmitted messages (e.g. `req_conn(x)`) due to message loss
- message reordering
- can’t “see” other side

---

Transport Layer 3-78
Agreeing to establish a connection

2-way handshake failure scenarios:

1. Client chooses x, sends req_conn(x), server returns ESTAB, server forgets x.
2. Client terminates connection x completes.
3. Half open connection! (no client!)

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Client terminates.
3. Server forgets x.

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Server responds with acc_conn(x).
3. Client sends req_conn(x), server returns ESTAB.
4. Client sends data(x+1), server accepts data(x+1).
5. Client completes connection x completes.

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Server responds with acc_conn(x).
3. Server sends data(x+1), client accepts data(x+1).
4. Client completes connection x completes.

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Server responds with acc_conn(x).
3. Server sends data(x+1), client accepts data(x+1).
4. Client completes connection x completes.

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Server responds with acc_conn(x).
3. Server sends data(x+1), client accepts data(x+1).
4. Client completes connection x completes.

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Server responds with acc_conn(x).
3. Server sends data(x+1), client accepts data(x+1).
4. Client completes connection x completes.

1. Client chooses x, sends req_conn(x), server returns ESTAB.
2. Server responds with acc_conn(x).
3. Server sends data(x+1), client accepts data(x+1).
4. Client completes connection x completes.
TCP 3-way handshake

**client state**

- **LISTEN**
  - choose init seq num, x
  - send TCP SYN msg

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

- **ESTAB**

**server state**

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

- **SYN RCVD**
  - SYNbit=1, Seq=x

- **ESTAB**
  - SYNbit=1, Seq=y
  - ACKbit=1; ACKnum=x+1
  - received ACK(y) indicates client is live
TCP 3-way handshake: FSM

Socket connectionSocket = welcomeSocket.accept();

SYN(x)

SYNACK(seq=y,ACKnum=x+1)
create new socket for communication back to client

Socket clientSocket = newSocket("hostname","port number");
SYN(seq=x)

SYN rcvd

ACK(ACKnum=y+1)

SYN sent

SYNACK(seq=y,ACKnum=x+1)
ACK(ACKnum=y+1)

ESTAB

closed

listen

Λ

Λ
TCP: closing a connection

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

**client state**

- **ESTAB**
  - clientSocket.close()
- **FIN_WAIT_1**
  - can no longer send but can receive data
- **FIN_WAIT_2**
  - wait for server close
- **TIMED_WAIT**
  - timed wait for 2*max segment lifetime
- **CLOSED**

**server state**

- **ESTAB**
- **CLOSE_WAIT**
  - can still send data
- **LAST_ACK**
  - can no longer send data
- **CLOSED**

- **FINbit=1, seq=x**
  - FINbit=1, seq=y
  - ACKbit=1; ACKnum=x+1
  - ACKbit=1; ACKnum=y+1

Transport Layer 3-83
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer

3.5 connection-oriented transport: TCP
- segment structure
- reliable data transfer
- flow control
- connection management

3.6 principles of congestion control
3.7 TCP congestion control
Principles of congestion control

congestion:

- informally: “too many sources sending too much data too fast for network to handle”
- different from flow control!
- manifestations:
  - lost packets (buffer overflow at routers)
  - long delays (queueing in router buffers)
- a top-10 problem!
Causes/costs of congestion: scenario 1

- two senders, two receivers
- one router, infinite buffers
- output link capacity: R
- no retransmission

- maximum per-connection throughput: R/2
- large delays as arrival rate, $\lambda_{in}$, approaches capacity
Causes/costs of congestion: scenario 2

- one router, *finite* buffers
- sender retransmission of timed-out packet
  - application-layer input = application-layer output: $\lambda_{in} = \lambda_{out}$
  - transport-layer input includes *retransmissions*: $\lambda_{in}' \geq \lambda_{in}$

![Diagram showing network flow and buffer management](image)
Causes/costs of congestion: scenario 2

idealization: perfect knowledge

- sender sends only when router buffers available

\[ \lambda_{in} : \text{original data} \]
\[ \lambda'_{in} : \text{original data, plus retransmitted data} \]

\[ \lambda_{out} \]

\[ \text{free buffer space!} \]

finite shared output link buffers
Causes/costs of congestion: scenario 2

**Idealization:** *known loss*

packets can be lost, dropped at router due to full buffers

- sender only resends if packet *known* to be lost

![Diagram showing network flow and congestion]

- \( \lambda_{in} \): original data
- \( \lambda_{in}' \): original data, *plus* retransmitted data
- \( \lambda_{out} \)
- *no buffer space!*

Host B
Causes/costs of congestion: scenario 2

**Idealization:** *known loss*
- Packets can be lost, dropped at router due to full buffers
  - Sender only resends if packet *known* to be lost

![Diagram](attachment:image.png)

- $\lambda_{\text{out}}$: original data
- $\lambda'_{\text{in}}$: original data, *plus* retransmitted data
- Free buffer space!

When sending at $R/2$, some packets are retransmissions but asymptotic goodput is still $R/2$ (why?)
Causes/costs of congestion: scenario 2

**Realistic: duplicates**

- packets can be lost, dropped at router due to full buffers
- sender times out prematurely, sending *two* copies, both of which are delivered

When sending at R/2, some packets are retransmissions including duplicated that are delivered!
Causes/costs of congestion: scenario 2

**Realistic: duplicates**

- packets can be lost, dropped at router due to full buffers
- sender times out prematurely, sending *two* copies, both of which are delivered

"costs" of congestion:

- more work (retrans) for given "goodput"
- unneeded retransmissions: link carries multiple copies of pkt
  - decreasing goodput
Causes/costs of congestion: scenario 3

- four senders
- multihop paths
- timeout/retransmit

Q: what happens as $\lambda_{in}$ and $\lambda_{in}'$ increase?

A: as red $\lambda_{in}'$ increases, all arriving blue pkts at upper queue are dropped, blue throughput $\to 0$
another “cost” of congestion:

- when packet dropped, any “upstream transmission capacity used for that packet was wasted!
Chapter 3 outline

3.1 transport-layer services
3.2 multiplexing and demultiplexing
3.3 connectionless transport: UDP
3.4 principles of reliable data transfer

3.5 connection-oriented transport: TCP
  • segment structure
  • reliable data transfer
  • flow control
  • connection management
3.6 principles of congestion control
3.7 TCP congestion control
TCP congestion control: additive increase multiplicative decrease

- **approach:** sender increases transmission rate (window size), probing for usable bandwidth, until loss occurs
  - *additive increase:* increase $cwnd$ by 1 MSS every RTT until loss detected
  - *multiplicative decrease:* cut $cwnd$ in half after loss

AIMD saw tooth behavior: probing for bandwidth

Additively increase window size ...
.... until loss occurs (then cut window in half)
TCP Congestion Control: details

sender limits transmission:
- \( \text{LastByteSent - LastByteAcked} \leq \text{cwnd} \)

\[ \text{sender sequence number space} \quad \text{cwnd} \]

TCP sending rate:
- roughly: send \( \text{cwnd} \) bytes, wait RTT for ACKs, then send more bytes

\[ \text{rate} \approx \frac{\text{cwnd}}{\text{RTT}} \text{ bytes/sec} \]
TCP Slow Start

- when connection begins, increase rate exponentially until first loss event:
  - initially $cwnd = 1$ MSS
  - double $cwnd$ every RTT
  - done by incrementing $cwnd$ for every ACK received

- **summary:** initial rate is slow but ramps up exponentially fast
TCP: detecting, reacting to loss

- loss indicated by timeout:
  - `cwnd` set to 1 MSS;
  - window then grows exponentially (as in slow start) to threshold, then grows linearly

- loss indicated by 3 duplicate ACKs: TCP RENO
  - dup ACKs indicate network capable of delivering some segments
  - `cwnd` is cut in half window then grows linearly

- TCP Tahoe always sets `cwnd` to 1 (timeout or 3 duplicate acks)
Q: when should the exponential increase switch to linear?
A: when \( cwnd \) gets to 1/2 of its value before timeout.

**Implementation:**
- variable \( ss\text{thresh} \)
- on loss event, \( ss\text{thresh} \) is set to 1/2 of \( cwnd \) just before loss event

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

slow start
- cwnd = 1 MSS
- ssthresh = 64 KB
- dupACKcount = 0
- timeout

fast recovery
- dupACKcount == 3
- ssthresh = cwnd/2
- cwnd = ssthresh + 3
- retransmit missing segment

congestion avoidance
- cwnd = cwnd + MSS
- dupACKcount = 0
- timeout
- ssthresh = cwnd/2
- cwnd = 1 MSS
- dupACKcount = 0
- retransmit missing segment

- cwnd > ssthresh
- dupACKcount++

- dupACKcount == 3
- ssthresh = cwnd/2
- cwnd = ssthresh + 3
- retransmit missing segment

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

- new ACK
- cwnd = ssthresh
- dupACKcount = 0
- transmit new segment(s), as allowed

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

- dupACKcount == 3
- ssthresh = cwnd/2
- cwnd = ssthresh + 3
- retransmit missing segment
TCP throughput

- avg. TCP throughput as function of window size, RTT?
  - ignore slow start, assume always data to send
- $W$: window size (measured in bytes) where loss occurs
  - avg. window size (# in-flight bytes) is $\frac{3}{4} W$
  - avg. throughput is $\frac{3}{4}W$ per RTT

$$\text{avg TCP throughput} = \frac{3}{4} \frac{W}{\text{RTT}} \text{ bytes/sec}$$
TCP Futures: TCP over “long, fat pipes”

- example: 1500 byte segments, 100ms RTT, want 10 Gbps throughput
- requires $W = 83,333$ in-flight segments
- throughput in terms of segment loss probability, $L$ [Mathis 1997]:

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

→ to achieve 10 Gbps throughput, need a loss rate of $L = 2 \cdot 10^{-10}$ — a very small loss rate!

- new versions of TCP for high-speed
TCP Fairness

*fairness goal:* if $K$ TCP sessions share same bottleneck link of bandwidth $R$, each should have average rate of $R/K$
Why is TCP fair?

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

```plaintext
Connection 1 throughput  
Connection 2 throughput  
R
```

- equal bandwidth share
- loss: decrease window by factor of 2
- congestion avoidance: additive increase
- congestion avoidance: additive increase

R
Fairness (more)

**Fairness and UDP**

- multimedia apps often do not use TCP
  - do not want rate throttled by congestion control
- instead use UDP:
  - send audio/video at constant rate, tolerate packet loss

**Fairness, parallel TCP connections**

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

network-assisted congestion control:

- two bits in IP header (ToS field) marked by network router to indicate congestion
- congestion indication carried to receiving host
- receiver (seeing congestion indication in IP datagram) sets ECE bit on receiver-to-sender ACK segment to notify sender of congestion
Chapter 3: summary

- principles behind transport layer services:
  - multiplexing, demultiplexing
  - reliable data transfer
  - flow control
  - congestion control

- instantiation, implementation in the Internet
  - UDP
  - TCP

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