Chapter 3: Transport Layer

Chapter goals:

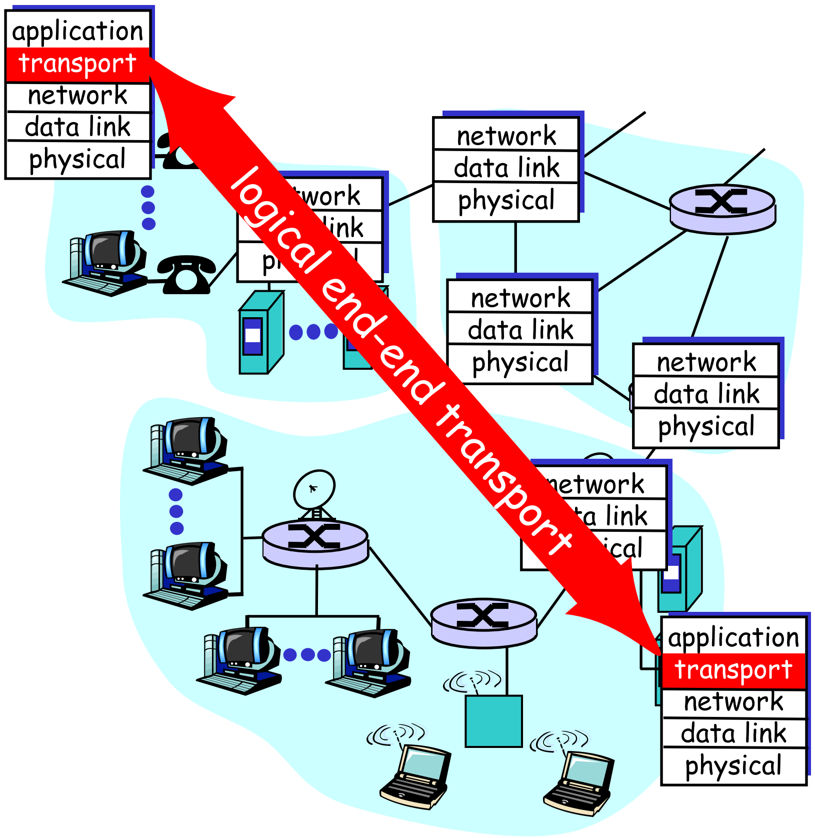
- understand principles behind transport layer services:
  - multiplexing/demultiplexing
  - reliable data transfer
  - flow control
  - congestion control
- instantiation and implementation in the Internet

Chapter Overview:

- transport layer services
- multiplexing/demultiplexing
- connectionless transport: UDP
- principles of reliable data transfer
- connection-oriented transport: TCP
  - reliable transfer
  - flow control
  - connection management
- principles of congestion control
- TCP congestion control

Transport services and protocols

- provide *logical communication* between app' processes running on different hosts
- transport protocols run in end systems
- transport vs network layer services:
- *network layer:* data transfer between end systems
- *transport layer:* data transfer between processes
  - relies on, enhances, network layer services



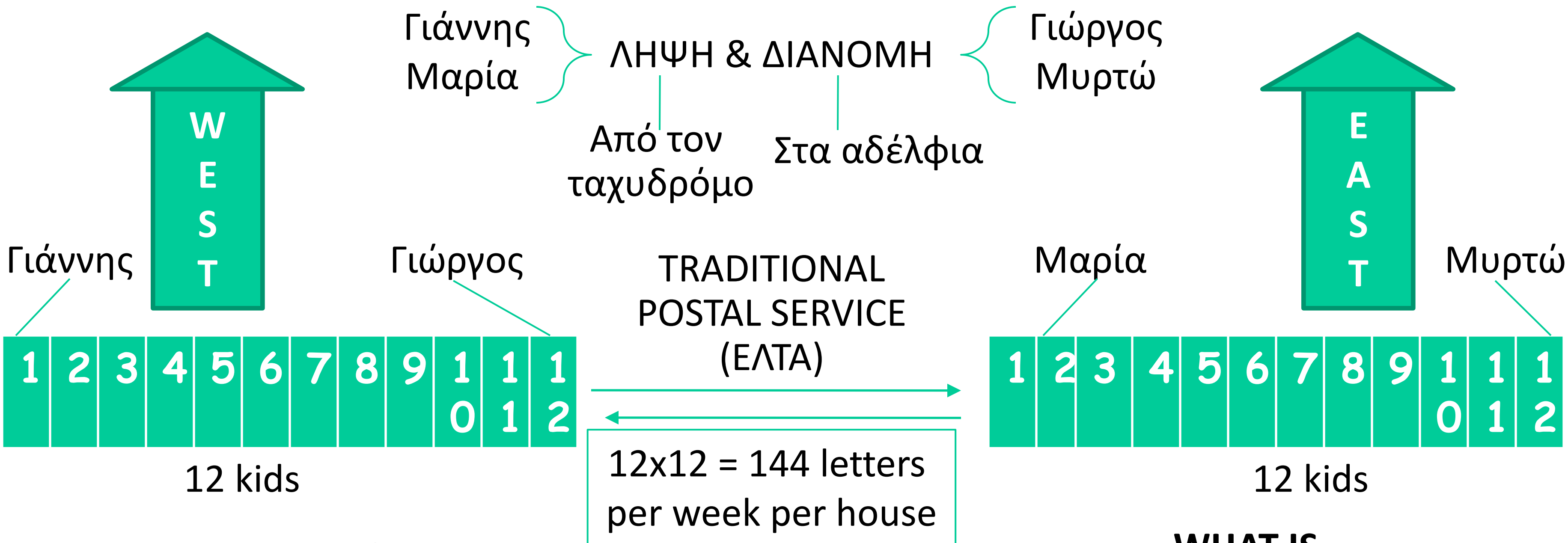
Introduction to Transport Layer Protocols

- Transport Layer Protocols provide logical communication between processes running on different hosts.
- Όταν σε έναν host καταφθάνουν πακέτα από διάφορες εφαρμογές (application processes), το πρωτόκολλο του Στρώματος Μεταφοράς είναι αυτό που ξεδιαλύνει σε ποια εφαρμογή θα οδηγηθεί κάθε πακέτο.
- Στο Διαδίκτυο έχουμε δύο πρωτόκολλα στο Στρώμα Μεταφοράς, το TCP (με αξιόπιστη λειτουργία) και το UDP (που θυσιάζει την αξιόπιστη λειτουργία χάριν της ταχύτητας – μεταφορά πληροφορίας χωρίς καθυστερήσεις).
- Ακολούθως:

  Ανθρώπινο ανάλογο που εξηγεί την λειτουργία των πρωτοκόλλων του Στρώματος Μεταφοράς και την σχέση τους με τα πρωτόκολλα του Στρώματος Δικτύου.

Ανθρώπινο ανάλογο

Δυό σπίτια, WEST – EAST, σε μια μεγαλούπολη με πολυμελείς οικογένειες - εξαδέλφια



WHAT IS

Houses – West & East

Kids

Letters (in the envelops)

ΕΛΤΑ - Ταχυδρόμοι

Γιάννης – Μαρία

Γιώργος – Μυρτώ

WHAT IS

Hosts – End systems

Application Processes

Messages

Network Layer Protocol

Transport Layer Protocol – TCP

Transport Layer Protocol – UDP

Ανθρώπινο ανάλογο (συνέχεια)

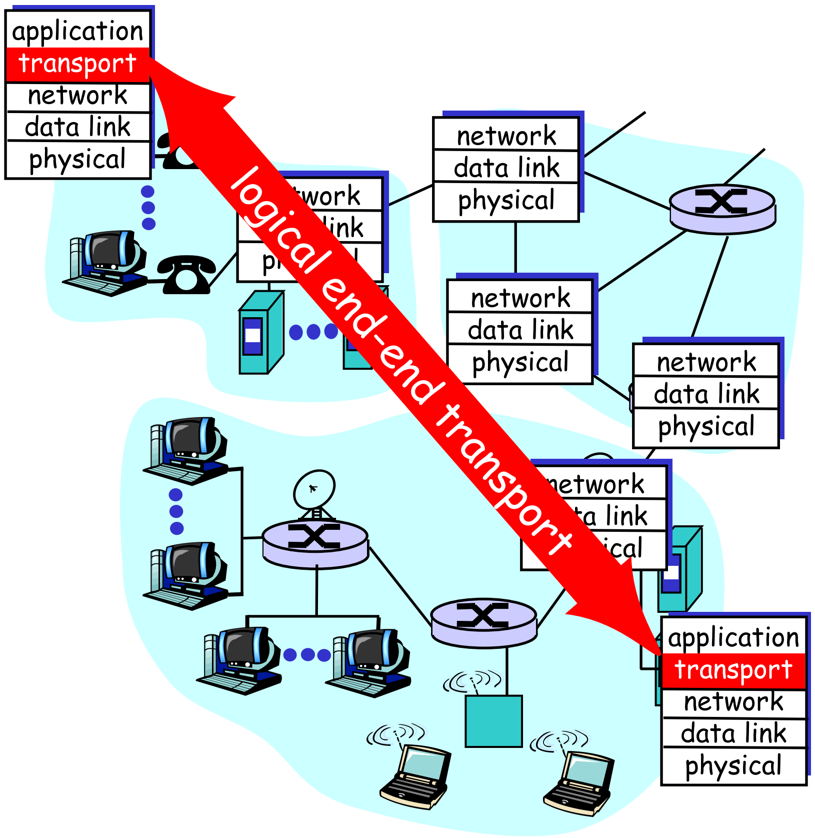
- The possible services that Γιάννης & Μαρία can provide are clearly constrained by the possible services that the ΤΑΧΥΔΡΟΜΕΙΟ provides.
- Αν το ΤΑΧΥΔΡΟΜΕΙΟ δεν εγγυηθεί max delay of (π.χ.) 3 days, ούτε ο Γιάννης & η Μαρία μπορούν να εγγυηθούν πότε θα παραδώσουν γράμματα στα αδέλφια τους.

- If the network layer protocol cannot provide delay or bandwidth guarantees for the segments (received from the transport layer and) sent to the hosts, thus
- the transport layer protocol cannot provide delay or bandwidth guarantees for the messages sent between application processes.

Transport-layer protocols

Internet transport services:

- reliable, in-order unicast delivery (TCP)
  - congestion
  - flow control
  - connection setup
- unreliable ("best-effort"), unordered unicast or multicast delivery: UDP
- services not available:
  - real-time
  - bandwidth guarantees
  - reliable multicast

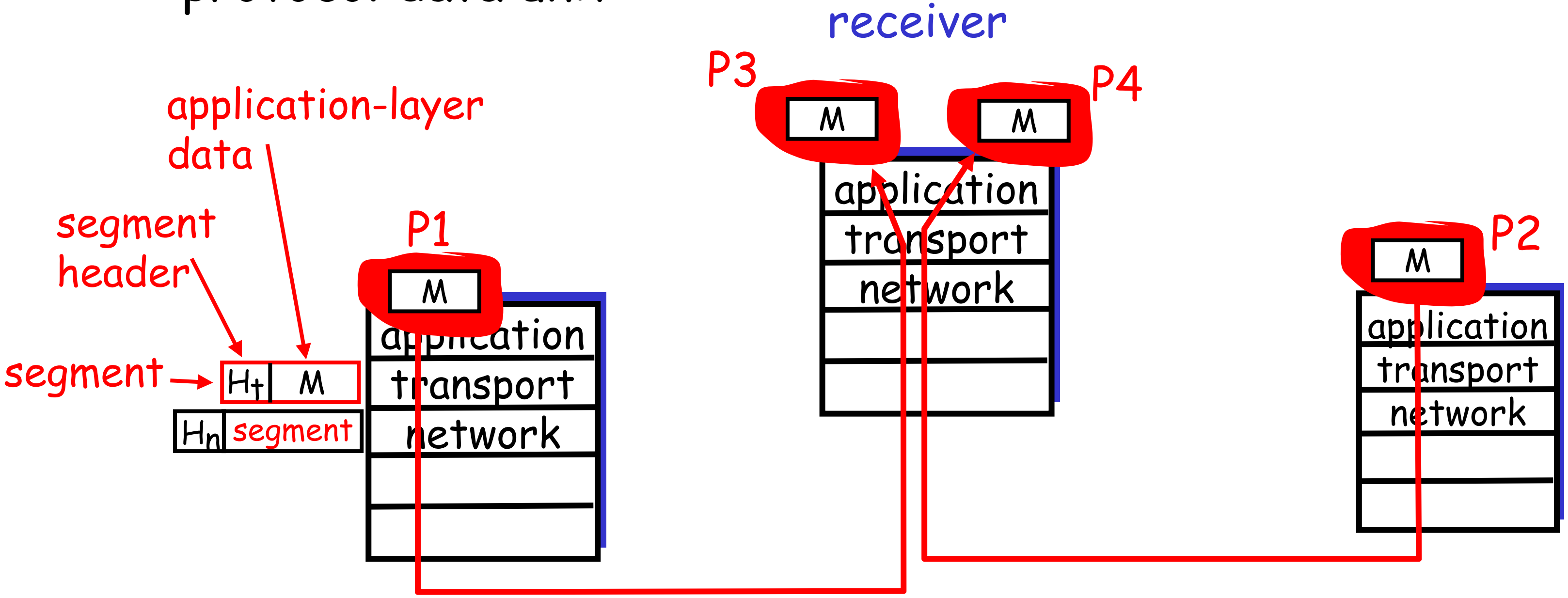


Multiplexing/demultiplexing

Recall: *segment* - unit of data exchanged between transport layer entities

- aka TPDU: transport protocol data unit

Demultiplexing: delivering received segments to correct app layer processes

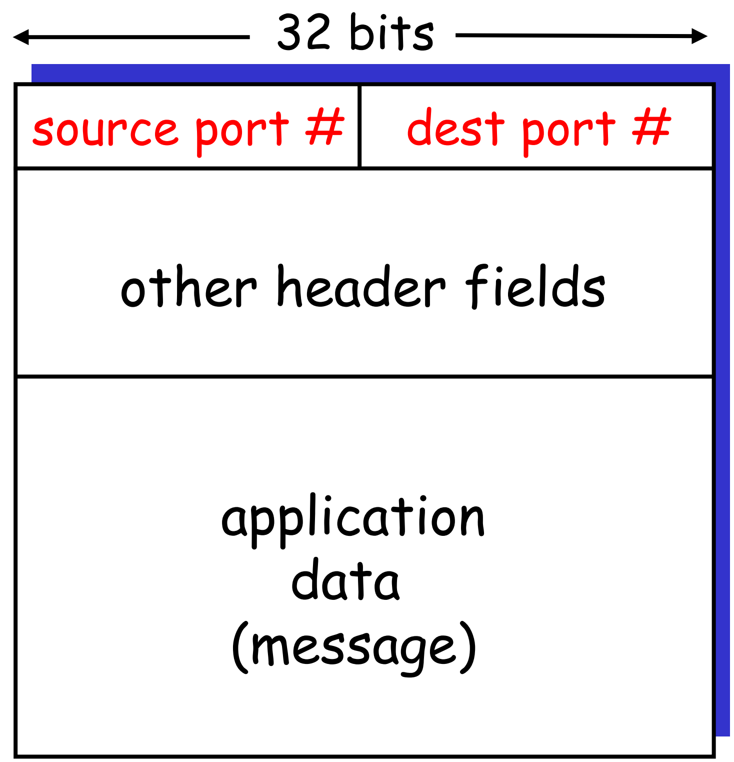


Multiplexing/demultiplexing

Multiplexing:
gathering data from multiple app processes, enveloping data with header (later used for demultiplexing)

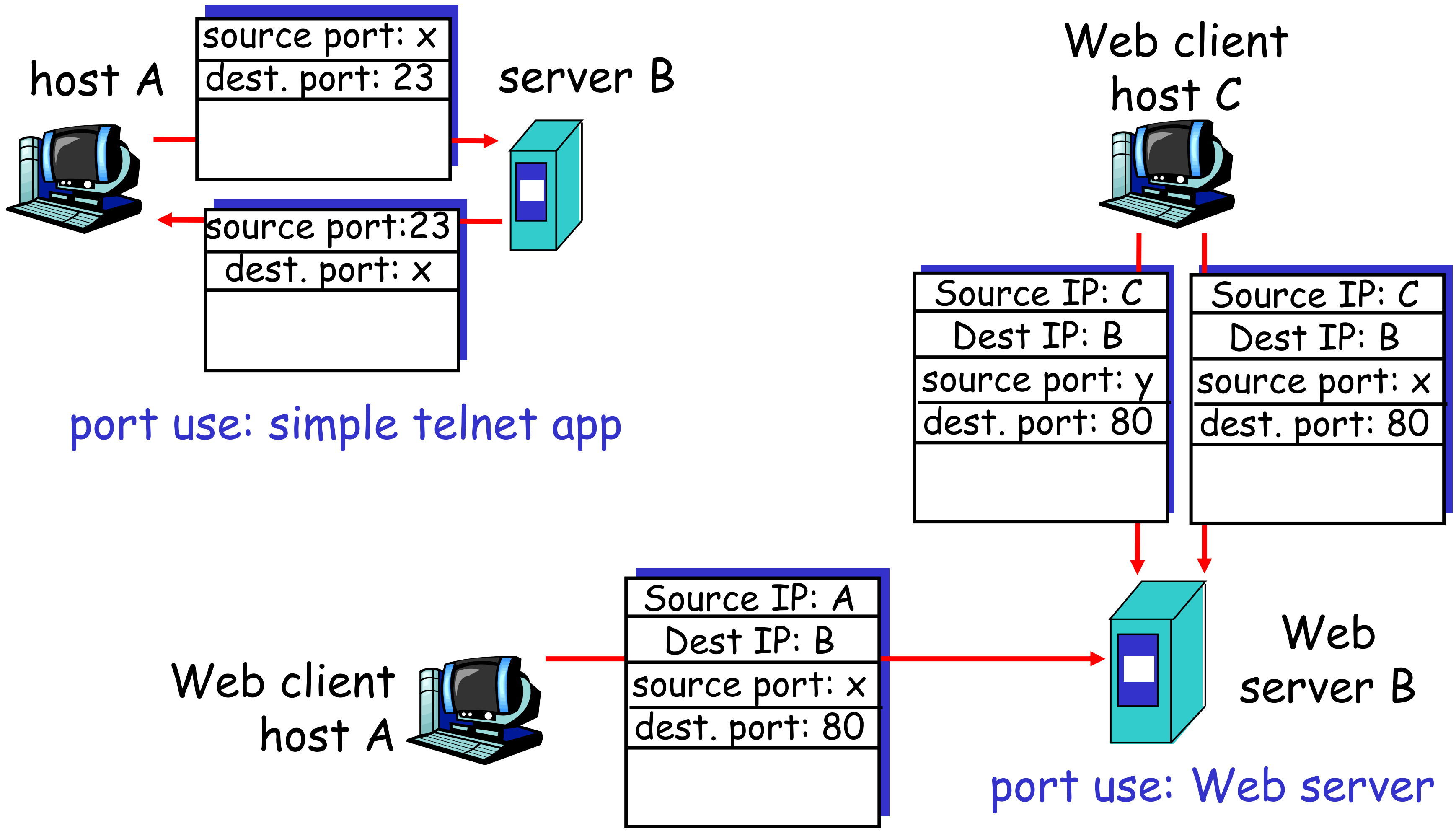
multiplexing/demultiplexing:

- based on sender, receiver port numbers, IP addresses
  - source, dest port #s in each segment
  - recall: well-known port numbers for specific applications



TCP/UDP segment format

Multiplexing/demultiplexing: examples



UDP: User Datagram Protocol [RFC 768]

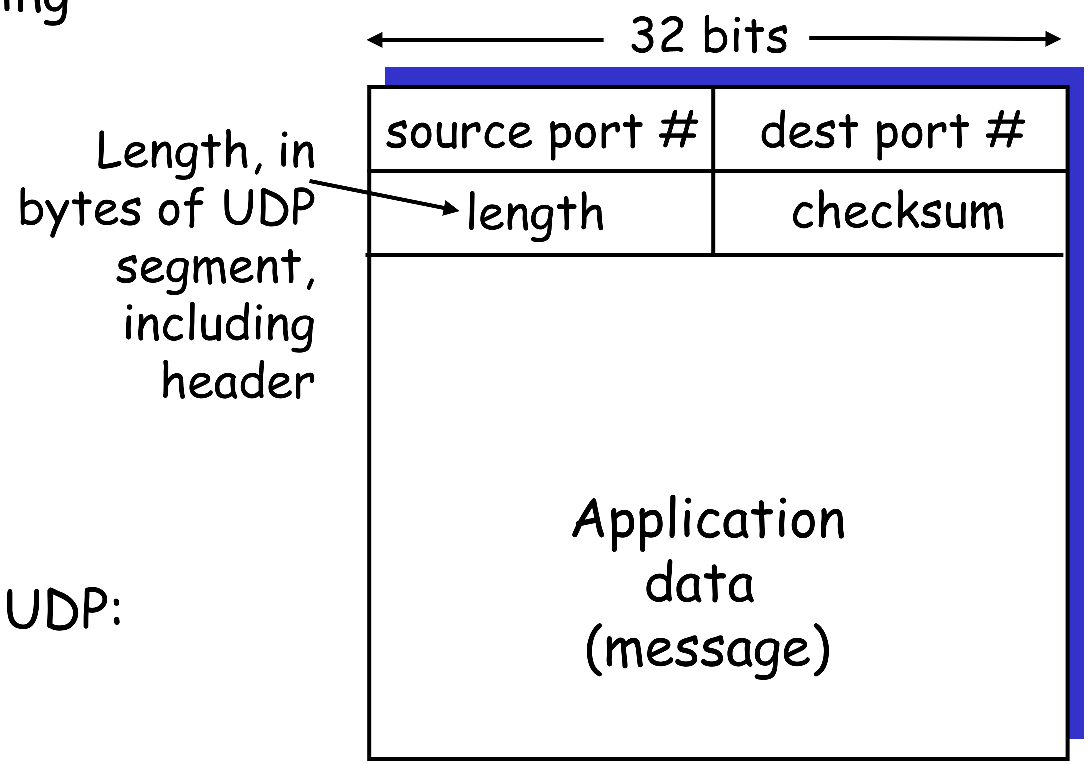
- "no frills," "bare bones" Internet transport protocol
- "best effort" service, UDP segments may be:
  - lost
  - delivered out of order to app
- *connectionless:*
  - no handshaking between UDP sender, receiver
  - each UDP segment handled independently of others

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

UDP: more

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



UDP segment format

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. *But maybe errors nonethless?* More later ....

Checksum Example

0	16	31
00001000 10001001	10000000 00010001	
10011111 00011111	11011000 01000101	
μήκος μηνύματος UDP	άθροισμα ελέγχου	
δεδομένα		

0000 1000 1000 1001
\+ 1000 0000 0001 0001
10001000 1001 1010
\+ 1001 1111 0001 1111
0010 0111 1011 1001
\+ κρατούμενο 1
0010 0111 1011 1010

Στον πομπό παίρνουμε το συμπλήρωμα του 1:

1101 1000 0100 0101

Στον δέκτη:

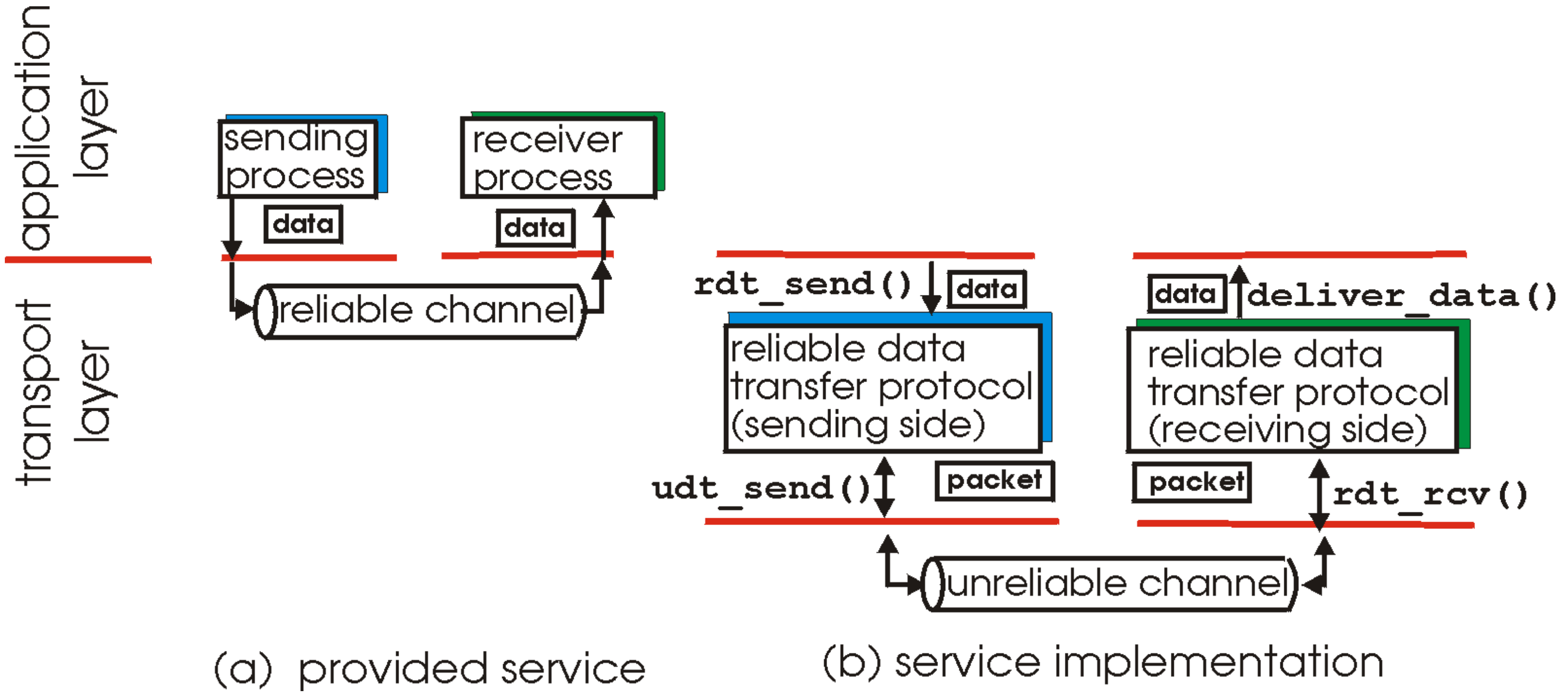
0000 1000 1000 1001
\+ 1000 0000 0001 0001
10001000 1001 1010
\+ 1001 1111 0001 1111
0010 0111 1011 1001
\+ κρατούμενο 1
0010 0111 1011 1010
\+ 1101 1000 0100 0101

1111 1111 1111 1111

Κανένα λάθος στη λήψη!

Principles of Reliable data transfer

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

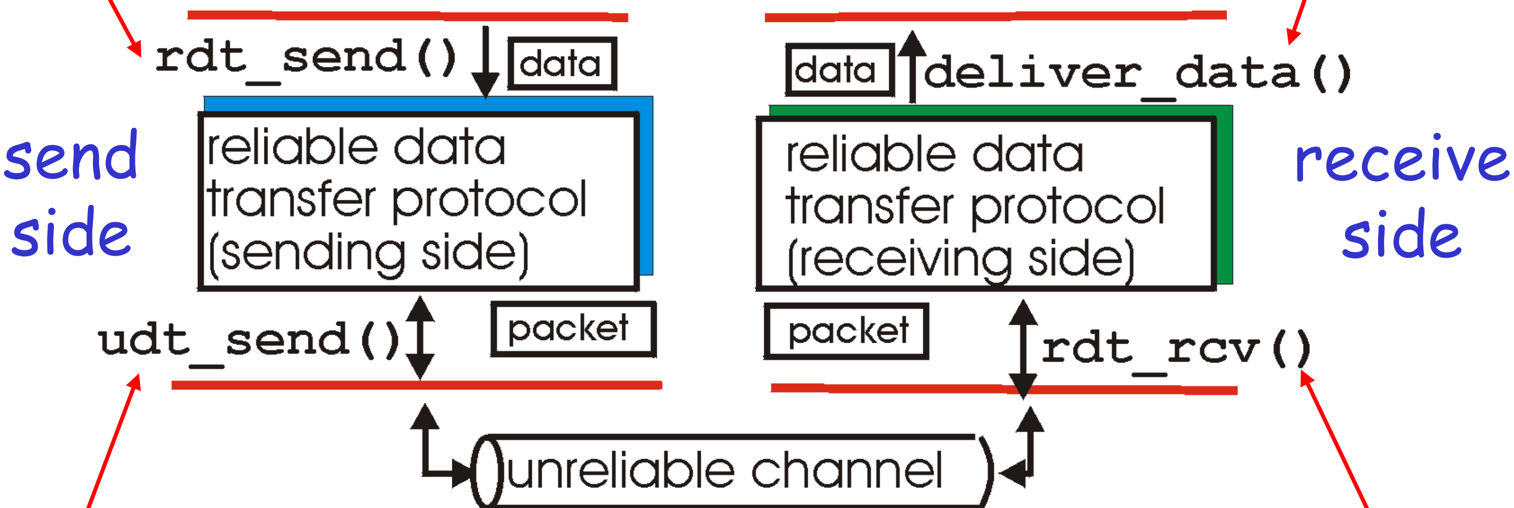


- characteristics of unreliable channel will determine complexity of reliable data transfer protocol (rdt)

Reliable data transfer: getting started

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

deliver_data() : called by **rdt** to deliver data to upper



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

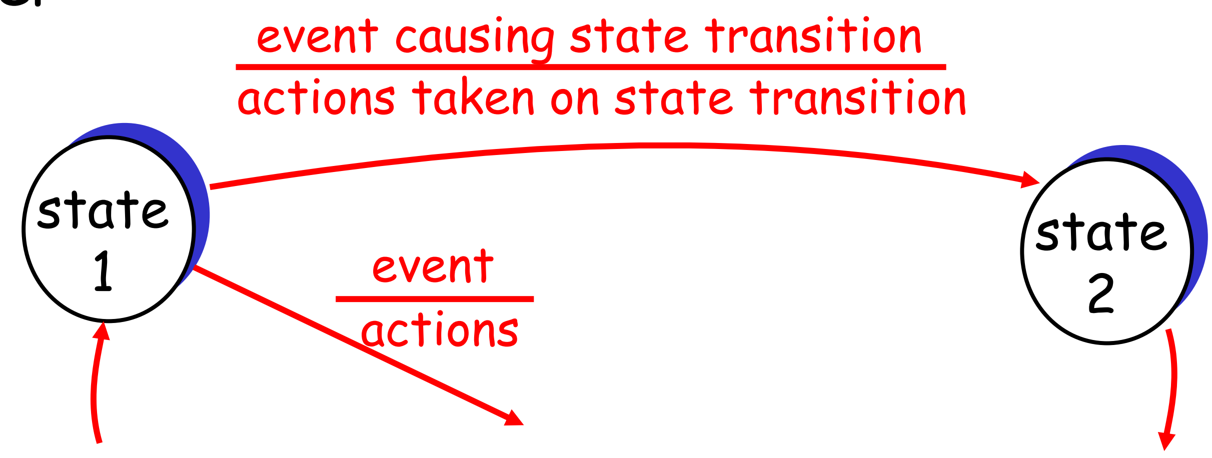
rdt_rcv() : called when packet arrives on rcv-side of channel

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 erros
  - no loss of packets
- separate FSMs for sender, receiver:
  - sender sends data into underlying channel
  - receiver read data from underlying channel

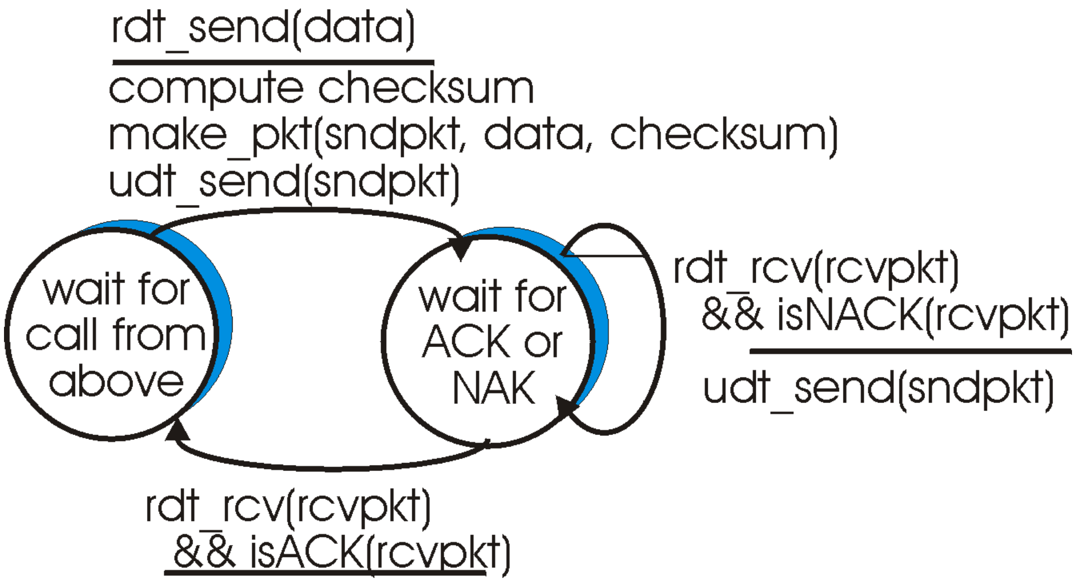
(a) rdt1.0: sending side

(b) rdt1.0: receiving side

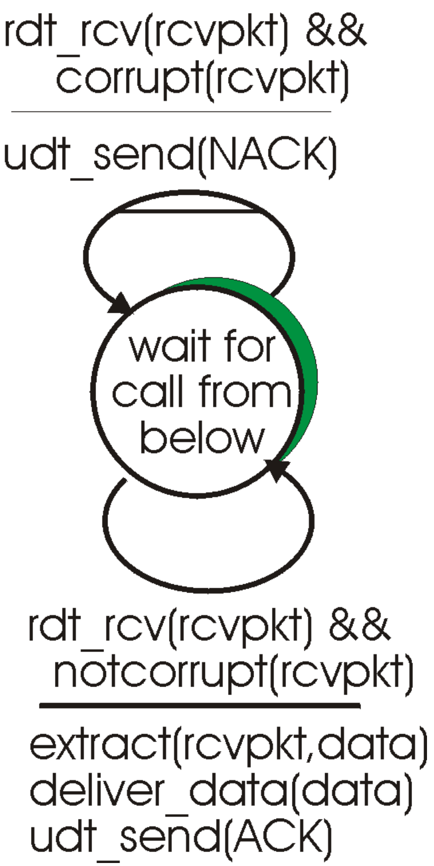
Rdt2.0: channel with bit errors

- underlying channel may flip bits in packet
  - recall: UDP 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
  - human scenarios using ACKs, NAKs?
- new mechanisms in **`rdt2.0`** (beyond **`rdt1.0`**):
  - error detection
  - receiver feedback: control msgs (ACK,NAK) rcvr->sender

rdt2.0: FSM specification

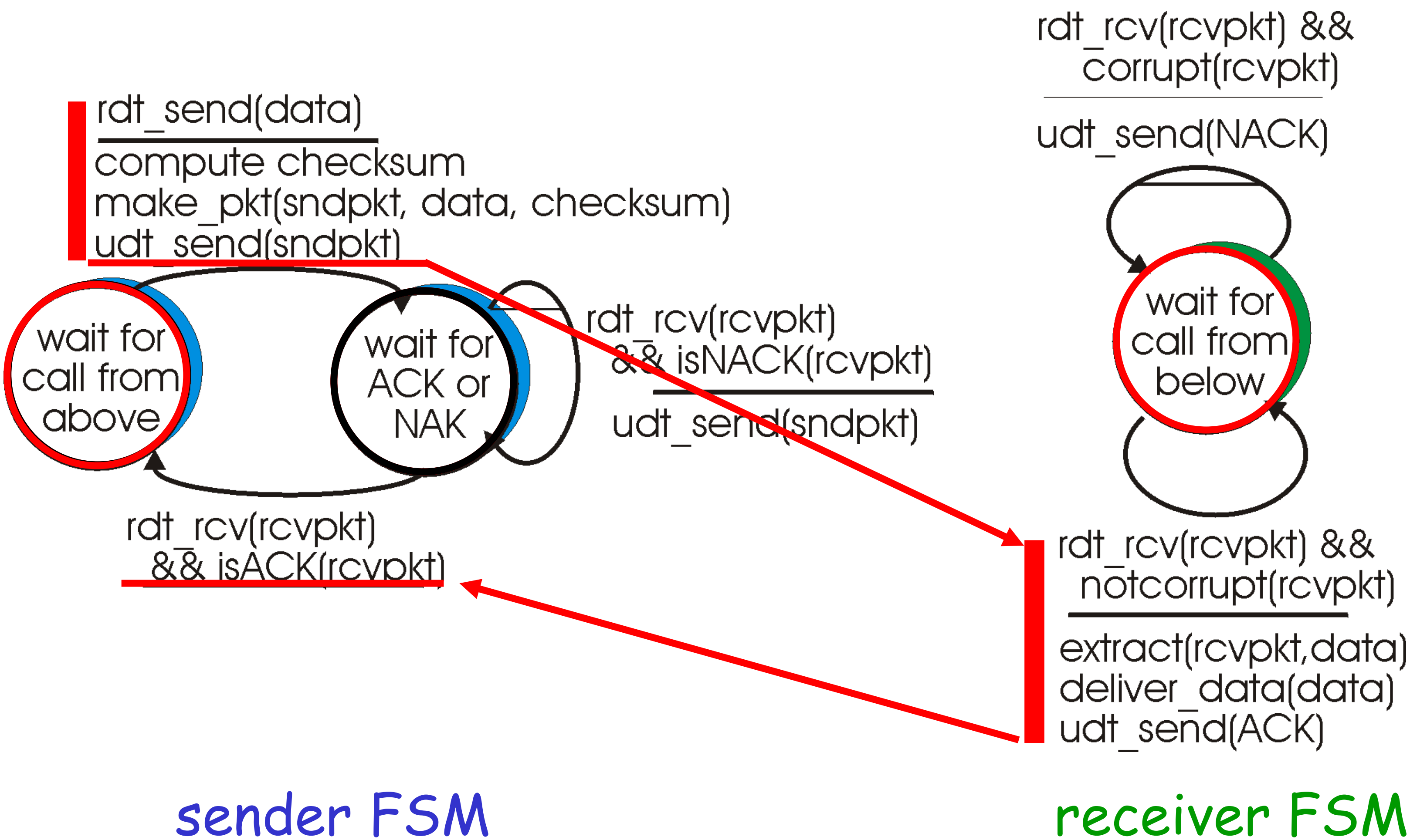


sender FSM

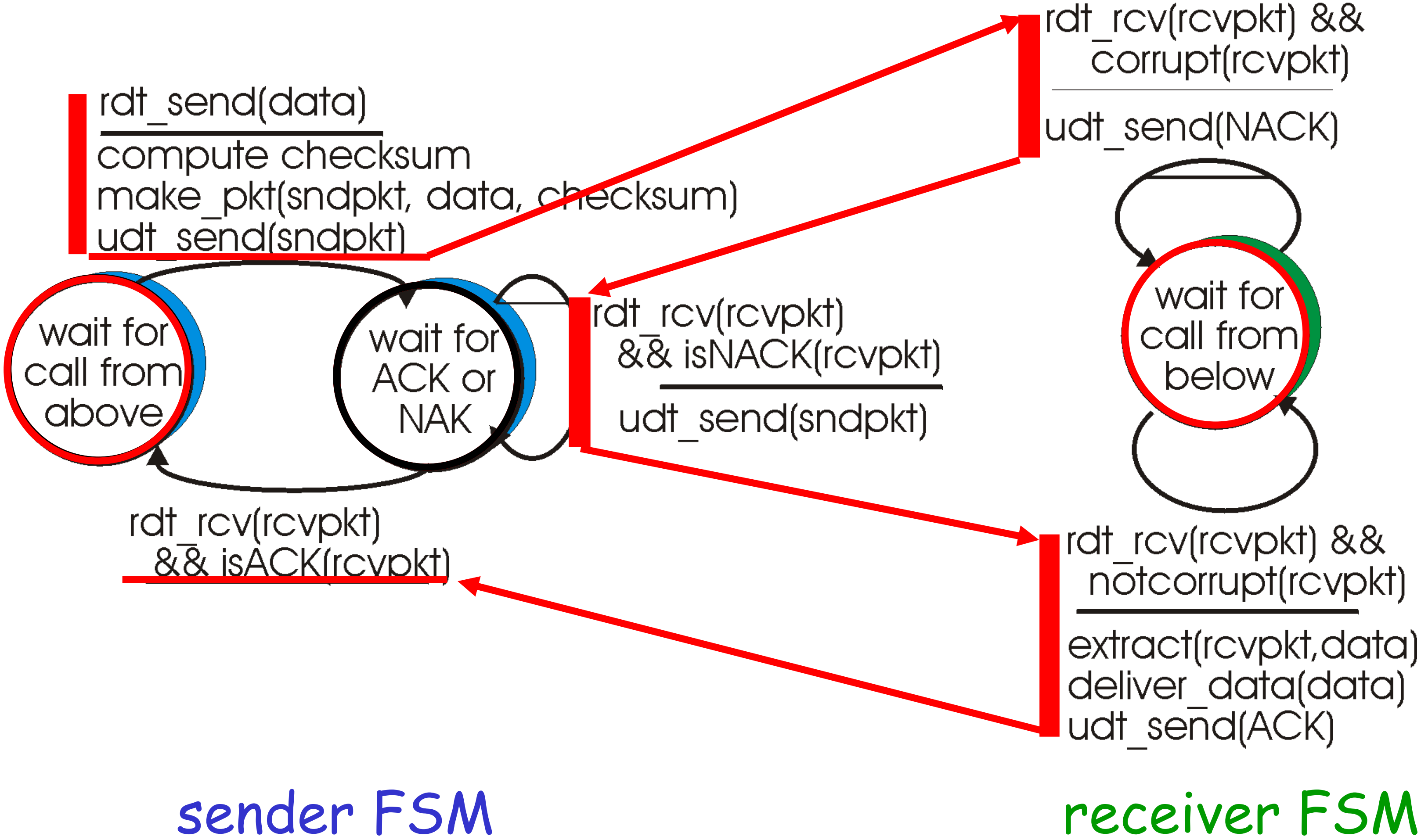


receiver FSM

rdt2.0: in action (no errors)



rdt2.0: in action (error scenario)



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

What to do?

- sender ACKs/NAKs receiver's ACK/NAK? What if sender ACK/NAK lost?
- retransmit, but this might cause retransmission of correctly received pkt!

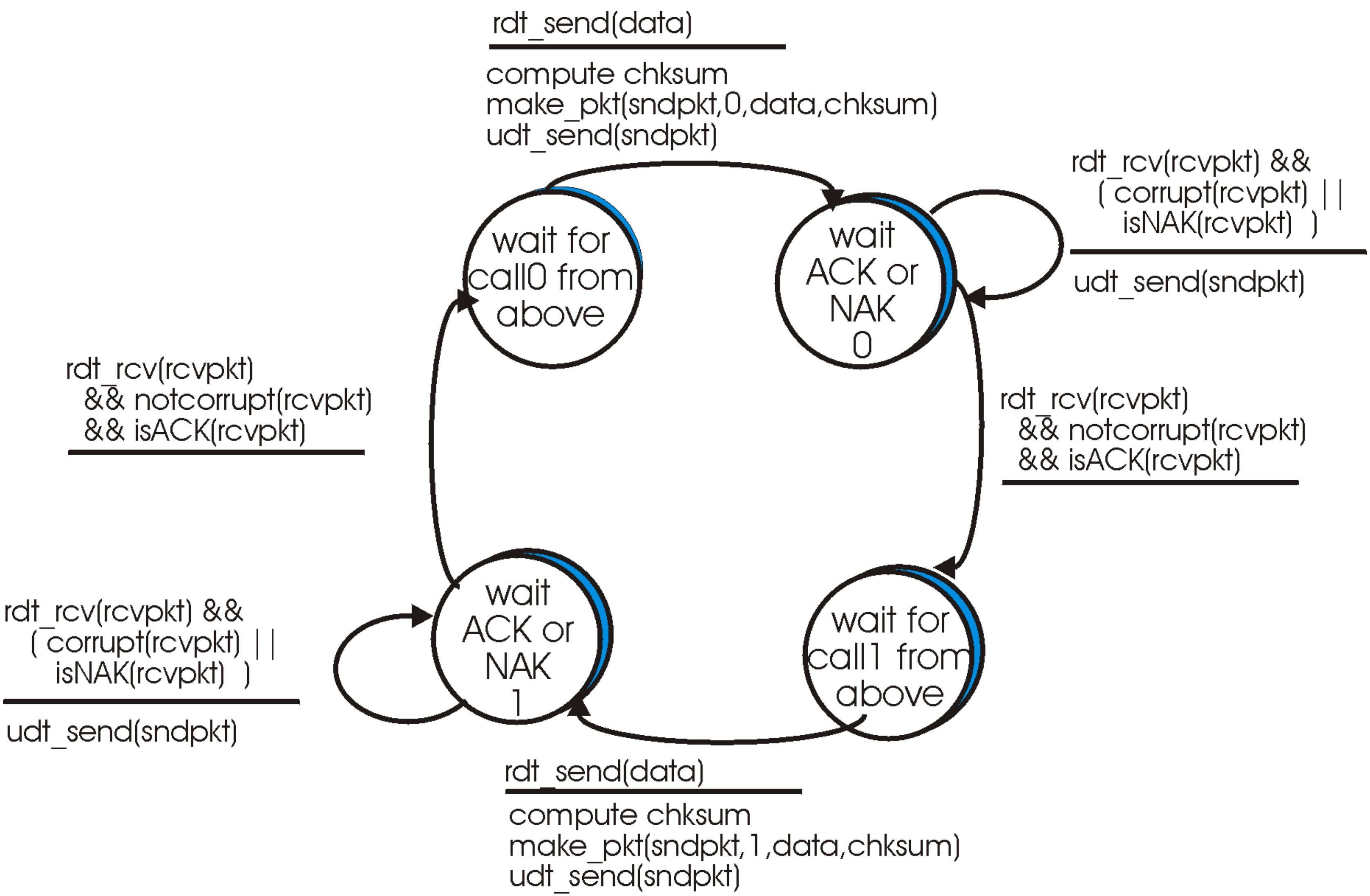
Handling duplicates:

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

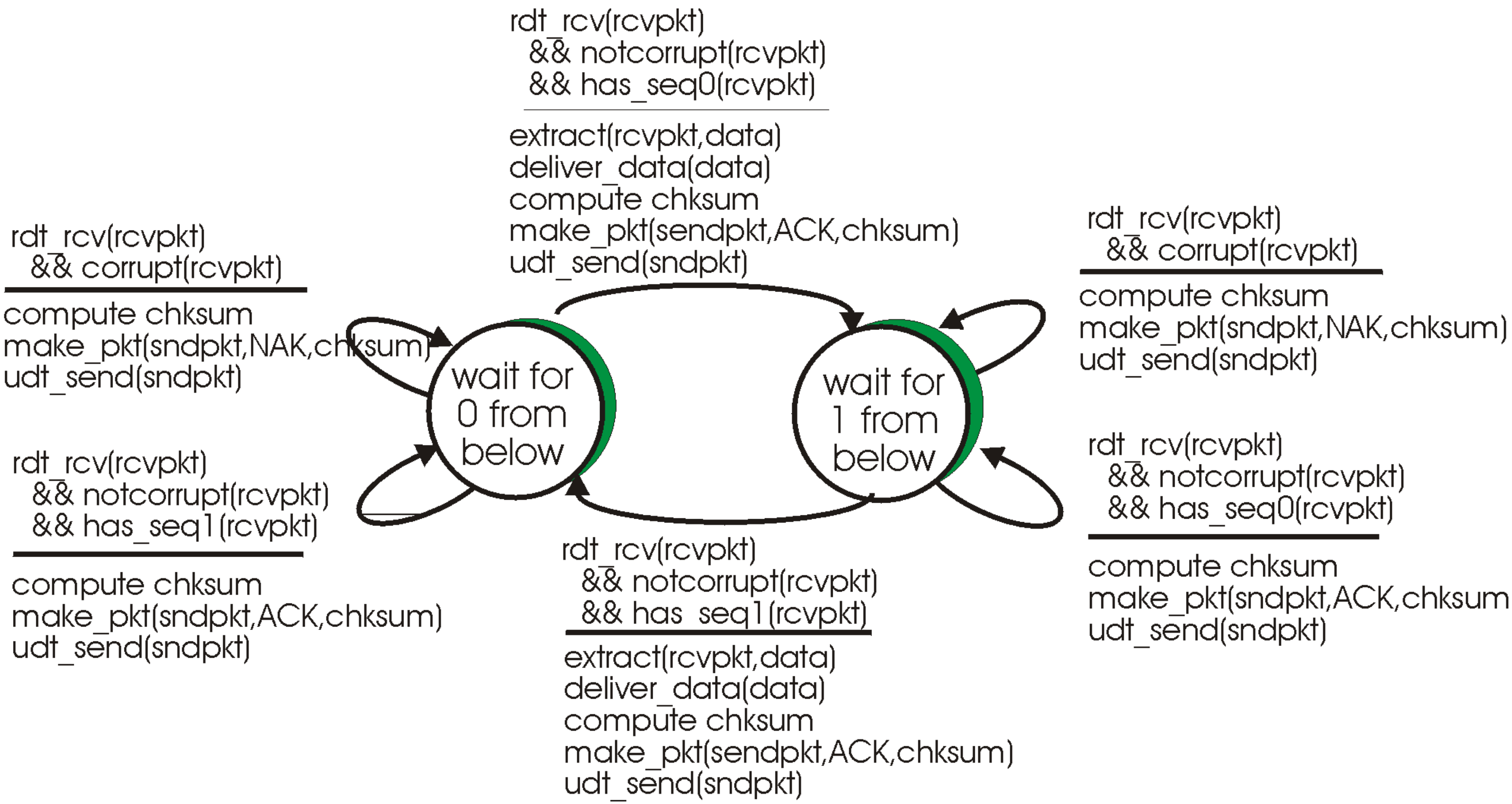
stop and wait

Sender sends one packet, then waits for receiver response

rdt2.1: sender, handles garbled ACK/NAKs



rdt2.1: receiver, handles garbled ACK/NAKs



rdt2.1: discussion

Sender:

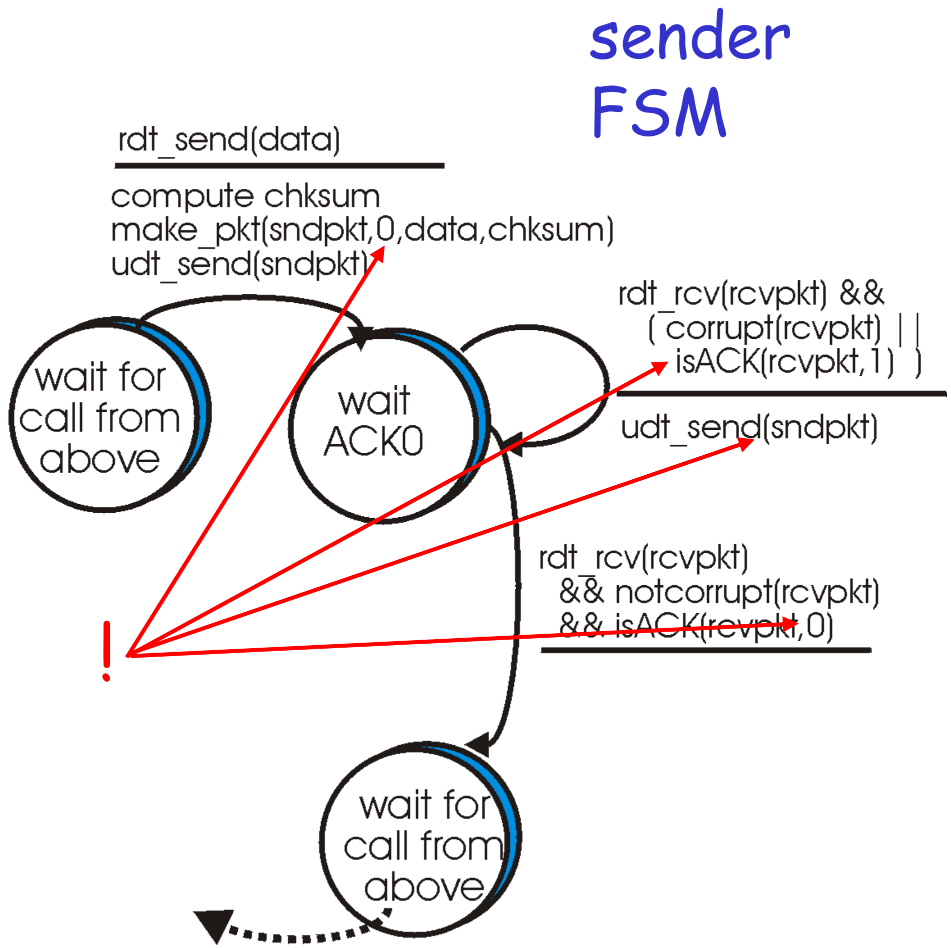
- seq # added to pkt
- two seq. #'s (0,1) will suffice. Why?
- must check if received ACK/NAK corrupted
- twice as many states
  - state must "remember" whether "current" pkt has 0 or 1 seq. #

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 NAKs 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*



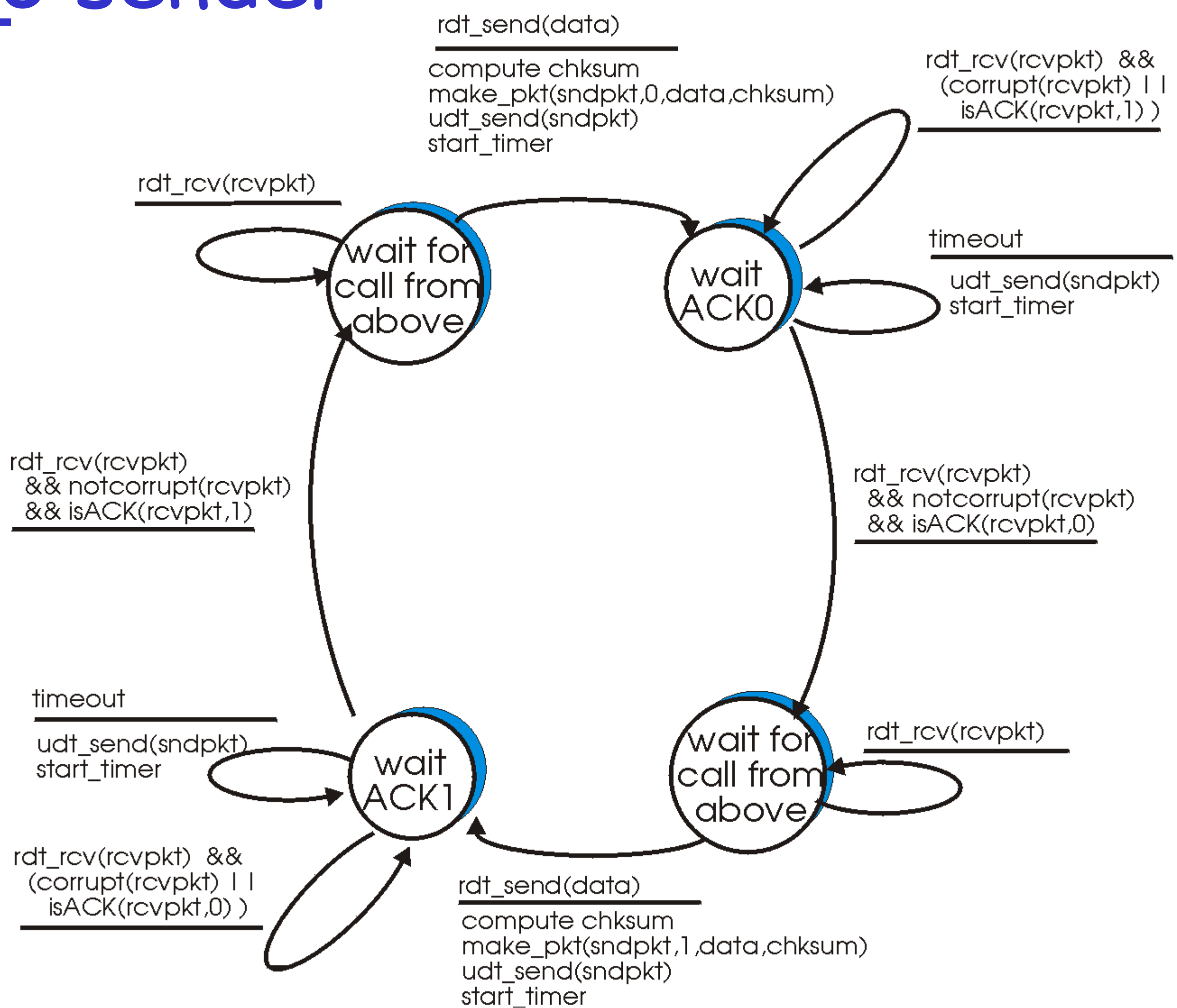
rdt3.0: channels with errors *and* loss

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

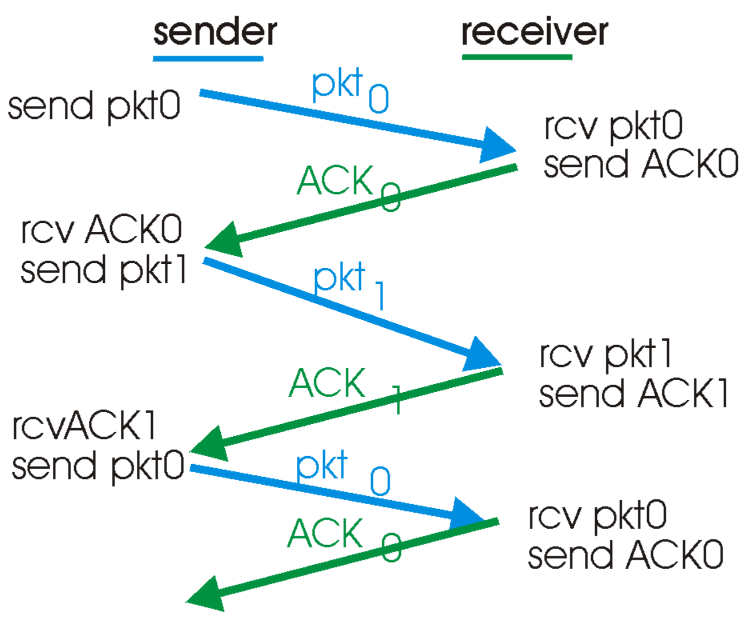
Q: how to deal with loss?
- sender waits until certain data or ACK lost, then retransmits
- yuck: drawbacks?

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 use of seq. #'s already handles this
  - receiver must specify seq # of pkt being ACKed
- requires countdown timer

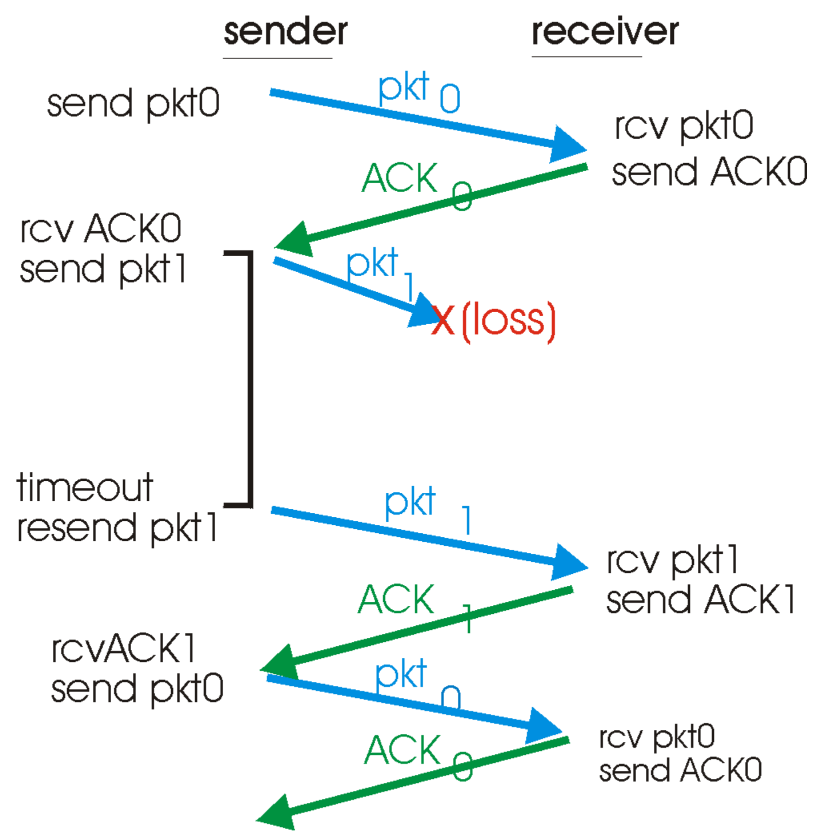
rdt3.0 sender



rdt3.0 in action

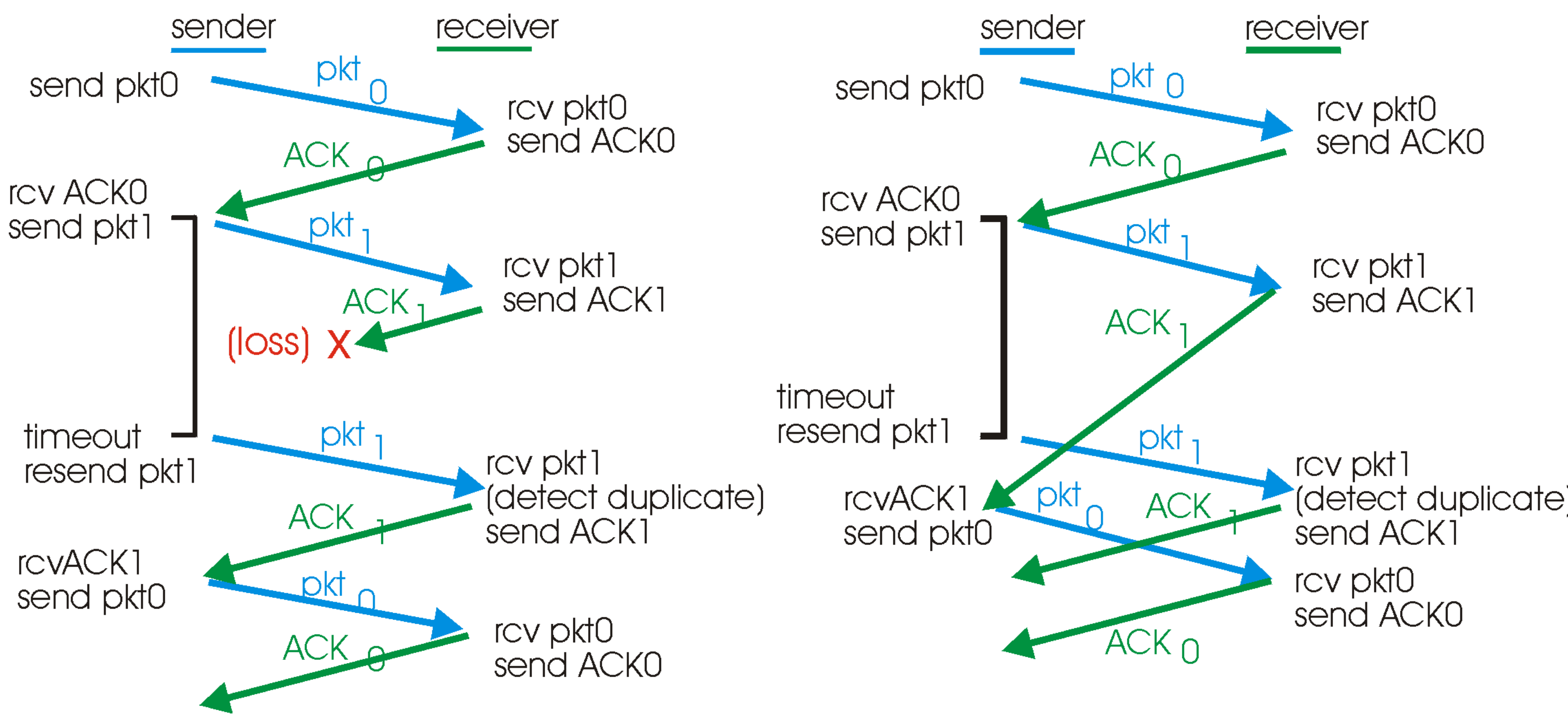


(a) operation with no loss



(b) lost packet

rdt3.0 in action



(c) lost ACK

(d) premature timeout

Performance of rdt3.0

- rdt3.0 works, but performance stinks
- example: 1 Gbps link, 15 ms e-e prop. delay, 1KB packet:

$$T_{transmit} = \frac{8\text{kb/pkt}}{10^{**}9\text{ b/sec}} = 8 \text{ microsec}$$

$$\text{Utilization} = U = \frac{\text{fraction of time}}{\text{sender busy sending}} = \frac{8 \text{ microsec}}{30.008 \text{ msec}} = 0.00027$$

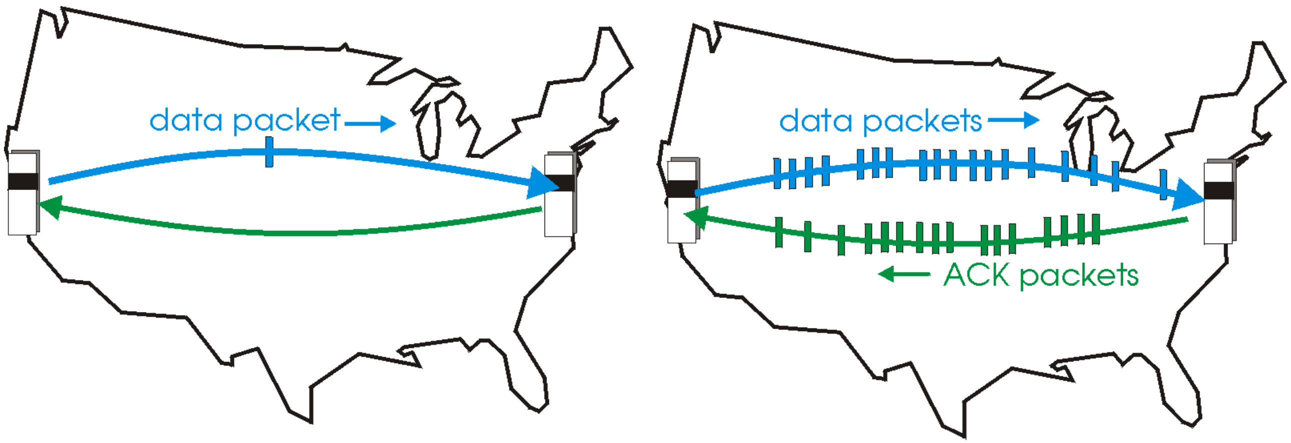
30.008 msec = total propagation delay + transfer delay, when ignoring the transmission delay of ACK

- 1KB pkt every 30 msec -> 33kB/sec throughput over 1 Gbps link
- network protocol limits use of physical resources!

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



(a) a stop-and-wait protocol in operation

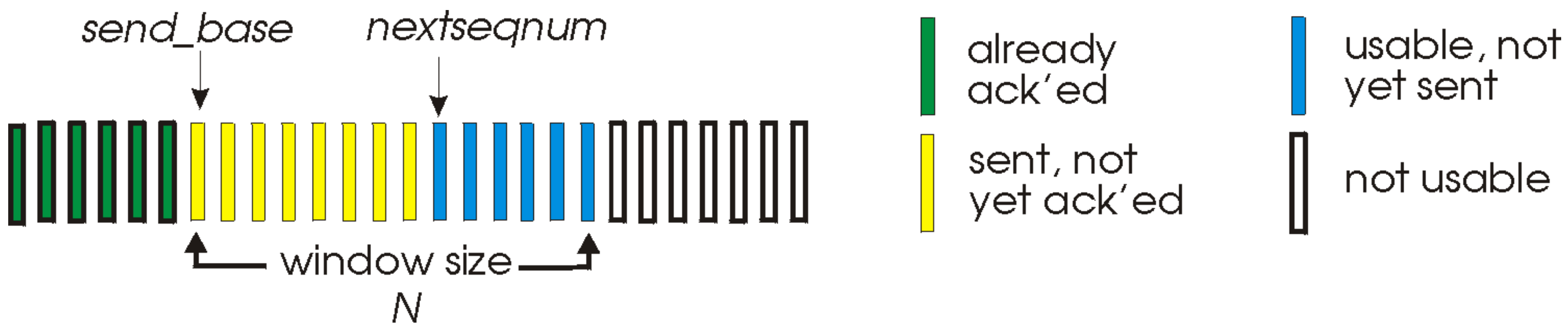
(b) a pipelined protocol in operation

- Two generic forms of pipelined protocols: *go-Back-N, selective repeat*

Go-Back-N

Sender:

- k-bit seq # in pkt header
- "window" of up to N, consecutive unack'ed pkts allowed

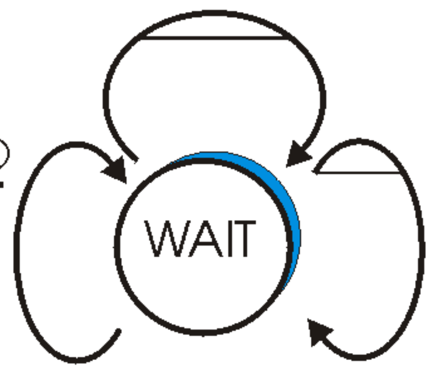


- ACK(n): ACKs all pkts up to, including seq # n - "cumulative ACK"
  - may deceive duplicate ACKs (see receiver)
- timer for each in-flight pkt
- *timeout(n)*: retransmit pkt n and all higher seq # pkts in window

GBN: sender extended FSM

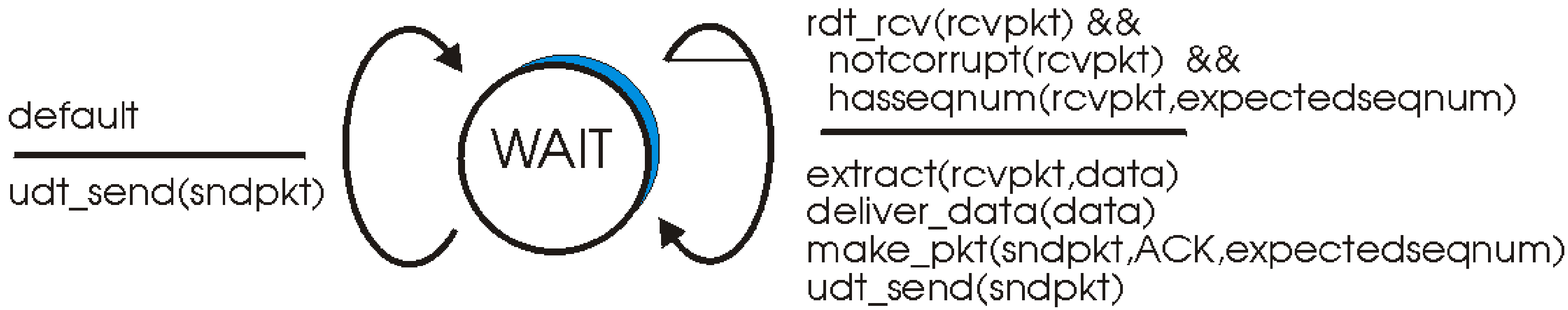
```
rdt_send(data)
______________
if (nextseqnum < base+N) {
   compute chksum
   make_pkt(sndpkt(nextseqnum)),nextseqnum,data,chksum)
   udt_send(sndpkt(nextseqnum))
   if (base == nextseqnum)
      start_timer
   nextseqnum = nextseqnum + 1
   }
else
  refuse_data(data)
```

```
rdt_rcv(rcv_pkt) && notcorrupt(rcvpkt)
______________
base = getacknum(rvcpkt)+1
if (base == nextseqnum)
    stop_timer
  else
    start_timer
```




```
timeout
______________
start_timer
udt_send(sndpkt(base))
udt_send(sndpkt(base+1))
......
udt_send(sndpkt(nextseqnum-1))
```

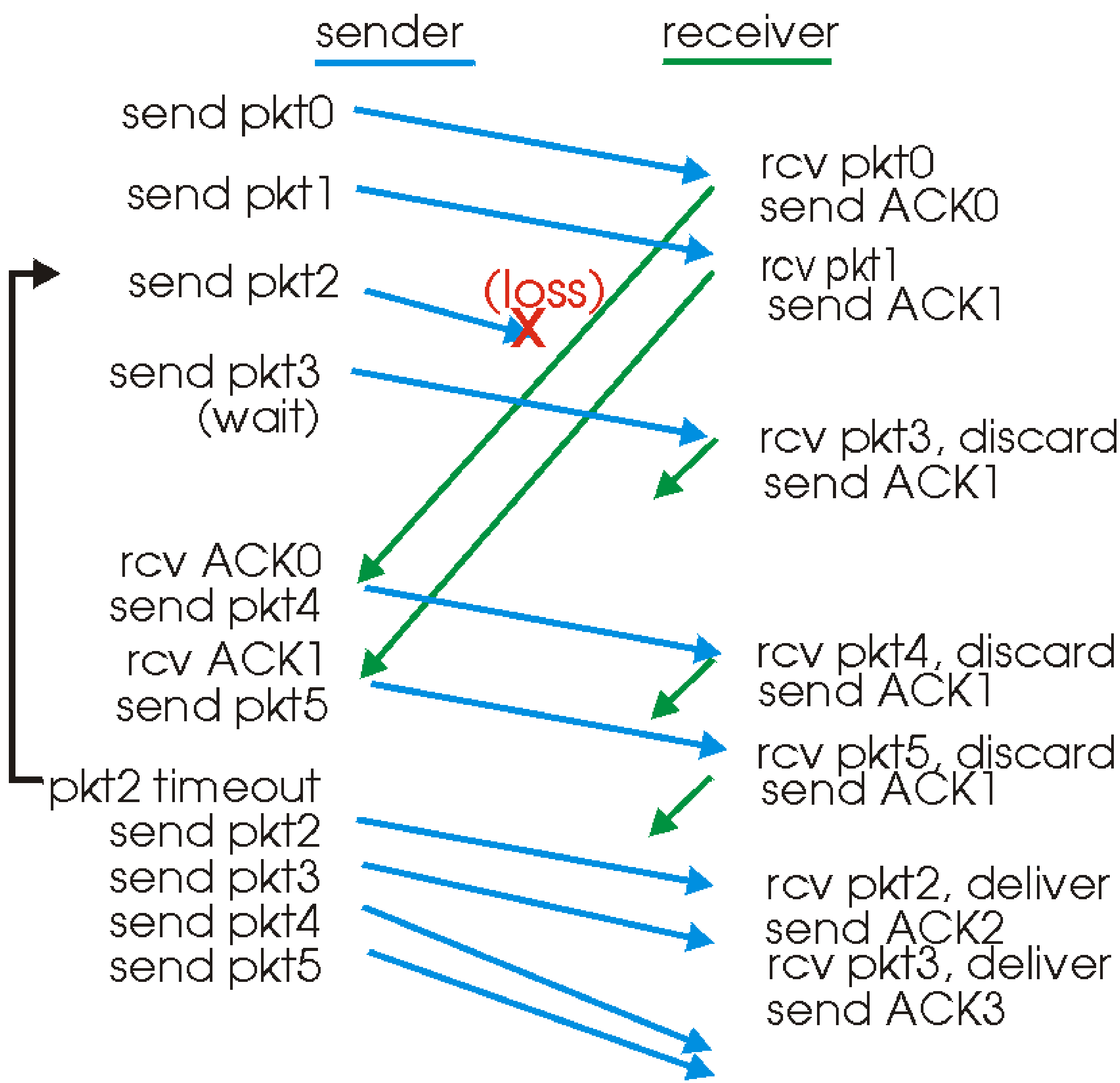
GBN: receiver extended FSM



receiver simple:

- ACK-only: always send ACK for correctly-received pkt with highest *in-order* seq #
  - may generate duplicate ACKs
  - need only remember **expectedseqnum**
- out-of-order pkt:
  - discard (don't buffer) -> no receiver buffering!
  - ACK pkt with highest in-order seq #

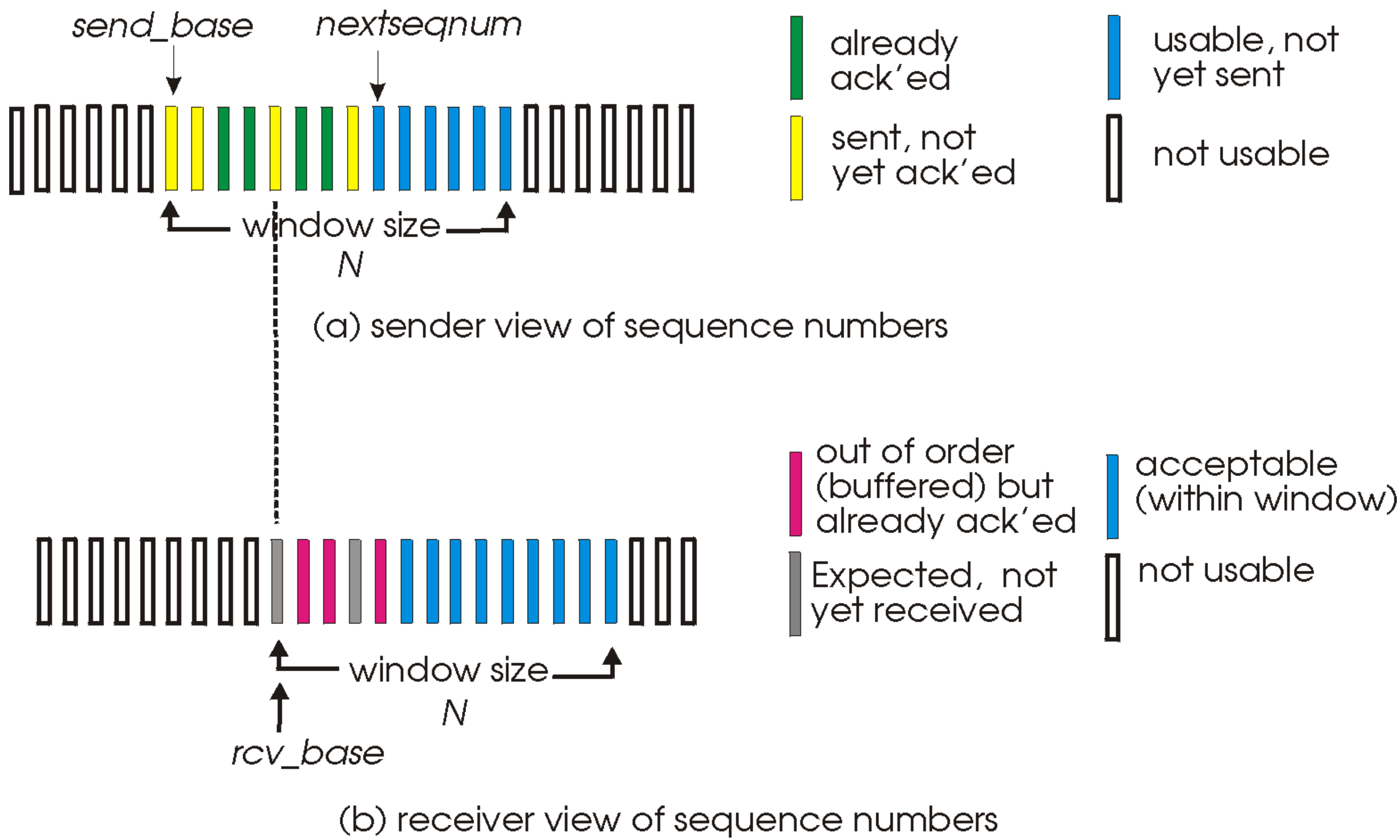
GBN in action



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
  - again limits seq #s of sent, unACKed pkts

Selective repeat: sender, receiver windows



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

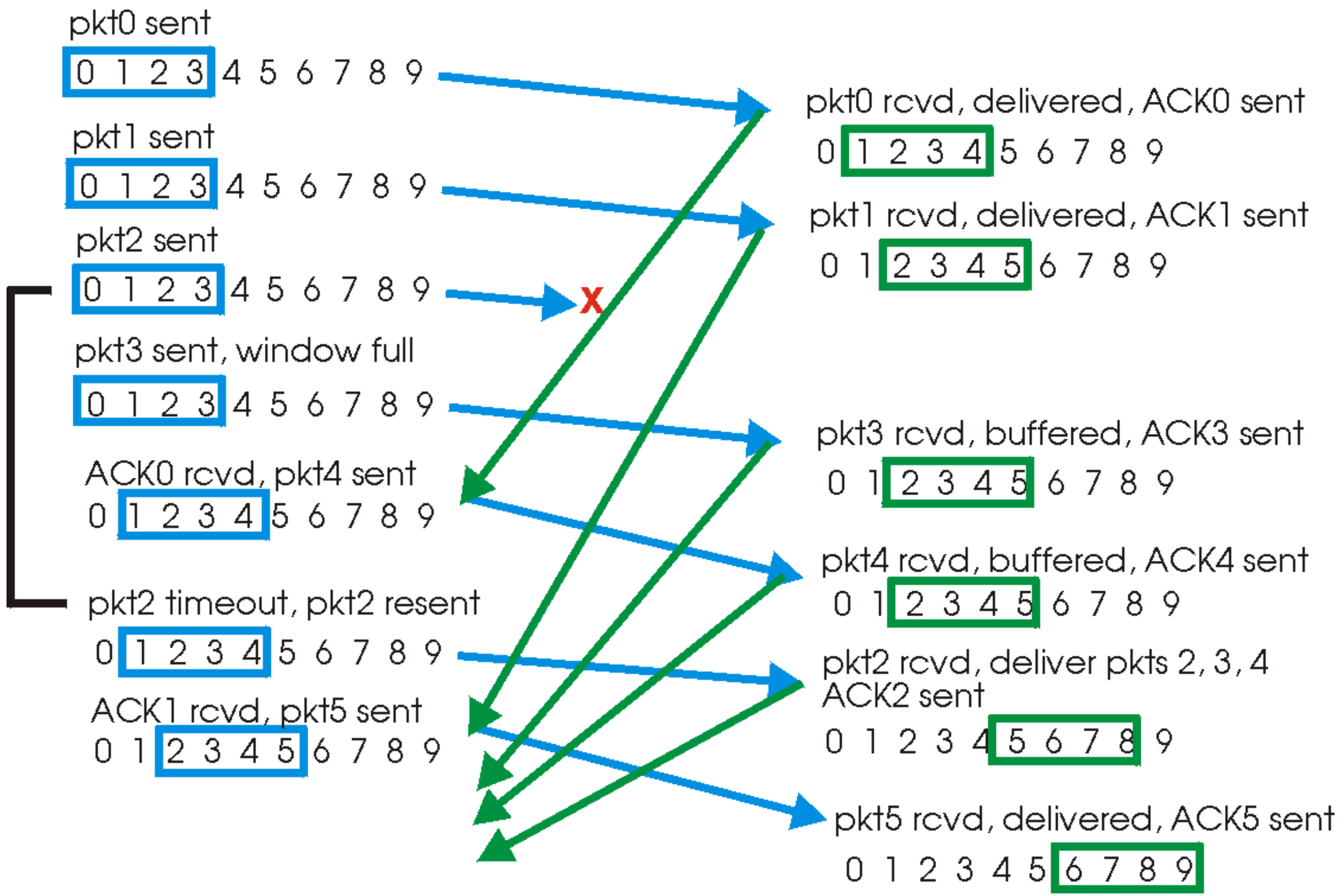
pkt n in [rcvbase-N,rcvbase-1]

- ACK(n)

otherwise:

- ignore

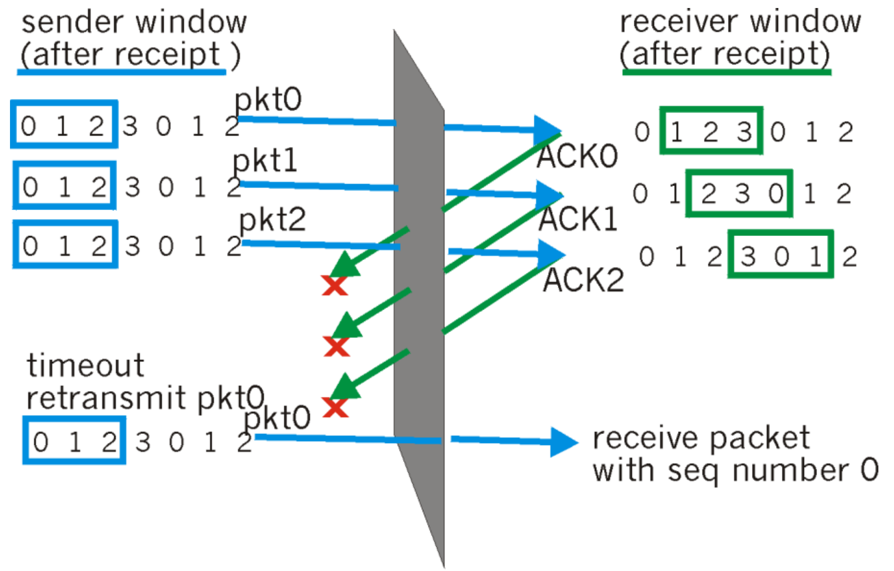
Selective repeat in action



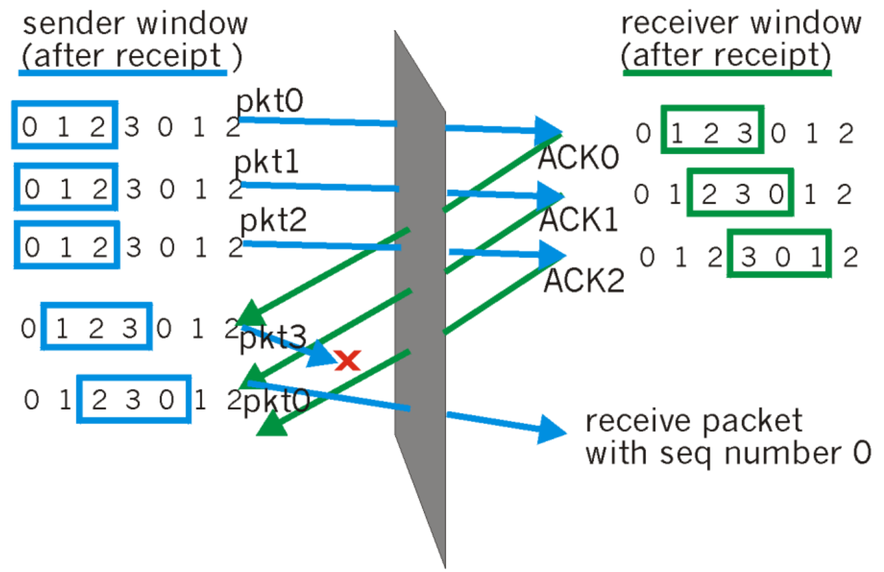
Selective repeat: dilemma

Example:

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



- receiver sees no difference in two scenarios!
- incorrectly passes duplicate data as new in (a)



Q: what relationship between seq # size and window size?