CSE/EE 461: Introduction to Computer Communications Networks
Autumn 2007

Module 3
Direct Link Networks – Part A

John Zahorjan
zahorjan@cs.washington.edu
534 Allen Center

This Module's Topics

Overview of Computer Networking

1. Two slides from last time
2. Overview – Scope of today's discussion
3. Encoding / Framing / Error Detection
4. Reliable Transmission

OSI “Seven Layer” Reference Model

Application
Presentation
Session
Transport
Network
Link
Physical

Their functions:

- Up to the application
- Encode/decode messages
- Manage connections
- Reliability, congestion control
- Routing
- Framing, multiple access
- Symbol coding, modulation

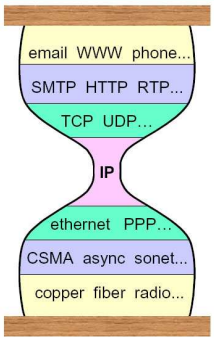
Internet Protocol Framework

Application
Transport
Network
Link

Model

Many (HTTP,SMTP)
TCP / UDP
IP
Many (Ethernet, …)

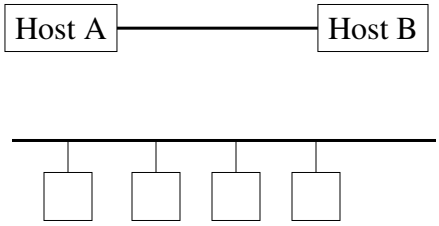
Protocols



The “narrow waist”

Direct Link Networks

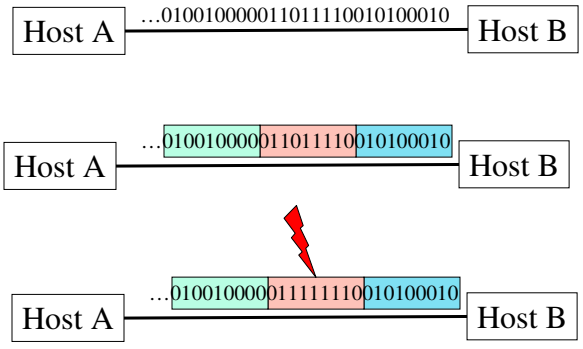
"Direct link" ⇒ no switching/routing



Point-to-point

Broadcast / shared

Direct Link Networks

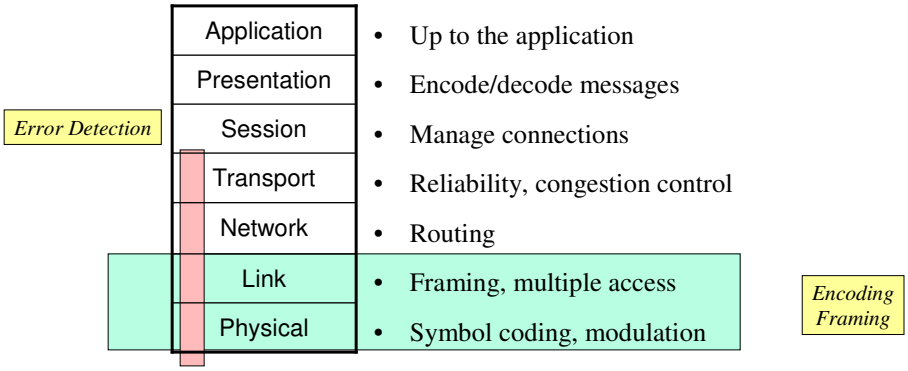


Encoding

Framing

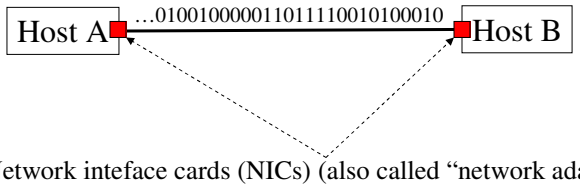
Error Detection

Relationship to the Protocol Stack



Remember, this is an idealization of what actually goes on (and the organization of the book is explicitly non-layerist).

Relationship to the hardware

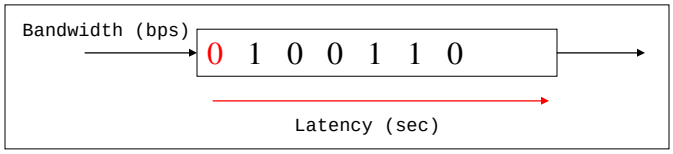


Network inteface cards (NICs) (also called “network adaptors”)

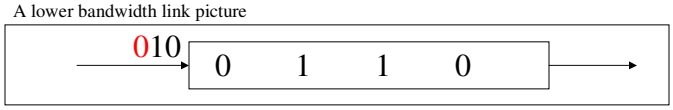
e.g., Ethernet card or 802.11g card

Link Properties

- Bandwidth
  - How many bits per second can be put onto the link.
  - (That's also the rate at which they come out, assuming there are any there to come out.)
- Latency (aka propogation delay or just delay)
  - How long it takes a bit to make it from the sender to the receiver



A lower bandwidth link picture



Bandwidth-Delay Product

- *Propogation delay* is

  `L = link distance (m) / propogation speed (m/sec)`

- *Bandwidth* is

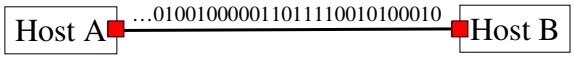
  `B = k (bits/sec)`

- *Bandwidth-Delay Product* is the number of bits that would be in flight if the sender were constantly sending

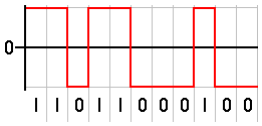
  `B*D ((bits/sec) * (m / (m/sec)) = bits)`

- Increasing bandwidth increases the bandwidth-delay product.

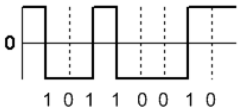
Encoding



- Modulate something – amplitude, frequency, phase
- A key issue is clocking
  - Higher transmission rates require better synch
- Some example encodings (thanks, *wikipedia)*:



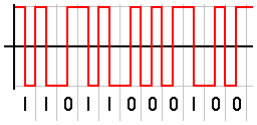
NRZ
(RS-232)



NRZI
(CDs, USB, Fast Ethernet)

Encoding: Self-Clocking

- Receiver can derive clock from the data signal
- Example 1:



Manchester
(10Mbps Ethernet)

- Example 2: Use NRZI, but make sure there are transitions
  - 4B/5B multi-level transition (MLT)
    - 100Mbps Ethernet, with 3 levels of signal
  - 8B/10B MLT
    - 1000Mbps Ethernet, with 5 levels of signal
  - (MLT is used to limit the required signal bandwidth to what can be carried on cheap, CAT 5 cable (100MHz).)

Example 4B/5B Code

Name	4b	5b	Description
0	0000	11110	hex data 0
1	0001	01001	hex data 1
2	0010	10100	hex data 2
3	0011	10101	hex data 3
4	0100	01010	hex data 4
5	0101	01011	hex data 5
6	0110	01110	hex data 6
7	0111	01111	hex data 7
8	1000	10010	hex data 8
9	1001	10011	hex data 9
A	1010	10110	hex data A
B	1011	10111	hex data B
C	1100	11010	hex data C
D	1101	11011	hex data D
E	1110	11100	hex data E
F	1111	11101	hex data F
I	-NONE-	11111	Idle
J	-NONE-	11000	SSD #1
K	-NONE-	10001	SSD #2
T	-NONE-	01101	ESD #1
R	-NONE-	00111	ESD #2
H	-NONE-	00100	Halt

(100BASE-TX Ethernet) |
SSD= Start of Stream Delimiter
ESD= End of Stream Delimiter

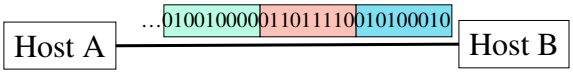
Separate Clock Distribution

- Self-clocking consumes bandwidth
  - Manchester: two transitions per bit
  - 4B/5B and 8B/10B: overhead of additional bits
- Alternative: send explicit clock
  - SONET (Synchronous Optical NETwork)
    - Clock can be carried explicitly from one network element to another
    - Nodes can all use clock from GPS
    - Various fallbacks

Table A: Stratum Clock Requirements and Hierarchy

Stratum	Accuracy/Adjust Range	Pull-In-Range	Stability	Time To First Frame Slip *
1	1×10^{-11}	N/A	N/A	72 Days
2	1.6×10^{-8}	Must be capable of synchronizing to clock with accuracy of $+/-1.6 \times 10^{-8}$	1×10^{-10}/day	7 Days
3E	1.0×10^{-6}	Must be capable of synchronizing to clock with accuracy of $+/-4.6 \times 10^{-6}$	1×10^{-8}/day	3.5 Hours
3	4.6×10^{-6}	Must be capable of synchronizing to clock with accuracy of $+/-4.6 \times 10^{-6}$	3.7×10^{-7}/day	6 Minutes (255 in 24 Hrs)
4E	32×10^{-6}	Must be capable of synchronizing to clock with accuracy of $+/-32 \times 10^{-6}$	Same as Accuracy	Not Yet Specified
4	32×10^{-6}	Must be capable of synchronizing to clock with accuracy of $+/-32 \times 10^{-6}$	Same as Accuracy	N/A

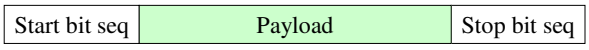
Framing



- Need to know where a frame starts
  - Special bit sequence marks start of frame
- Need to know where frame ends
  - Special bit sequence, or
  - Length of frame is transmitted, or
  - Fixed length frame

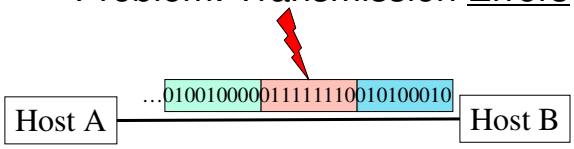
Framing (cont.)

- The generic view



- Because the payload may contain the start or stop sequence, may have to "stuff" payload at sender, and unstuff at receiver
  - Something like putting a quote inside a quoted string in a programming language
    - Suppose start bit sequence is 0x7E.
    - Replace 0x7E in payload with 0x7D 0x5E
    - Replace 0x7D in payload with 0x7D 0x5D
    - At receiver, 0x7D 0x5E replaced with 0x7E
- We'll see more frame formats when we look at specific link level protocols in a bit…

Problem: Transmission Errors



- Noise can flip some of the bits we receive
  - We must be able to detect when this occurs!
- Basic approach: add redundant data
  - Error detection codes allow errors to be recognized
  - Error correction codes allow (some) errors to be repaired too
- Questions we'll delay for a bit:
  - What should happen if an uncorrectable error is detected?
  - Which layer(s) should do whatever it is?

Patterns of Errors

- Q: Suppose you expect a bit error rate of about 1 bit per 1000 sent. What fraction of packets would be corrupted if they were 1000 bits long (and you could detect all errors but correct none)?
- A: It depends on the pattern of errors
  - Bit errors occur at random
    - Packet error rate is about $1 - 0.999^{1000} = 63\%$
  - Errors occur in bursts, e.g., 100 consecutive bits every 100,000 bits
    - Packet error rate ≤ 2%

Detection/Correction Codes: Redundancy

- A scheme maps D bits of data into D+R bits – i.e., it uses only 2^D distinct bit strings of the 2^{D+R} possible.

```
ECC = f(data)
```

- The sender computes the ECC bits based on the data, and sends them along with the data.
- The receiver takes the data bits it received and computes the ECC bits for them. It then computes the ECC it computed with the one it received.
  - Detection occurs when what the receiver computed and received don't match
    - Note: corruption of ECC bits triggers detection
  - That is, *detection occurs when the D+R total bits are not one of the 2^D messages valid using the code*

Detection/Correction Codes: Redundancy

- Detection/correction schemes are characterized in two ways:
  - Overhead: ratio of total bits sent to data bits, minus 1
    - Example: 1000 data bits + 100 code bits = 10% overhead
  - The errors they detect/correct
    - E.g., all single-bit errors, all bursts of fewer than 3 bits, etc.

The Hamming Distance

- The Hamming distance of a code is the smallest number of bit differences that turn any one codeword into another
  - e.g, code 000 for 0, 111 for 1, Hamming distance is 3
- For code with distance d+1:
  - d bit errors can be detected, e.g, 001, 010, 110, 101, 011
- For code with distance 2d+1:
  - d errors can be corrected, e.g., 001 000

Specific Schemes

- We'll briefly touch on the three schemes mentioned in the book
  - They're organized from least to most expensive to compute
- Scheme 1: parity
  - A single parity bit is associated with each K bits of the data, for some K. It is set so that the XOR of the data bits + the parity bit = 0 (for even parity)
    - Example: K=8, one parity bit per byte
      - Detects all odd numbers of errored bits
    - Example: 2-dimensional parity: one parity bit for each bit in a byte, another for each of the eight bit positions in 8 consecutive bytes
      - Detects all 1-, 2-, and 3- bit errors, plus many >3-bit errors

2-d parity example

0101001	1
1101001	0
1011110	1
0001110	1
0110100	1
1011111	0
1111011	0

Specific Schemes

- Scheme 2: checksum
  - General idea: Sum successive blocks of K-bits of the data, as though they were integers
  - Internet checksum: K=16, use 1's-complement arithmetic, take 1's complement of result as checksum
    - Example: data is 01 00 F2 03 F4 F5 F6 F7
      - 0100 + F203 = [0] F303
      - F303 + F4F5 = [1] E7F8 = E7F9
      - E7F9 + F6F7 = [1] DEF0 = DEF1
      - Checksum is 1's complement of DEF1: 210E
      - Transmit 01 00 F2 03 F4 F5 F6 F7 21 0E
    - Why use 1's-complement is a bit arcane (e.g., endian-ness of machine doesn't matter), and not terribly crucial
  - Note: overhead (in bits) of this scheme is independent of the number of data bits sent

Specific Schemes

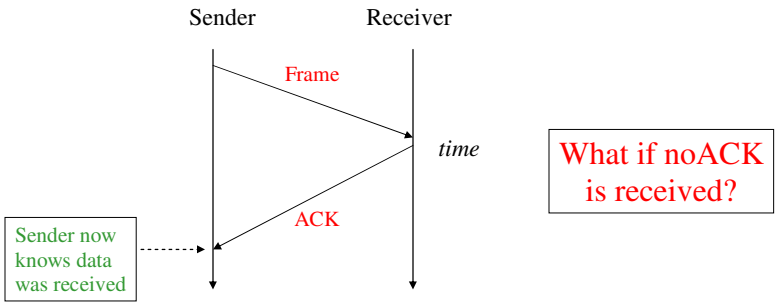
- Scheme 3: CRCs (Cyclic Redundancy Check)
  - Stronger protection than checksums
    - Used widely in practice, e.g., Ethernet CRC-32
    - Implemented in hardware (XORs and shifts)
- Based on mathematics of finite fields
  - "numbers" correspond to polynomials, use modulo arithmetic
  - e.g, interpret 10011010 as $x^7 + x^4 + x^3 + x^1$
- Algorithm: Given n bits of data, generate a k bit check sequence that gives a combined n + k bits that are divisible by a chosen divisor C(x)

How is C(x) Chosen?

- Mathematical properties:
  - All 1-bit errors if non-zero x^k and x^0 terms
  - All 2-bit errors if C(x) has a factor with at least three terms
  - Any odd number of errors if C(x) has (x + 1) as a factor
  - Any burst error < k bits
- There are standardized polynomials of different degree that are known to catch many errors
  - Ethernet CRC-32: 10000010011000001000111011011011 1

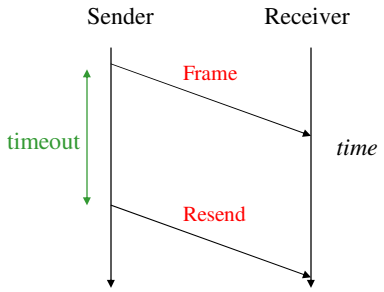
Reliable Transmission

- Because there may be uncorrectable errors (no matter what ECC scheme is used), how can the sender be sure that the receiver got the data?
  - The sender must receive an acknowledgement (ACK) from the sender



Timeouts / Automatic Repeat Request (ARQ)

- If no ACK comes back, the sender must re-send the data (ARQ)
  - When is the sender sure that no ACK is coming back?
    - Because as a practical matter delays are very difficult to bound, in some sense it can never be sure
    - Sender chooses some reasonable timeout – if the ACK isn't back in that much time, it assumes it will never see an ACK, and re-sends

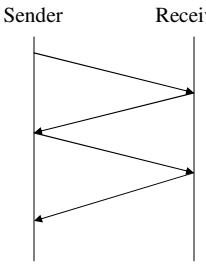


What if original frame arrived, but ACK was lost?

Duplicate Detection: Sequence Numbers

- So that the receiver can detect (and discard) duplicates, distinct frames are given distinct sequence numbers
  - E.g., 0, 1, 2, 3, …
- When a frame is re-sent, it is re-sent with the same sequence number as the original
- The receiver keeps some information about what sequence numbers it has seen, and discards arriving packets that are duplicates

Stop-and-Wait Protocol

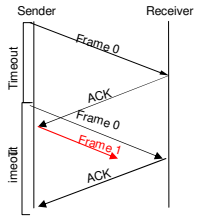


Here's what it looks like when things are going well (no transmission errors).

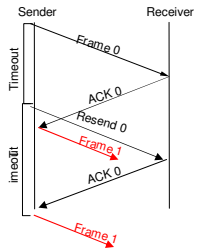
- Sender doesn't send next packet until he's sure receiver has last packet
- The packet/ACK sequence enables reliability
- Sequence numbers help avoid problem of duplicate packets

Stop & wait sequence numbers

- Sequence numbers enable the receiver to discard duplicates
- ACKs must carry sequence number info as well



The Problem Scenario



The Solution

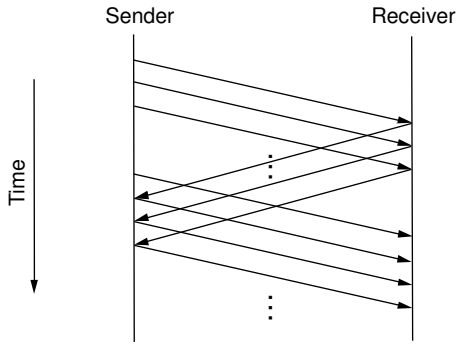
- Stop & wait allows one outstanding frame, requires two distinct sequence numbers

Problem with Stop-And-Wait: Performance

- Problem: “keeping the pipe full”
  - If the bandwidth-delay product is much larger than a packet size, the sender will be unable to keep the link busy
- Example
  - 1.5Mbps link x 45ms RTT = 67.5Kb (8KB)
  - 1KB frames imples 1/8th link utilization
- Solution: allow multiple frames “in flight”

Solution: Allow Multiple Frames in Flight

- This is a form of pipelining

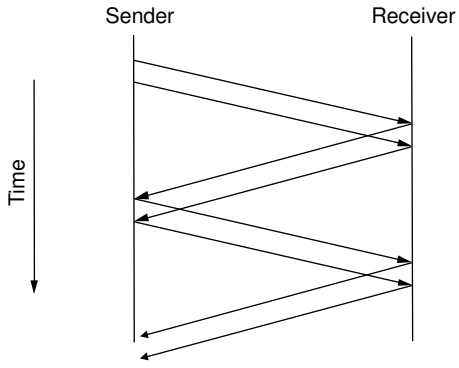


Flow Control

- Why can't we allow the sender to send as fast as it can, timing out and re-sending each frame as necessary?
- Flow control:
  - Receiver needs to buffer data until it can be delivered to higher layers
    - If the sender is much faster than the receiver, it will overwhelm it, causing the receiver to run out of buffer space
  - Additionally, if a frame is lost, the receiver will receive frames "out of order". It wants to buffer those frames to avoid retransmission, but cannot deliver them to the client until the missing frame is re-sent and received
  - Finally, sender needs to buffer frames in case it has to resend them
- Flow control is the notion that the sender must limit the rate at which it transmits to something below the raw bandwidth of the link
- A common, important approach to flow control is the *sliding window protocol*

Sliding Window Protocol

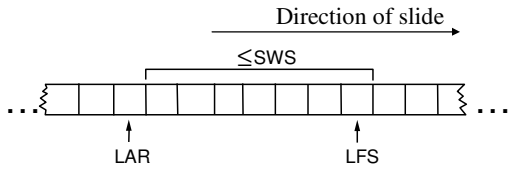
- There is some maximum number of un-ACK'ed frames the sender is allowed to have in flight
  - We call this "the window size"
  - Example: window size = 2



Once the window is full, each ACK'ed frame allows the sender to send one more frame

Sliding Window: Sender

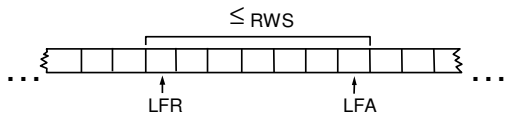
- Assign sequence number to each frame (**SeqNum**)
- Maintain three state variables:
  - send window size (**SWS**)
  - last acknowledgment received (**LAR**)
  - last frame sent (**LFS**)
- Maintain invariant: **LFS** - **LAR** <= **SWS**



- Advance **LAR** when ACK arrives
- Buffer up to **SWS** frames

Sliding Window: Receiver

- Maintain three state variables
  - receive window size (**RWS**)
  - largest frame sequence number acceptable (**LFA**)
  - last frame received (**LFR**)
- Maintain invariant: **LFA** - **LFR** <= **RWS**

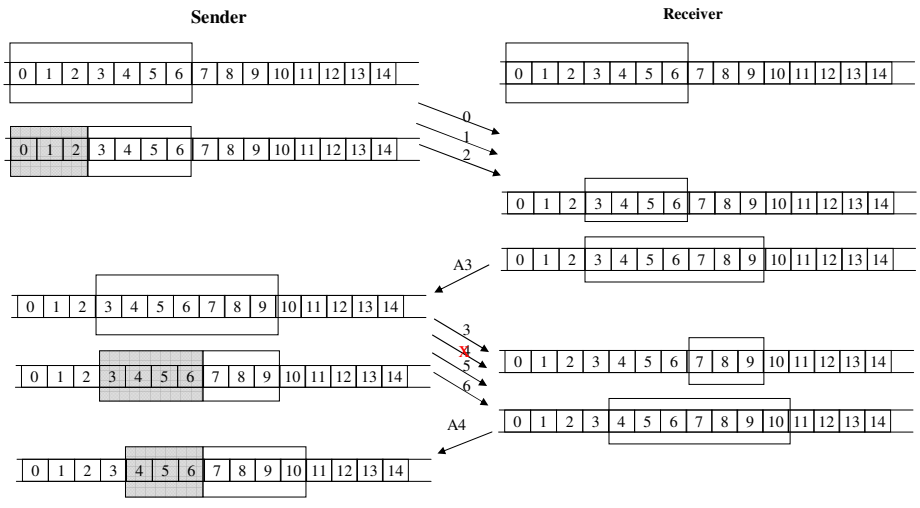


- Frame **SeqNum** arrives:
  - if **LFR < SeqNum ≤ LFA** ⇒ accept + send ACK
  - if **SeqNum ≤ LFR** or **SeqNum > LFA** ⇒ discard

ACKs

- Send *cumulative* ACKs
  - send ACK for largest frame such that all frames less than this have been received
  - Why?
- Send an ACK each time a packet with SeqNum in the window arrives
  - even if you've seen that packet already
  - Why?

Sliding Window Example



Sequence Number Space

- **`SeqNum`** field is finite; sequence numbers wrap around
- Sequence number space must be larger then number of outstanding frames
- **`SWS <= MaxSeqNum-1`** is not sufficient
  - suppose 3-bit **`SeqNum`** field (0..7)
  - **`SWS=RWS=7`**
  - sender transmit frames 0..6
  - arrive successfully, but ACKs lost
  - sender retransmits 0..6
  - receiver expecting 7, 0..5, but receives the original incarnation of 0..5
- **`SWS < (MaxSeqNum+1)/2`** is correct rule
- Intuitively, **`SeqNum`** "slides" between two halves of sequence number space

Sliding Window Summary

- Sliding window is best known algorithm in networking
- First role is to enable reliable delivery of packets
  - Timeouts and acknowledgements
- Second role is to enable in order delivery of packets
  - Receiver doesn't pass data up to app until it has packets in order
- Third role is to enable flow control
  - Prevents server from overflowing receiver's buffer