

Internet Protocol Stack

application: supporting network applications

- HTTP, SMTP, FTP, etc.

transport: endhost-endhost data transfer

- TCP, UDP

network: routing of datagrams from source to destination

- IP, routing protocols

link: data transfer between neighboring network elements

- Ethernet, WiFi

physical: bits “on the wire”



1

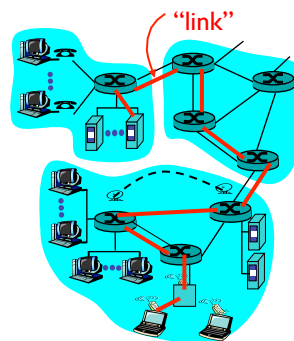
Data Link Layer

The data-link layer has the responsibility of transferring packets from one node to an adjacent node over a link

At the link layer, a packet is called a **frame**, and it encapsulates a network-layer datagram

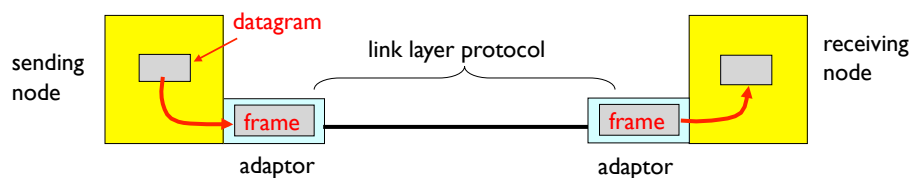
A network datagram may be transferred by different link protocols over different links:

- e.g., Ethernet on the first link, frame relay on intermediate links, and 802.11 on the last link



2

Adaptors Communicating



Link layer implemented in
“adaptor” (a.k.a. NIC)

- Ethernet card, PCMCIA card, 802.11 card

Sending side:

- encapsulates datagram in a frame
- adds error checking bits, flow control, etc.

Receiving side

- looks for errors, flow control, etc
- extracts datagram, passes to receiving node

Adaptor is semi-autonomous
link & physical layers

3

Ethernet

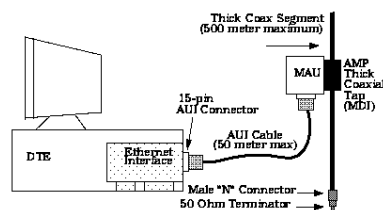
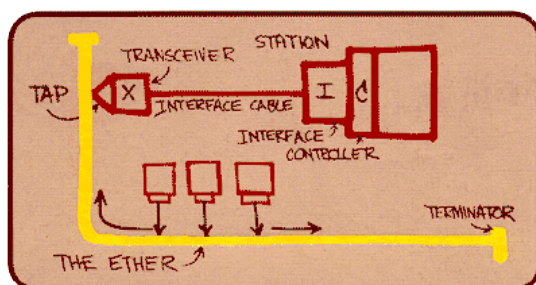
“Dominant” wired LAN technology:

cheap \$20 for 100Mbps!

First widely used LAN technology

Simpler, cheaper than token LANs and ATM

Kept up with speed race: 10 Mbps – 10 Gbps



Metcalfe's
Ethernet sketch

4

Data Link Layer

The Data Link layer can be further subdivided into:

1. Logical Link Control (LLC): error and flow control
2. Media Access Control (MAC): framing and media access

different link protocols may provide different services, e.g., Ethernet doesn't provide reliable delivery (error recovery)

MAC topics:

- framing and MAC address assignment
- LAN forwarding
- IP to MAC address resolution
 - IP to MAC: Address Resolution Protocol (ARP)
 - MAC to IP: Reverse ARP (RARP), BOOTstrap Protocol (BOOTP), Dynamic Host Configuration Protocol (DHCP)
- media access control



5

Framing

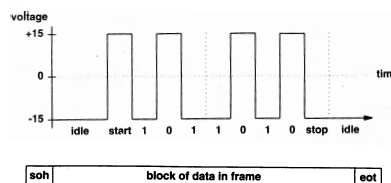
Why packetize/frame data?

- minimize retransmission (upon error)
- resource sharing, example:
 - 5 MB file takes 12 min to transmit on a 56 kbps line
 - 1 KB packet takes 143 ms

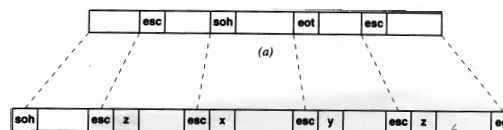
Framing allows sources with small amount of data to send (e.g., VoIP) to finish promptly

Framing is done by using a special bit pattern to denote start & end of frame (soh & eot)

Bit stuffing: if soh & eot shows up in data, they must be protected/escaped



Character In Data	Characters Sent
soh	esc x
eot	esc y
esc	esc z



Frame Transmission and MAC Addresses

Frame transmission on a LAN:

- frames are tagged with destination MAC address
- frames sent to *all hosts* on the LAN
- the NIC on each host makes a copy of frame
- if the frame is addressed to the host, the NIC sends the frame up to the CPU
- a frame can also have a broadcast or multicast address
- NICs could be put in promiscuous mode (e.g., tcpdump, ethereal, network sniffer, network analyser)

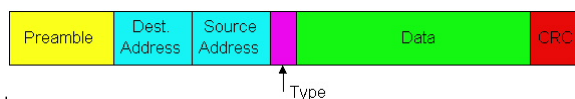
MAC address assignment

- static: Ethernet (48-bits): requires global address assignment
- configurable: requires DIP switch, EPROM
- dynamic (random number):
 - advantage: only need to be unique within a LAN
 - disadvantage: address changes between reboots

7

Ethernet Frame Structure

Sending adaptor encapsulates IP datagram (or other network layer protocol packet) in Ethernet frame



Preamble: 7 bytes with pattern 10101010 followed by one byte with pattern 10101011, used to synchronize receiver-sender clock rates

Addresses: 6 bytes

- if adaptor receives frame with matching destination MAC address, or with broadcast address (e.g., ARP packet), it passes data in the frame to network-layer protocol
- otherwise, adaptor discards frame

Type: indicates the higher layer protocol

CRC: checked at receiver, if error is detected, the frame is simply dropped

8

Address Resolution

IP routing on a LAN: assume hosts know their own network number and subnet mask:

- send directly to the destination if it is determined to be on the same LAN
- send to a default router otherwise

either case, the host must know the MAC address of the destination or the default router

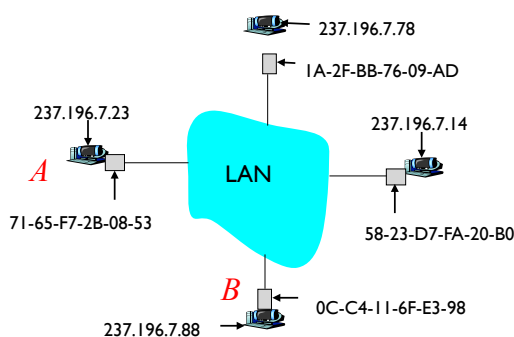
Given a node's IP address, how can a host determine its MAC address?

- MAC address can be inferred from the IP address (IPv6)
- from a statically configured table
- ask a server
- use the Address Resolution Protocol (ARP)

9

Address Resolution Protocol (ARP)

Question: how would *A* find out *B*'s MAC address, knowing *B*'s IP address?



Each IP node (host, router) on the LAN has an ARP table

ARP Table: IP/MAC address mappings for some LAN nodes

< IP address; MAC address; ttl >

- ttl (time to live): time after which, address mapping will be flushed (typically 20 min)
- ARP table is maintained in an LRU manner

10

ARP Protocol: Same LAN

A wants to send datagram to *B*, and *B*'s MAC address not in *A*'s ARP table

A broadcasts ARP query packet, containing *B*'s IP address

- dest MAC address = FF-FF-FF-FF-FF-FF
- all machines on LAN receive ARP query
- query packet also contains *A*'s own IP and MAC addresses

B receives ARP packet, replies to *A* with its (*B*'s) IP and MAC addresses

- frame sent to *A*'s MAC address (unicast)
- *B* caches (saves) *A*'s IP to MAC address mapping in its own ARP table, or refreshes *A*'s entry if it already exists

A caches *B*'s IP-to-MAC address pair in its ARP table until ttl expires, at which time it will be flushed

- soft state: information that times out (goes away) unless refreshed

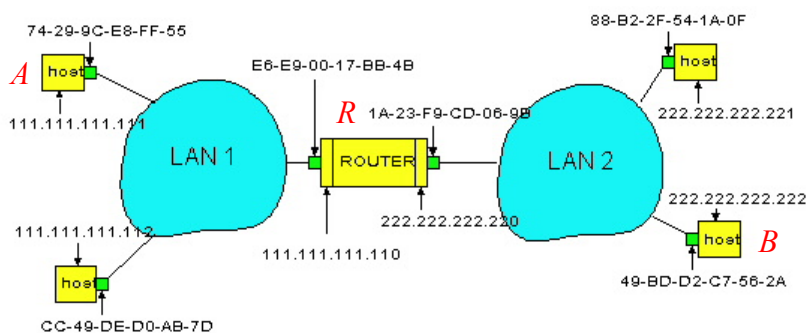
ARP is “plug-and-play”:

- nodes create their ARP tables without intervention from net administrator
- Try out `arp(8)` (may need root/administrator permission)

11

Forwarding to Another LAN

Want: send datagram from *A* to *B* via *R*, assume *A* knows *B*'s IP address

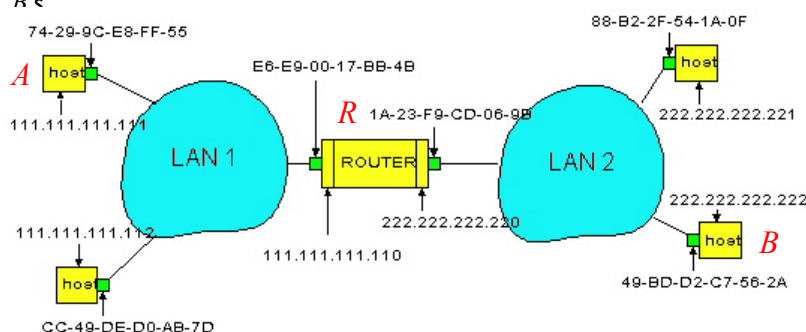


- Router *R* has two ARP tables: one for each LAN
- *A* knows that its default router (*R*) has IP address 111.111.111.110
- *A* looks up *R*'s MAC address E6-E9-00-17-BB-4B from its ARP table, or if the mapping doesn't exist, it sends out an ARP request packet to resolve it

12

Forwarding to Another LAN

- *A* creates datagram with source IP *A*, destination IP *B*
- *A* creates link-layer frame with *R*'s MAC address as dest, frame containing *A*-to-*B* IP datagram
- *A*'s adaptor sends frame to *R*
- *R*'s adaptor receives frame, extracts IP datagram from the frame, sees that its destination is *B*
- *R* uses ARP to get *B*'s MAC address, and creates a new frame containing *A*-to-*B* IP datagram with MAC destination address set to *R*'s



13

Obtaining an IP Address

How does a host obtain its IP address?

1. Hard-coded by system admin in a file
 - Wintel: control-panel->network->configuration->tcp/ip->properties
 - UNIX: /etc/rc.config
2. Ask a server:
 - Reverse ARP (RARP) (obsolete)
 - BOOT Protocol (BOOTP) (obsolete)
 - Dynamic Host Configuration Protocol (DHCP):
 - dynamically request an address from a server when the host boot
 - "plug-and-play"

14

RARP

- Sender broadcasts a RARP packet with its own MAC address
- One or more RARP server respond with the sender's IP address
- If no reply, server may be down or busy, retry later

To prevent too many replies:

- each host can be assigned a primary server
- on repeated query, non-primary servers wait a random time for response from other servers before replying

RARP may also be used to find out the IP address of a 3rd party host

Disadvantages of RARP:

- can't be used with dynamic MAC addresses
 - limited information sent
 - limited to physical segment
 - requires one RARP server per segment
- (last two due to the use of broadcasting)

15

Need for a More General Bootstrap Protocol

Information a newly booted machine may need:

- IP address
- subnet mask
- default router's address
- boot file (name and size)
- time of day
- DNS server
- print server
- file server (if thin/diskless client)
- etc.

Each piece of the required info can be requested and sent separately

However, this would be inefficient because:

- it causes a lot of broadcast traffic
- each request/reply must be padded out to min frame size (due to MAC layer broadcast requirement)

16

BOOTP and DHCP

BOOTP

- batched query and response
- uses UDP/IP with IP broadcasting
 - limiting BOOTP use within a LAN, but beyond a physical segment
- hosts' IP addresses are assigned statically
 - requires database update for each new host

DHCP: BOOTP with a pool of shared host identities

- if MAC address of a querying host is not in the database of permanent identities, assigns (leases) it a temporary identity from pool
- clients wait a random time before sending Discover or Request messages after booting, to prevent storming the LAN
- advantage: doesn't require manual configuration
- shortcoming: DHCP's interaction with DNS unspecified (dynamic DNS not widely deployed)

17

DHCP (and BOOTP) Packet Format

Opcode:

- BOOTREQUEST
- BOOTREPLY

Hardware Type:

- Ethernet (1), FireWire (24), etc.

Hardware address length (hlen)

Transaction ID:

- random number chosen by client to associate messages and responses

Seconds Elapsed: since client began an address acquisition or renewal process

0	8	16	24	31
OP	HTYPE	HLEN	HOPS	
TRANSACTION IDENTIFIER				
SECONDS ELAPSED		FLAGS		
CLIENT IP ADDRESS				
YOUR IP ADDRESS				
SERVER IP ADDRESS				
ROUTER IP ADDRESS				
CLIENT HARDWARE ADDRESS (16 OCTETS)				
⋮				
SERVER HOST NAME (64 OCTETS)				
⋮				
BOOT FILE NAME (128 OCTETS)				
⋮				
OPTIONS (VARIABLE)				

18

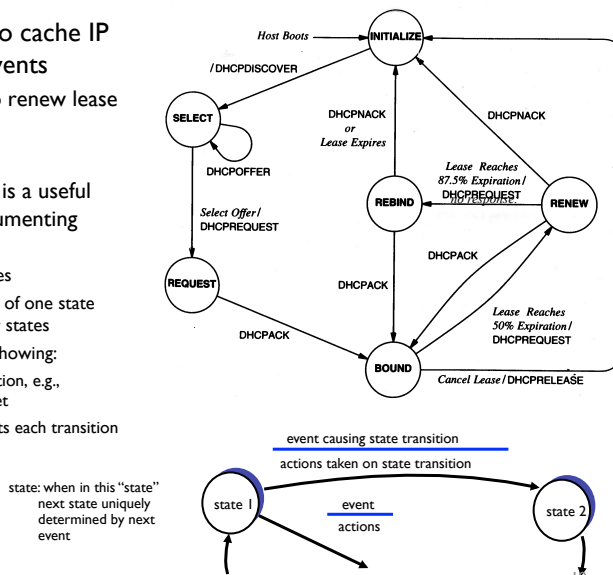
DHCP Simplified Finite State Machine

Also allows for client to cache IP address across boot events

- upon boot, client tries to renew lease of cached address

Finite state machine (FSM) is a useful tool for designing and documenting protocol:

- it consists of a number of states
- a graph showing the transition of one state into one or more of the other states
- and labels on the graph edge showing:
 - what event causes each transition, e.g., receiving certain type of packet
 - and what actions or side effects each transition may cause, if any



Multiple Access Problem

Broadcast channel of rate R bps, shared medium

- if two users send at the same time, *collision* results in no packet being received (interference)
- if no users send, channel goes idle
- thus, want to have only one user send at a time

Media Access Control:

- determines who gets to send next
- what to do if more than one hosts send at the same time and there's collision

Duplex mode:

- half duplex: only one end can send at a time
- full duplex: both ends can send simultaneously

Ideal Multiple Access Protocol

- when one node wants to transmit, it can send at rate R
- when M nodes want to transmit, each can send at average rate R/M
- fully decentralized:
 - no special node to coordinate transmissions
 - no synchronization of clocks, slots
 - distributed algorithm that determines how nodes share channel, i.e., determine when node can transmit
 - communication about channel sharing must use channel itself!
 - no out-of-band channel for coordination

21

Categorization of MAC Protocols

1. Random access:
 - Slotted ALOHA
 - ALOHA
 - Carrier Sense Multiple Access with Collision Detection (CSMA/CD)
 - CSMA/CAvoidance
2. Token passing
3. Channel partitioning: TDMA, FDMA, CDMA

Standards:

- 802.3 (CSMA/CD), 802.3a? (GigE)
- 802.4 (token bus)
- 802.5 (token ring)
- 802.11 [bagn] (WiFi)

22

Random Access MAC Protocol

Characteristics:

- sender xmits bits on the wire at full channel rate R bps
 - no prior coordination among nodes
- bits are propagated along the entire network
- destination recognizes that frame is for itself
- destination grabs frame
- while one host is xmitting, all others must wait

Random access means:

- relies on collision to control access
- how to detect collisions
- how to recover from collisions

23

Ethernet: CSMA/CD

Carrier Sensing:

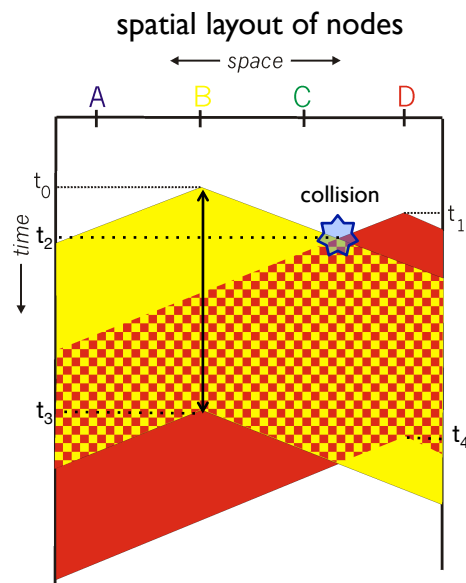
1. check for presence of electrical signal (carrier) on wire before transmission
2. presence of carrier means someone else is sending, wait
3. start transmission if no carrier detected

Problem: collision

24

CSMA Collisions

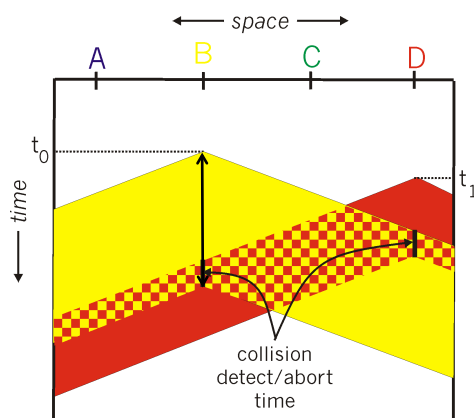
- collisions occur because propagation delay means two nodes may not hear of each other's transmission when they start transmitting (A at t_0 , D at t_1)
- when collision occurs (at t_2), entire frame transmission time ($t_3 - t_0$ or, equivalently, $t_4 - t_1$) is wasted
- note the role distance & propagation delay play in determining collision probability
- a collision is detected if power received is larger than power transmitted



25

Collision Detection

- sender must continue to detect collision after transmission
 - on collision, frames must be retransmitted
 - problem: more collision
4. if adaptor detects collision while transmitting, aborts and sends **jam signal**
 5. after aborting, adaptor enters **exponential backoff**



26

Jam Signal and Exponential Back-off

Jam signal: make sure all other transmitters are aware of collision; 48 bits

Exponential back-off: senders pick a uniformly distributed random delay between $[0, 2^0d]$ before retransmission. Why random?

If collision occurs again, pick another random delay between $[0, 2^1d]$, $[0, 2^2d]$, $[0, 2^3d]$, . . . hence (*binary*) *exponential back-off*

Bit time: .1 μ sec on a 10 Mbps Ethernet

→ for $2^{10}d$, wait time is about 50d msec

27

CSMA/CD Summary

The algorithm:

1. listen for carrier
2. if no carrier, send frame
3. listen for collision or jamming signal
4. if collision detected, send jamming signal
5. if collision or jamming signal detected, retransmit after exponential back-off

Historical Note:

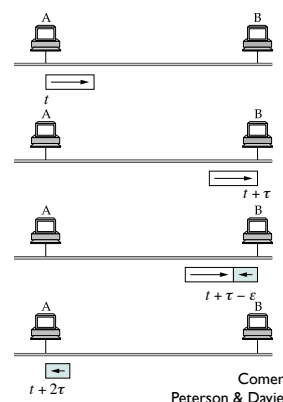
Collision detection and retransmission with back-off was first used in the ALOHA MAC algorithm from the University of Hawaii (1970) for access to satellite channels

28

Collision Detection Time

How long must a sender listen for collision?

- let τ be the propagation time from one end of the wire to the other
- within τ time after the transmission of a frame (t), all nodes on the segment would have sensed carrier
- worst case scenario for collision: a node at the other end of the wire starts transmitting at time $t + \tau - \epsilon$
- the node closest to the collision sends out a jamming signal to ensure collision is detected by the other node
- it takes another τ period for the collision to get back to the original sender



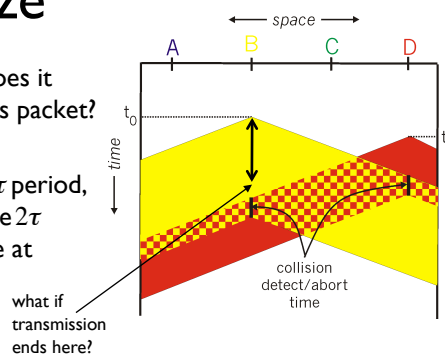
Hence the original sender must listen for 2τ period

29

Minimum Frame Size

When a sender detects collision how does it know that the collision was caused by its packet?

Answer: sender must hold carrier for 2τ period, i.e., it must be transmitting for the whole 2τ period $\Rightarrow$ each Ethernet frame must be at least $2\tau \times \text{link speed}$ long



Example:

- 10 Mbps Ethernet allows maximum of 5 segments, each 500 m long
- speed of light $3 \times 10^8 \text{ m/s}$, but coax propagation $2 \times 10^8 \text{ m/s}$
- round-trip propagation delay (2τ) on 2.5 km coax is 25 μs
- allowing for 4 repeaters makes end-to-end delay 50 μs
- 50 μs means 62.5 bytes
- 802.3 standard requires stations to hold carrier for $64 \text{ bytes} / 10 \text{ Mbps} = 51.2 \mu\text{s}$

30

CSMA/CD Efficiency (η)

t_{prop} = max propagation time between 2 nodes in the LAN

t_{trans} = time to transmit maximum-size frame

$$\eta = \frac{t_{trans}}{t_{trans} + 5t_{prop}} = \frac{1}{1 + 5t_{prop} / t_{trans}}$$

$\eta \rightarrow 1$ as $t_{prop} \rightarrow 0$ or as $t_{trans} \rightarrow \infty$

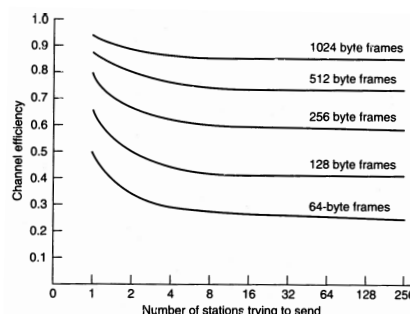
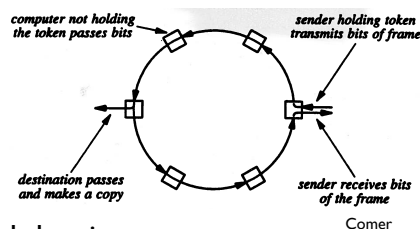


Fig. 4-23. Efficiency of 802.3 at 10 Mbps with 512-bit min frame size
Tanenbaum

31

Token Ring MAC Protocol

- a **token** goes around a ring network
- to send data, a node must first grab the token
- a frame sent from a source is passed from node to node around the ring
- destination recognizes own address and makes a copy of frame
- sender removes frame from ring
- each node can only transmit one frame at a time; must return token to the ring after each frame transmission



Why let the sender, instead of the receiver, remove frame from the ring?

32

Token Ring MAC Protocol

Token:

- a special bit pattern
- use bit-stuffing if data resembles token
- only one token on ring at a time (managed by a monitor)

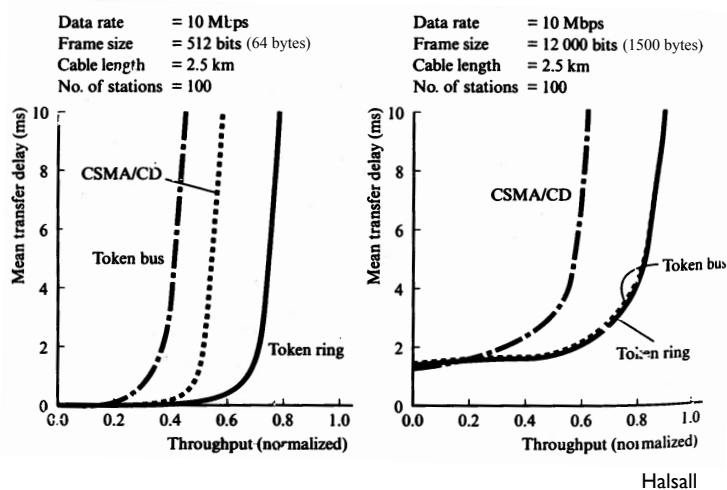
IBM's token ring link speed is 16 Mbps

Token ring:

- advantage: no collision
- disadvantage: failure of a node or link disables the whole network

33

Token Ring Performance



34

CSMA/CD Efficiency (η)

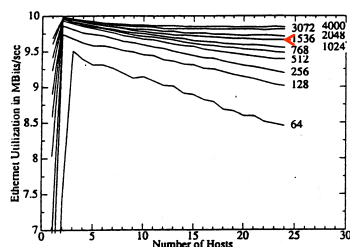


Figure 3-3: Total bit rate

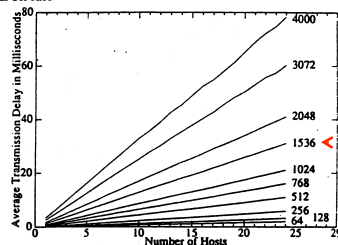


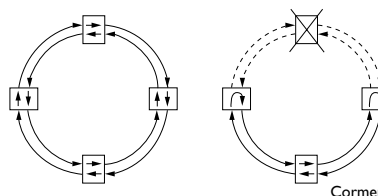
Figure 3-7: Average transmission delay

35

Other MAC Protocols

FDDI:

- operates at 100 Mbps
- uses the token ring MAC protocol
- for robustness, uses two counter-rotating rings
- if a link/node goes down, the dual-ring can be reconfigured to a single ring network (hence called **self-healing** network)



SLIP/PPP: serial line, point-to-point protocol, no need for media access control, just framing

ATM/Frame Relay/SONET: for backbone links

36