15-441 Computer Networks

Lecture 6

Link-Layer (2)

Dave Eckhardt

Roadmap

- **What's a link layer?**
- **Ethernet**
- **Things which aren't Ethernet**
  - Token Bus, Token Ring, FDDI, Frame Relay
  - 802.11
  - PPP, DSL, cable modems
- **A word on approach**
  - We will discuss many "obsolete" technologies
  - This can be a good way to grasp the underlying ideas
    - ...which keep turning up in different contexts
    - A good arrangement of ideas is an easier advance than a genuinely new thing

Reminder: Medium Access Control (MAC)

- **Share a communication medium among multiple**
- **Arbitrate between connected hosts**
- **Goals:**
  - High resource utilization
  - Avoid starvation
  - Simplicity (non-decentralized algorithms)
- **Approaches**
  - Taking turns, random access, really-random access (SS)
  - Random access = allow collisions
    - Manage & recover from them

Outline

- **Ethernet**
  - Conceptual history
  - Carrier sense, Collision detection
  - Ethernet history, operation (CSMA/CD)
  - Packet size
  - Ethernet evolution
  - Connecting Ethernets
- **Not Ethernet**
  - FDDI, wireless, ...

Hui Zhang, Dave Eckhardt

Ethernet in Context

- **ALOHA**
  - When you're ready, transmit
  - Detect collisions by waiting (a long time)
  - Recover from collision by trying again
    - – ...after a random delay...
      - » Too short, entire network collapses
      - » Too long, every user gets bored
- **Things to try**
  - Slotted ALOHA – reduce collisions (some, not enough)
  - Listen before transmit
  - True collision detection

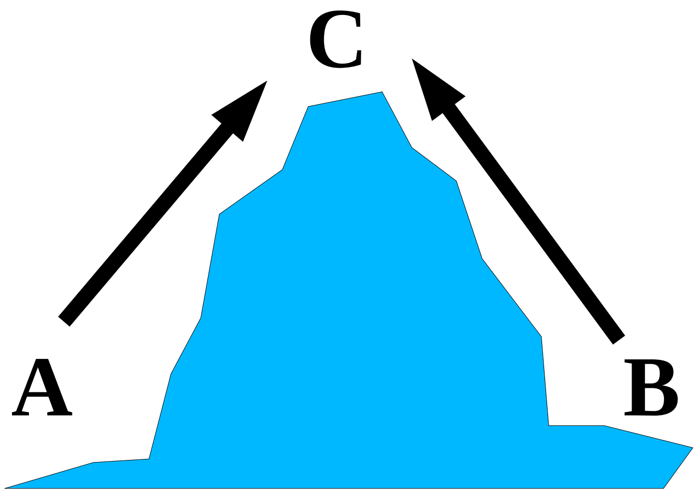
Hui Zhang, Dave Eckhardt

Listen Before Transmit

- **Basic idea**
  - Detect, avoid collisions – *before they happen*
  - Listen before transmit (officical name: "Carrier Sense")
    - – Don't start while anybody else is already going
- **Why didn't ALOHA do this?**
  - "Hidden terminal problem"

Hidden Terminal Problem

- **A and B are deaf to each other**
  - Can't sense each other's carrier
  - Carrier sense "needs help" in this kind of environment
- **But CS can work really well in an enclosed environment (wire)**



Collision Detection

- **Is Carrier sense enough?**
  - Sometimes there is a “race condition”
    - Two stations listen at the same time
    - Both hear nothing, start to transmit
    - Result: collision
      - » Could last “for a while”
      - » Can we detect it while it's happening?
- **Collision Detection**
  - Listen while you transmit
  - If your signal is “messed up”, assume it's due to a collision
- **Great idea! Why didn't ALOHA do it?**

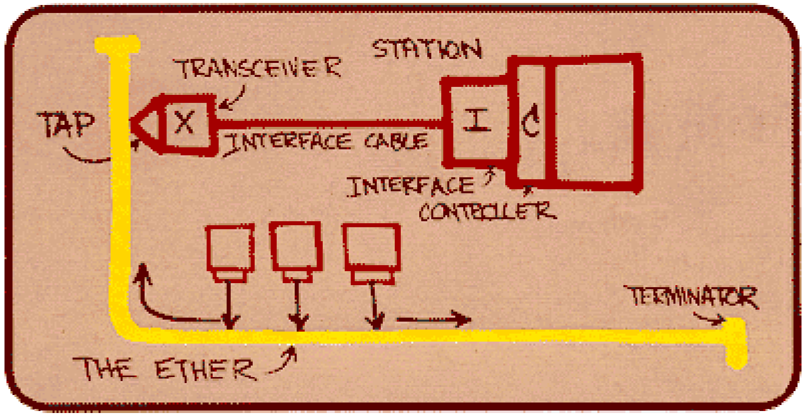
Hui Zhang, Dave Eckhardt

Collision Detection

- **Collision detection difficult for radios**
  - “Inverse-square law” relates power to distance
    - At A, A's transmission drowns out B's
    - At B, B's transmission drowns out A's
    - Neither can hear each other, C hears mixture (collision)
  - Many radios disable receiver while transmitting
    - Huge power of local transmitter may damage receiver
- **Collision detection *can* be done inside a wire**

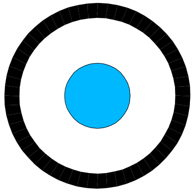
Hui Zhang, Dave Eckhardt

Original Xerox PARC Ethernet Design



www.ethermanage.com/ethernet

Coaxial cable



- **Medium – one long cable snaked through your building**
- **Transceiver – fancy radio with collision detection**
- **“Vampire tap”**
  - Drill hole into cable (carefully!)
  - Insert pin to touch center connector (carefully!!)

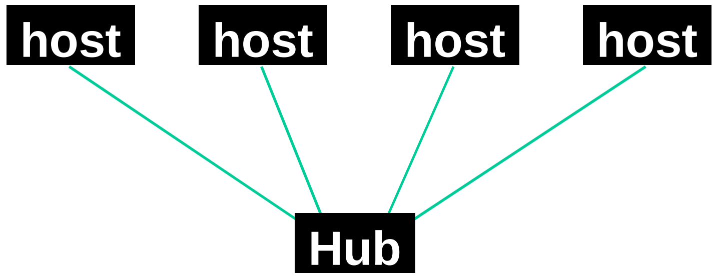
Hui Zhang, Dave Eckhardt

Original Xerox PARC Ethernet Design

- **Carrier-sense multiple access with collision detection (CSMA/CD).**
  - MA = multiple access
  - CS = carrier sense
  - CD = collision detection
- **PARC Ethernet parameters**
  - 3 Mb/s (to match Xerox Alto workstation RAM throughput)
  - 256 stations (1-byte destination, source addresses)
  - 1 kilometer of cable

802.3 Ethernet

Broadcast technology



- **DEC/Intel/Xerox ("DIX") Ethernet standardized by IEEE**
  - 3 Mb/s ⇒ 10 Mb/s
  - Station addresses 1 byte ⇒ 6 bytes
- **Growth over the years**
  - Hubs, bridges, switches
  - 100Mbps, 1Gbps, 10Gbps
  - Thin coax, twisted pair, fiber, wireless

CSMA/CD Algorithm

- **Listen for carrier**
- **If carrier sensed, wait until carrier ends.**
  - Sending would force a collision and waste time
- **Send packet and listen for collision.**
- **If no collision detected, consider packet delivered.**
- **Otherwise**
  - Abort immediately
    - Transmit "jam signal" (32 bits) to fill cable with errors
  - Perform “exponential back-off” to try packet again.

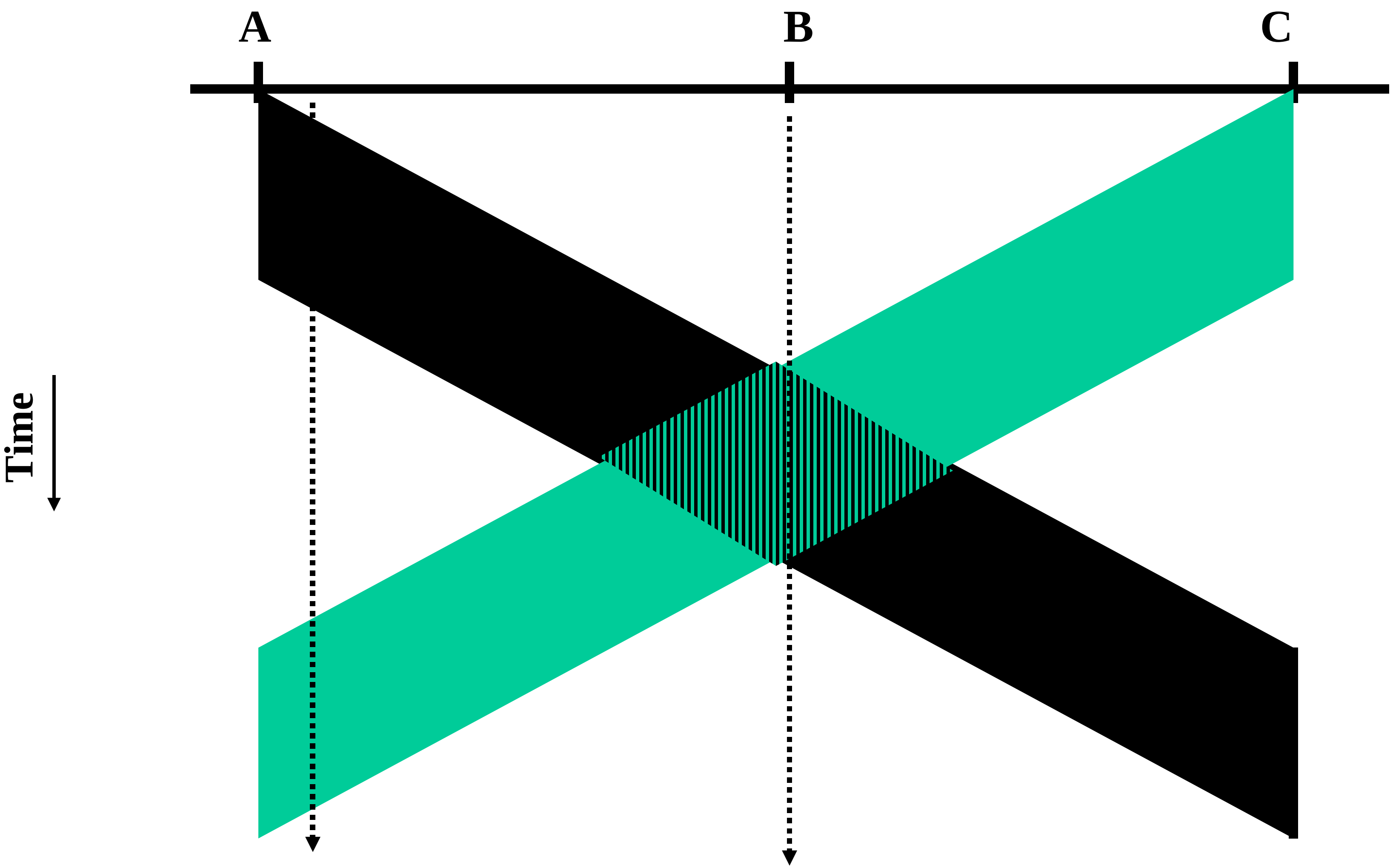
Exponential Back-off

- **Basic idea**
  - Choose a random interval (e.g., 8 small-frame times)
    - – Delay that long, try again
  - How long should interval be?
    - – ...roughly 1 time per station contending for medium...
    - – ...can't tell, must guess

Exponential Back-off

- **Exponential Back-off**
  - First collision: delay 0 or 1 periods (512 bits)
    - 50/50% probability
    - Appropriate if two stations contending for medium
  - Second collision: delay 0...3 periods
    - Will work well if "roughly 4" stations contending
  - Third collision: delay 0...7 times
  - Ten collisions?
    - Give up, tell device driver "transmission failed"

Collision Detection



Hui Zhang, Dave Eckhardt

Collision Detection: Implications

Goal: every node detects collision as it's happending

Any node can be sender

So: need short wires, or long packets.

Or a combination of both

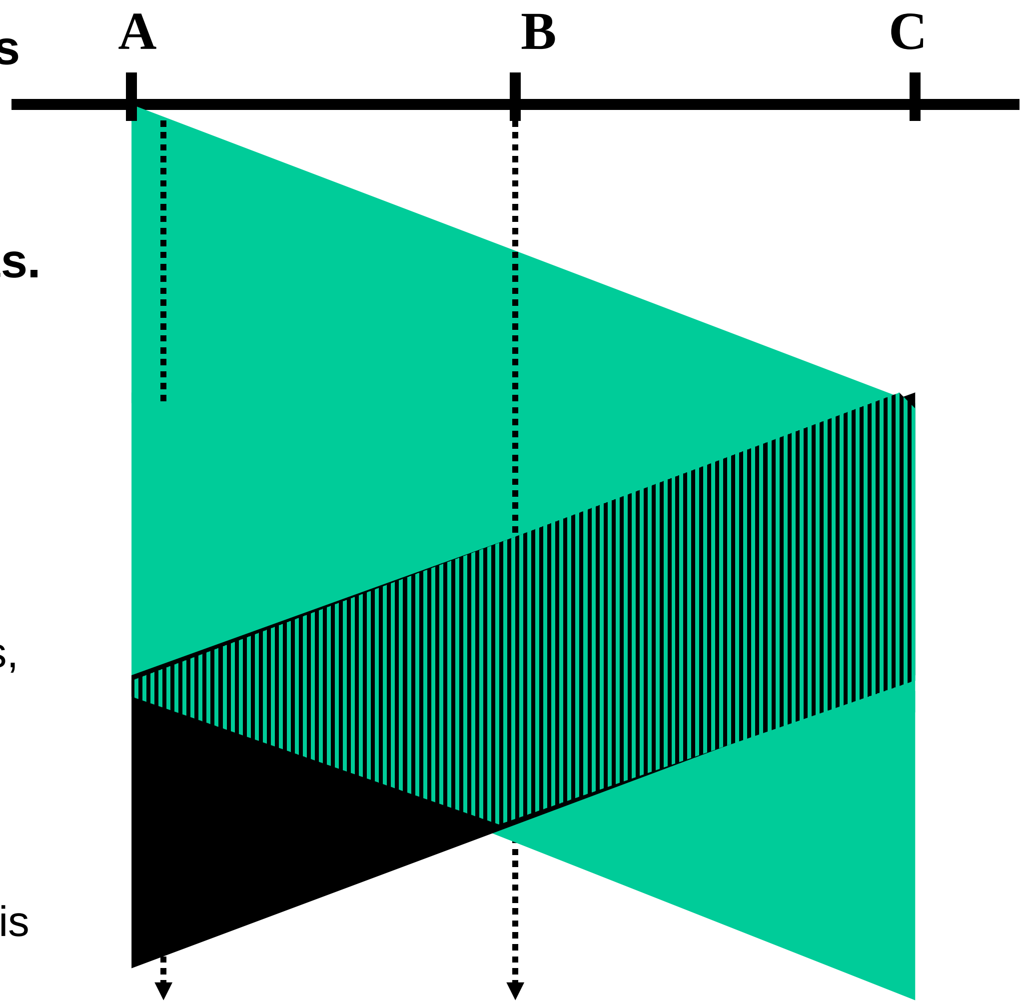
Can calculate length/distance based on transmission rate and propagation speed.

Messy: propagation speed is medium-dependent, low-level protocol details, ..

Minimum packet size is 64 bytes

Cable length ~256 bit times

Example: maximum coax cable length is 2.5 km



Hui Zhang, Dave Eckhardt

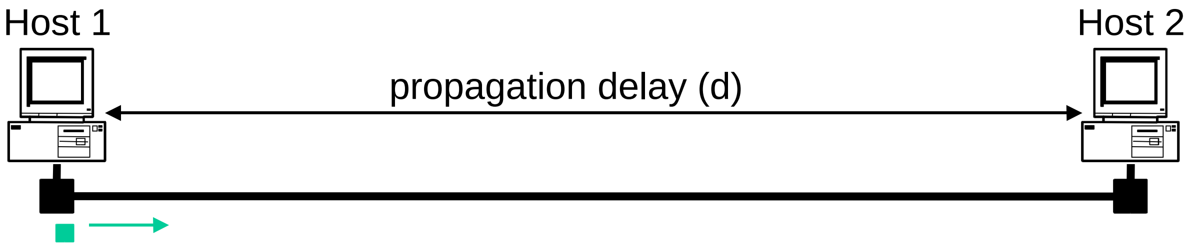
Minimum Packet Size

- Why put a minimum packet size?
- Give a host enough time to detect collisions
- In Ethernet, minimum packet size = 64 bytes (two 6-byte addresses, 2-byte type, 4-byte CRC, and 46 bytes of data)
- If host has less than 46 bytes to send, the adaptor pads (adds) bytes to make it 46 bytes
- What is the relationship between minimum packet size and the length of the LAN?

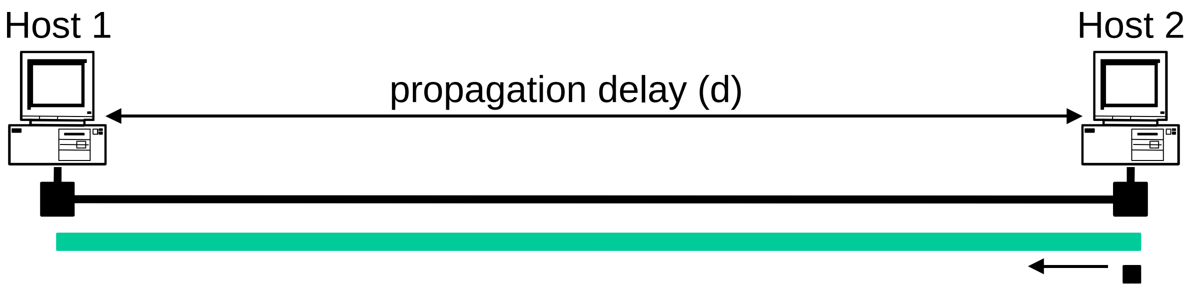
Hui Zhang, Dave Eckhardt

Minimum Packet Size (more)

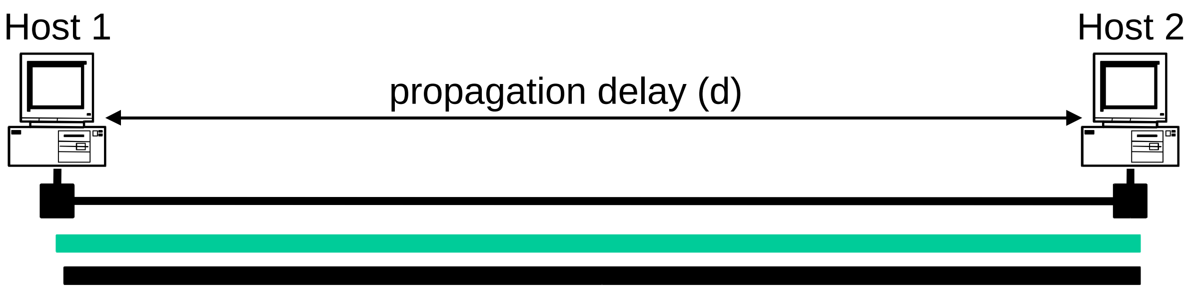
a) Time = t; Host 1 starts to send frame



b) Time = t + d; Host 2 starts to send a frame just before it hears from host 1's frame



c) Time = t + 2*d; Host 1 hears Host 2's frame → detects collision



LAN length = (min_frame_size)*(light_speed)/(2*bandwidth) =

= $(8*64b)*(2*10^8 mps)/(2*10^7$ bps$)$ = 5.12 km

Hui Zhang, Dave Eckhardt

Ethernet Frame Format

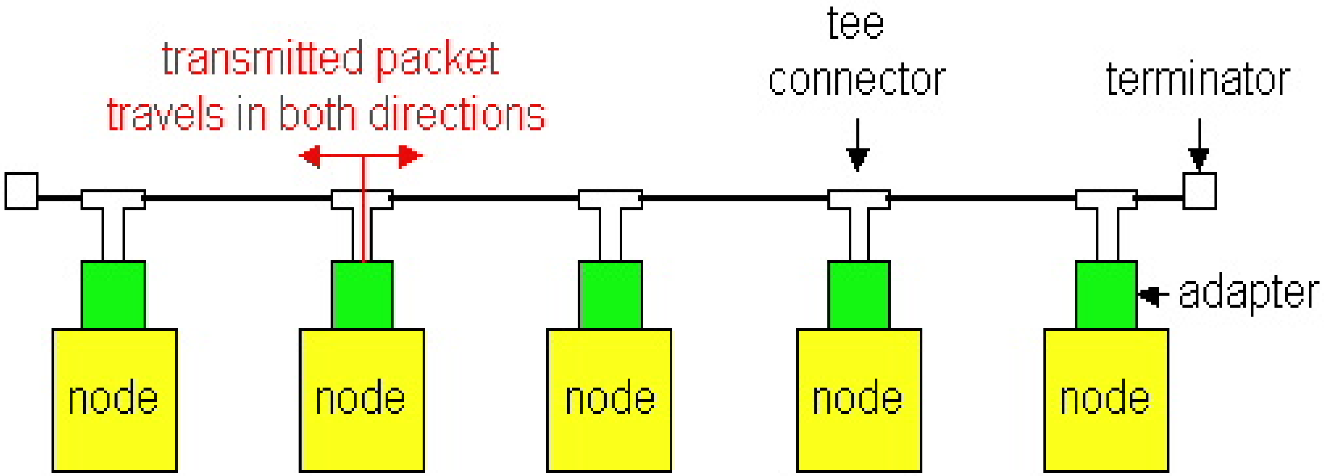
8	6	6	2			4
Preamble	Dest	Source	Type	Data	Pad	CRC

- **Preamble marks the beginning of the frame.**
  - Also provides clock synchronization
- **Source and destination are 48 bit IEEE MAC addresses.**
  - Flat address space
  - Hardwired into the network interface
- **Type field (DIX Ethernet) is a demultiplexing field.**
  - Which network (layer 3) protocol should receive this packet?
  - 802.3 uses field as length instead
- **CRC for error checking.**

Hui Zhang, Dave Eckhardt

Ethernet Technologies: 10Base2

- **10: 10Mbps; 2: under 200 meters max cable length**
- **Thin coaxial cable in a bus topology**

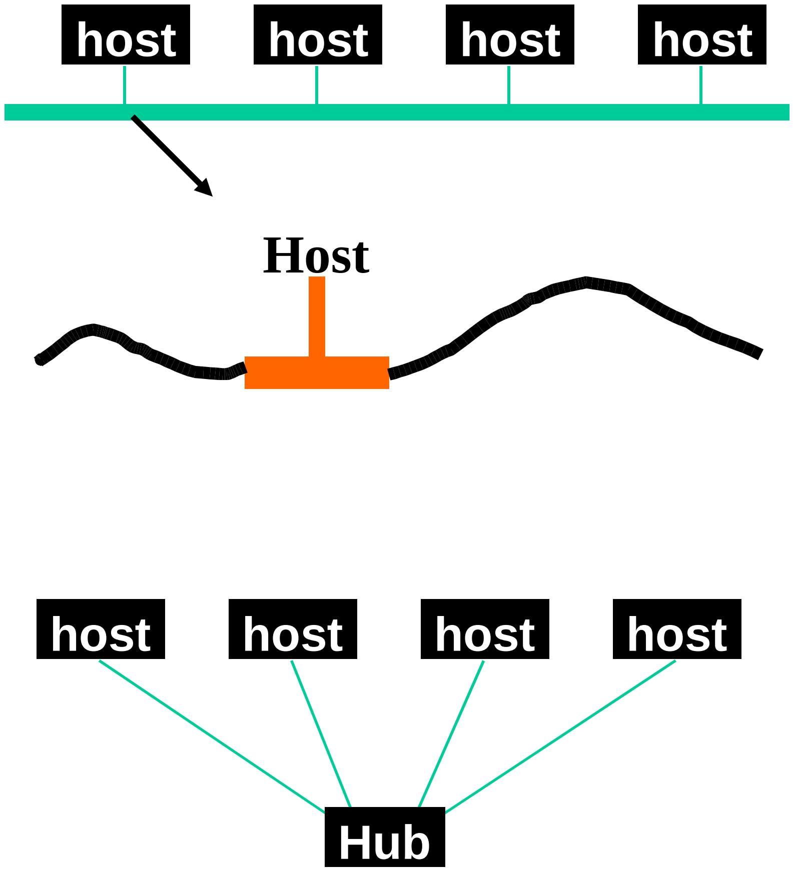


- **Repeaters used to connect up to multiple segments**
- **Repeater repeats bits it hears on one interface to its other interfaces: physical layer device only!**

Hui Zhang, Dave Eckhardt

Compatible Physical Layers

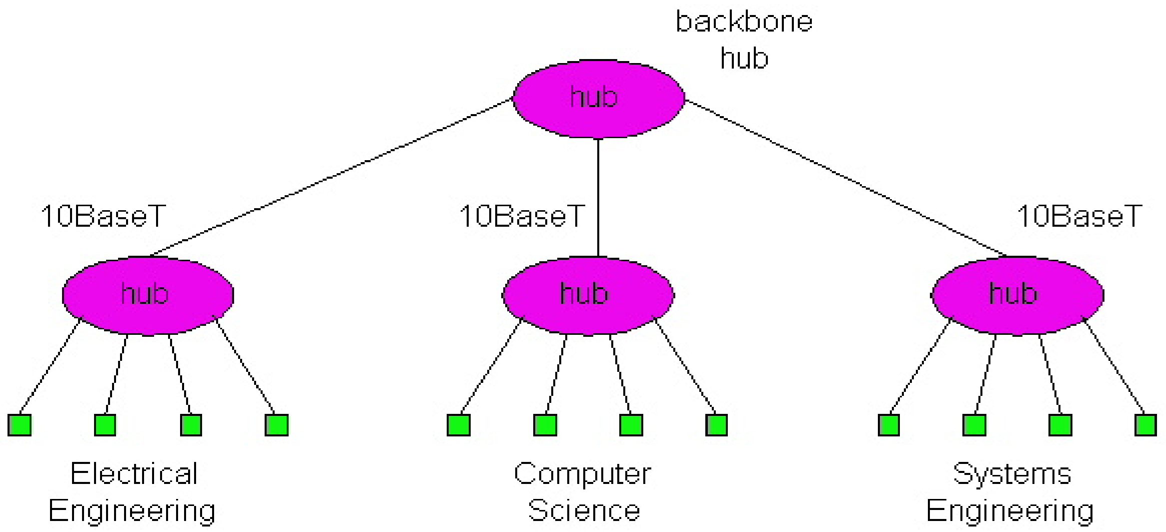
- **10Base2 standard**
  - Thin coax, point-to-point "T" connectors
  - Bus topology
- **10-BaseT: twisted pair**
  - Hub acts as a concentrator
- **3 layers, same protocol!**
  - Key: electrical connectivity between all nodes
  - Deployment is different



Hui Zhang, Dave Eckhardt

10BaseT and 100BaseT

- **10/100 Mbps rate; later called “fast ethernet”**
- **T stands for Twisted Pair**
- **Hub to which nodes are connected by twisted pair, thus “star topology”**



10BaseT and 100BaseT (more)

- **Max distance from node to Hub is 100 meters**
- **Hub can disconnect “jabbering” adapter**
- **Hub can gather monitoring information, statistics for display to LAN administrators**

- **Hubs still preserve one collision domain**
  - Every packet is forwarded to all hosts
- **Use *bridges* to address this problem**
  - Bridges forward a packet only to the port leading to the destination

Hui Zhang, Dave Eckhardt

Gbit Ethernet

- **Use standard Ethernet frame format**
- **Allows for point-to-point links and shared broadcast channels**
- **In shared mode, CSMA/CD is used; short distances between nodes to be efficient**
- **Uses hubs, called here "Buffered Distributors"**
- **Full-Duplex at 1 Gbps for point-to-point links**
  - "Full-duplex" means "both sides transmit simultaneously"

Traditional IEEE 802 Networks: MAC in the LAN and MAN

- **“Ethernet” often considered same as IEEE 802.3.**
  - Not quite identical
- **The IEEE 802.* set of standards defines a common framing and addressing format for LAN protocols.**
  - Simplifies interoperability
  - Addresses are 48 bit strings, with no structure
- **802.3 (Ethernet)**
- **802.5 (Token ring)**
- **802.X (Token bus)**
- **802.6 (Distributed queue dual bus)**
- **802.11 (Wireless)**

LAN Properties

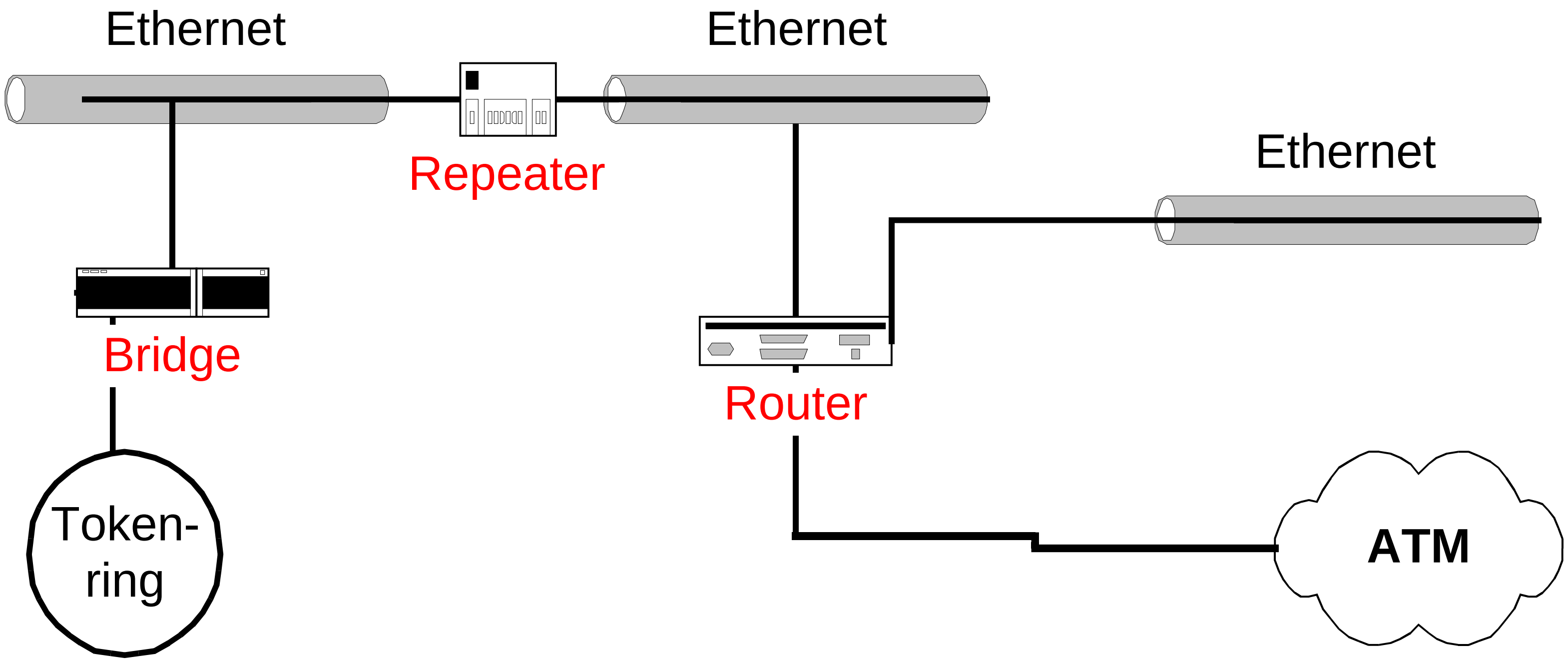
- **Exploit physical proximity.**
  - Typically there is a limitation on the physical distance between the nodes
  - E.g. to collect collisions in a contention based network
  - E.g. to limit the overhead introduced by token passing or slot reservations
- **Relies on single administrative control and some level of trust.**
  - Broadcasting packets to everybody and hoping everybody (other than the receiver) will ignore the packet
  - Token passing protocols assume everybody plays by the rules

Why Ethernet?

- **Easy to manage.**
  - You plug in the host and it basically works
  - No configuration at the datalink layer
- **Broadcast-based.**
  - In part explains the easy management
  - Some of the LAN protocols (e.g. ARP) rely on broadcast
    - Networking would be harder without ARP
  - Not having natural broadcast capabilities adds complexity to a LAN
    - Example: ATM
- **Drawbacks.**
  - Broadcast-based: limits bandwidth since each packet consumes the bandwidth of the entire network
  - Distance

Internetworking

- **There are many different devices for interconnecting networks.**



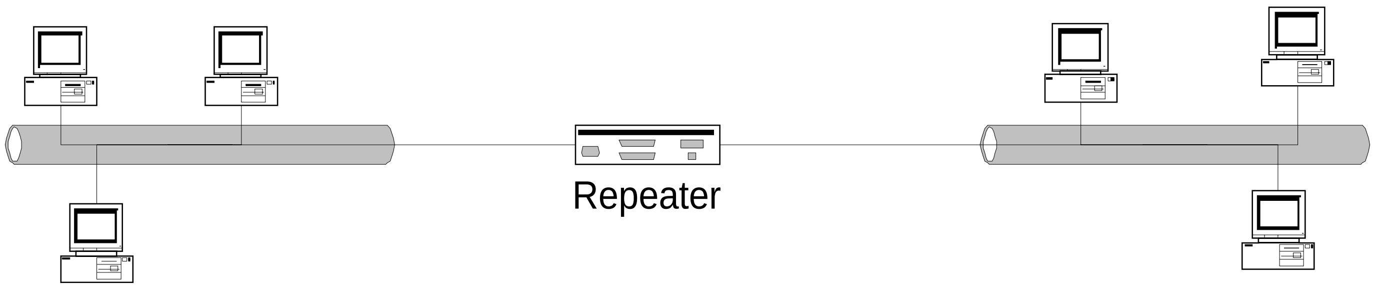
Hui Zhang, Dave Eckhardt

Repeaters

– Used to interconnect multiple Ethernet segments

– Merely extends the baseband cable

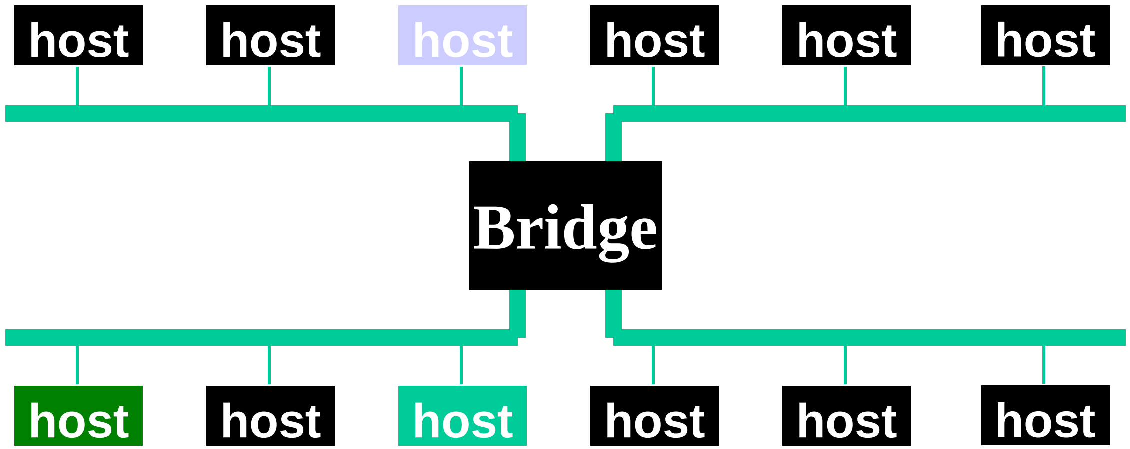
– Amplifies all signals including collisions



Hui Zhang, Dave Eckhardt

Building Larger LANs: Bridges

- **Repeaters, hubs rebroadcast packets**
- **Bridges connect multiple IEEE 802 LANs at layer 2.**
  - Forward packets only to the right port
  - Reduce collision domain compared with single LAN



Hui Zhang, Dave Eckhardt

Transparent Bridges

- **Overall design goal:** Complete transparency
  – "Plug-and-play"
  – Self-configuring without hardware or software changes
  – Bridges should not impact operation of existing LANs

- **Three parts to transparent bridges:**
  (1) Forwarding of Frames
  (2) Learning of Addresses
  (3) Spanning Tree Algorithm

Frame Forwarding

- **Each bridge maintains a** forwarding database **with entries**

  < MAC address, port, age>

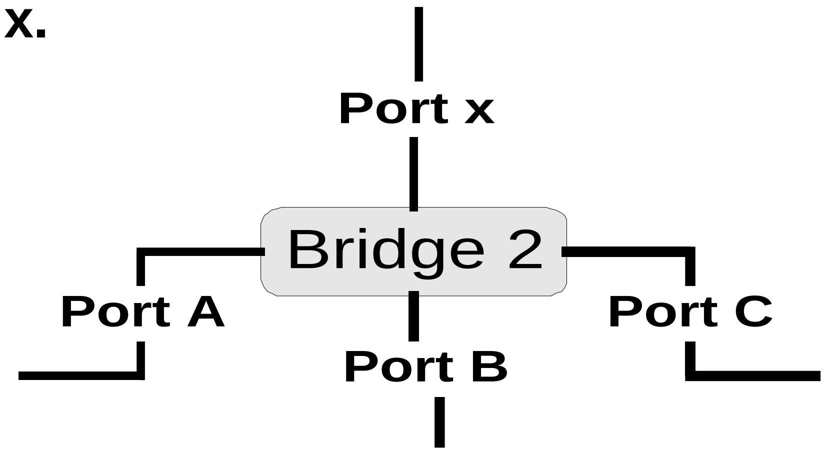
MAC address:	host name or group address
port:	port number of bridge
age:	aging time of entry

with interpretation:

  - a machine with **MAC address** lies in direction of the numbered **port** from the bridge. The entry is **age** time units old.

Frame Forwarding 2

- Assume a frame arrives on port x.



Search if MAC address of destination is listed for ports A, B, or C.

Found?

Forward the frame on the appropriate port

Not found ?

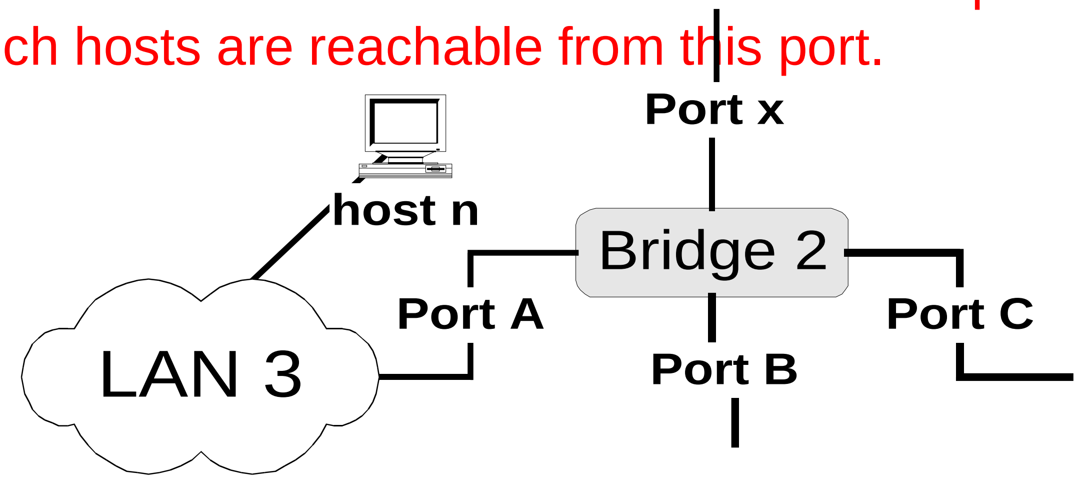
Flood the frame, i.e., send the frame on all ports except port x.

Hui Zhang, Dave Eckhardt

Address Learning

- **In principle, the forwarding database could be set statically (=static routing)**
- **In the 802.1 bridge, the process is made automatic with a simple heuristic:**

The source field of a frame that arrives on a port tells which hosts are reachable from this port.

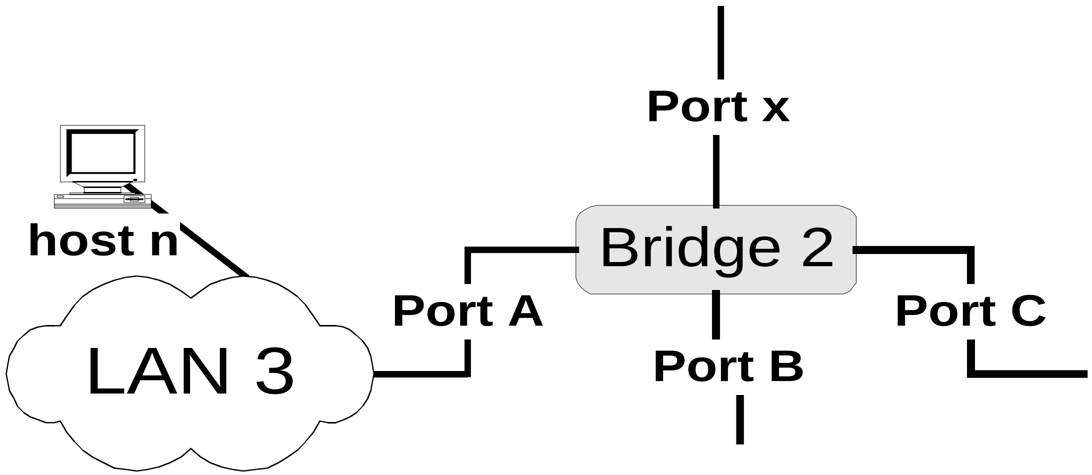


Hui Zhang, Dave Eckhardt

Address Learning 2

Algorithm:

- For each frame received, the source stores the source field in the forwarding database together with the port where the frame was received.
- All entries are deleted after some time (default is 15 seconds).



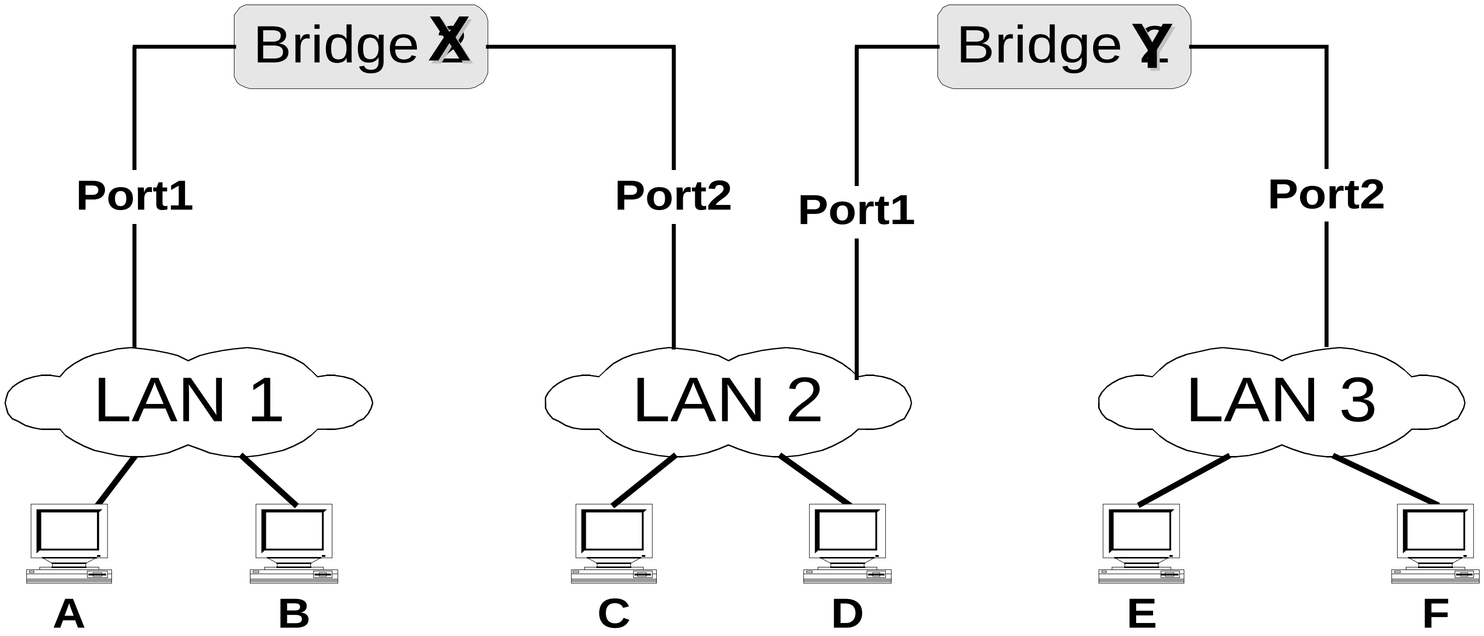
Hui Zhang, Dave Eckhardt

Example

•Consider the following packets:

<Src=A, Dest=F>, <Src=C, Dest=A>, <Src=E, Dest=C>

•What have the bridges learned?



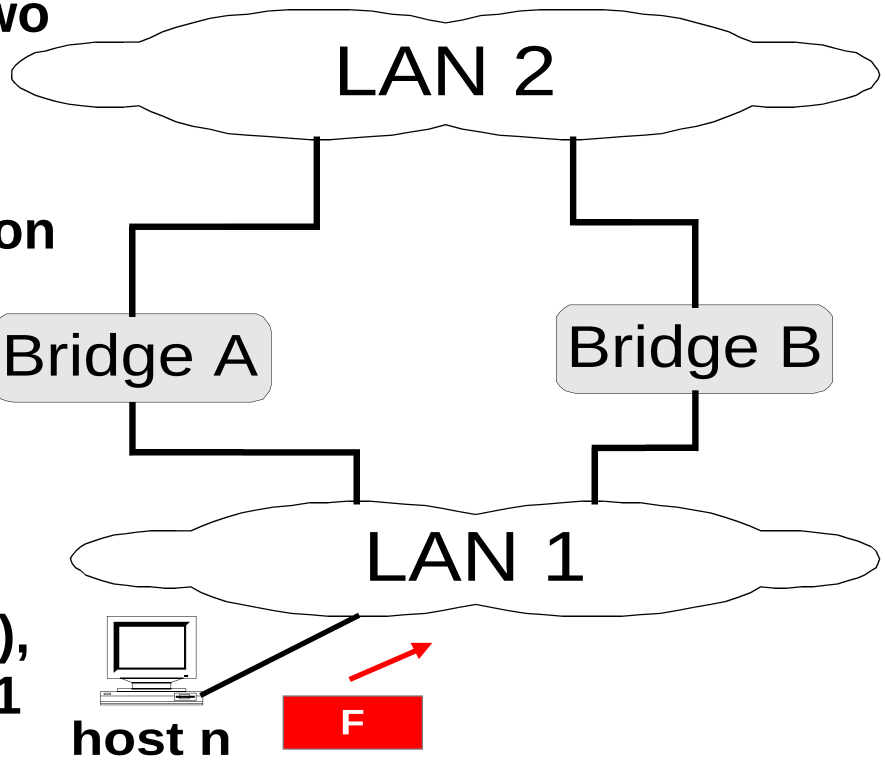
Danger of Loops

- **Two LANs connected by two bridges.**
- ***Host n* is transmits a frame F to unmapped station**

What happens?

- **Bridges A and B flood F to LAN 2.**
- **Bridge B sees F on LAN 2 (with unknown destination), and copies it back to LAN 1**
- **Bridge A does the same!**
- **The copying continues**

Where's the problem? What to do?



Spanning Trees

- A solution to the loop problem is to not have loops in the topology

- IEEE 802.1 has an algorithm that builds and maintains a spanning tree in a dynamic environment.

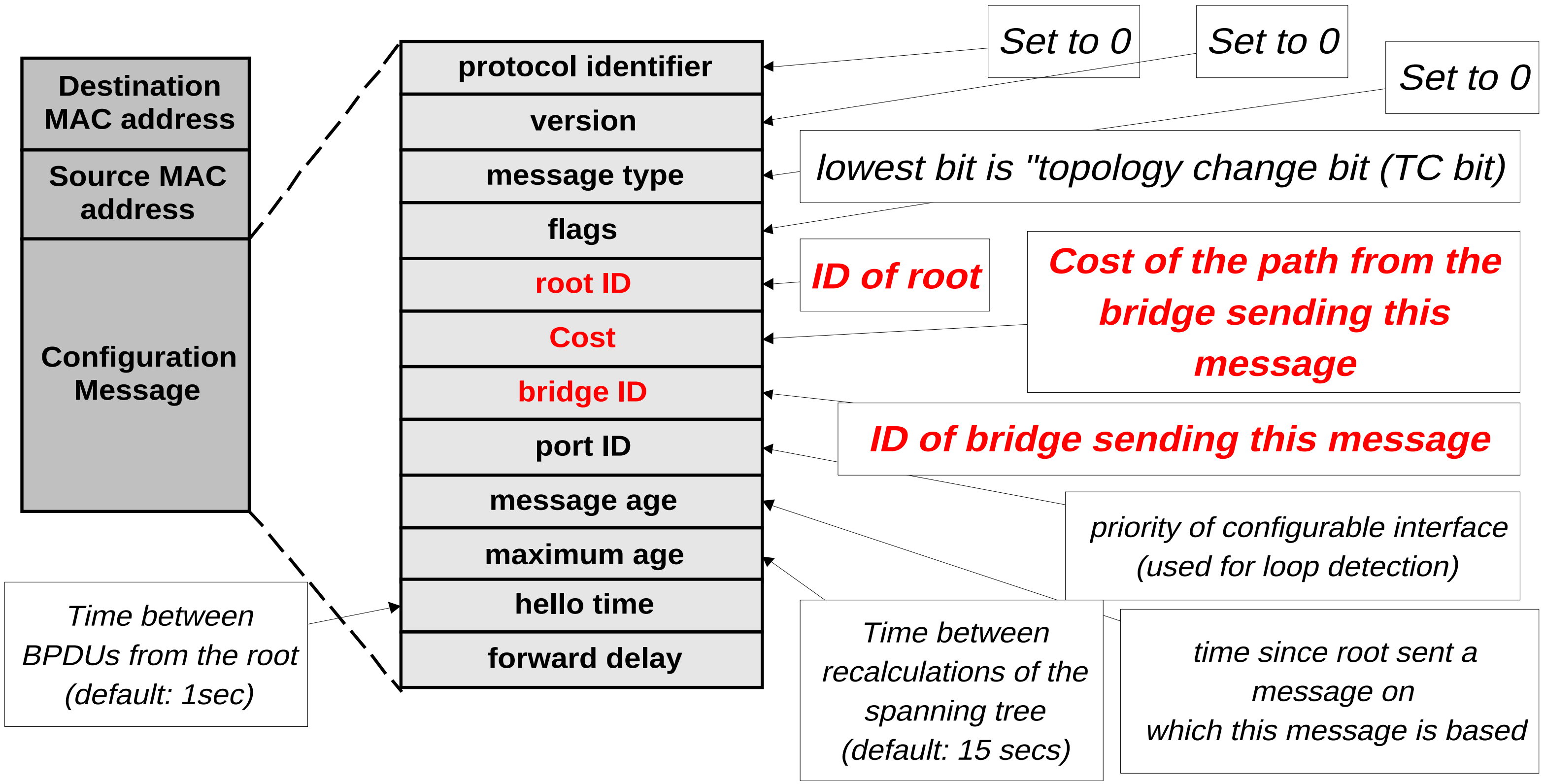
- Bridges exchange messages to configure the bridge (Configuration Bridge Protocol Data Unit, Configuration BPDUs) to build the tree.

What do the BPDUs do?

With the help of the BPDUs, bridges can:

- **Elect a single bridge as the** root bridge**.**
- **Calculate the distance of the shortest path to the root bridge**
- **Each LAN can determine a** designated bridge**, which is the bridge closest to the root. The designated bridge will forward packets towards the root bridge.**
- **Each bridge can determine a** root port**, the port that gives the best path to the root.**
- **Select ports to be included in the spanning tree.**

Configuration BPDUs



Hui Zhang, Dave Eckhardt

Concepts

- **Each bridge has a unique identifier:**

Bridge ID = <MAC address + priority level>

Note that a bridge has several MAC addresses
(one for each port), but only one ID

- **Each port within a bridge has a unique identifier (port ID).**
- Root Bridge: **The bridge with the lowest identifier is the root of the spanning tree.**
- Path Cost: **Cost of the least cost path to the root from the port of a transmitting bridge; Assume it is measured in #Hops to the root.**
- Root Port: **Each bridge has a root port which identifies the next hop from a bridge to the root.**

Hui Zhang, Dave Eckhardt

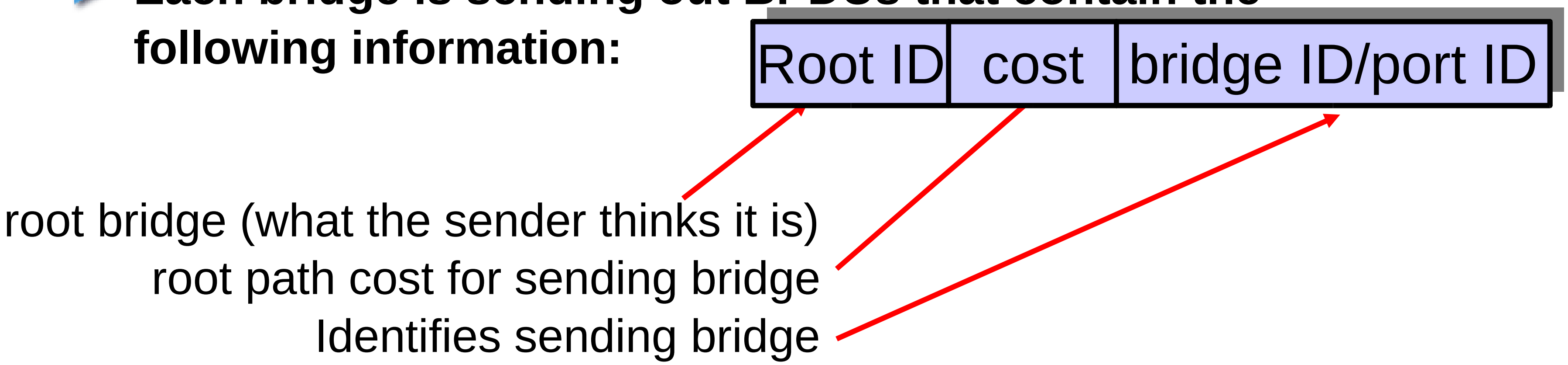
Concepts

- Root Path Cost: **For each bridge, the cost of the min-cost path to the root**
- Designated Bridge, Designated Port: **Single bridge on a LAN that provides the minimal cost path to the root for this LAN:**
  - if two bridges have the same cost, select the one with highest priority
  - if the min-cost bridge has two or more ports on the LAN, select the port with the lowest identifier
- Note: **We assume that "cost" of a path is the number of "hops".**

Steps of Spanning Tree Algorithm

1. Determine the root bridge
2. Determine the root port on all other bridges
3. Determine the designated port on each LAN

- **Each bridge is sending out BPDUs that contain the following information:**



Hui Zhang, Dave Eckhardt

Ordering of Messages

- **We can order BPDU messages with the following ordering relation "∠":**

If (R1 < R2)

M1 ∠ M2

elseif ((R1 == R2) and (C1 < C2))

M1 ∠ M2

elseif ((R1 == R2) and (C1 == C2) and (B1 < B2))

M1 ∠ M2

Hui Zhang, Dave Eckhardt

Determine the Root Bridge

- **Initially, all bridges assume they are the root bridge.**
- **Each bridge B sends BPDUs of this form on its LANs:**

B	0	B

- **Each bridge looks at the BPDUs received on all its ports and its own transmitted BPDUs.**
- **Root bridge is the smallest received root ID that has been received so far (Whenever a smaller ID arrives, the root is updated)**

Calculate the Root Path Cost
Determine the Root Port

- **<u>At this time</u>: A bridge B has a belief of who the root is, say R.**
- **Bridge B determines the Root Path Cost (Cost) as follows:**
  - *If B = R :* Cost = 0.
  - *If **B** ∠ P* Cost = {Smallest Cost in any of BPDUs that were received from R} + 1
- B's root port **is the port from which B received the lowest cost path to R (in terms of relation "∠").**
- **Knowing R and Cost, B can generate its BPDU (but will not necessarily send it out):**

R	Cost	B

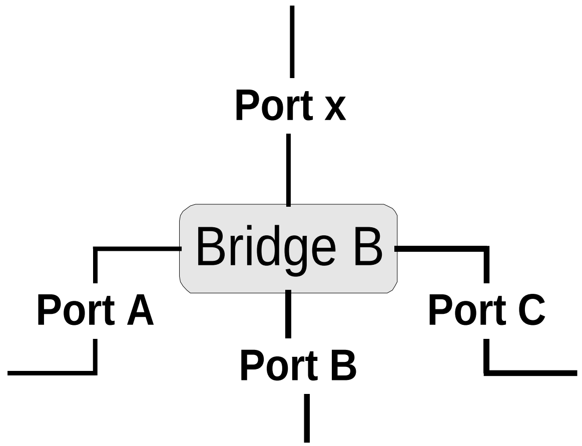
Calculate the Root Path Cost
Determine the Root Port

- **<u>At this time</u>: B has generated its BPDU**

R	Cost	B

- **B will send this BPDU on one of its ports, say** port x**, only if its BPDU is lower (via relation ∠) than any BPDU that B received from port x.**

- **In this case, B also assumes that it is the** designated bridge **for the LAN to which the port connects.**



Selecting the Ports for the Spanning Tree

- **<u>At this time</u>: Bridge B has calculated the root, the root path cost, and the designated bridge for each LAN.**
- **Now** B can decide which ports are in the spanning tree:
  - B's root port is part of the spanning tree
  - All ports for which B is the designated bridge are part of the spanning tree.
- **B's ports that are in the spanning tree will forward packets** (=forwarding state)
- **B's ports that are not in the spanning tree will not forward packets** (=blocking state)

Hui Zhang, Dave Eckhardt

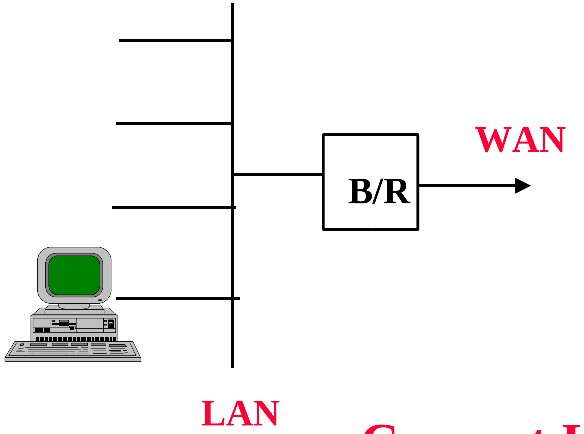
Ethernet Switches

- **Bridges make it possible to increase LAN capacity.**
  - Packets are no longer broadcasted - they are only forwarded on selected links
  - Adds a switching flavor to the broadcast LAN
- **Ethernet switch is a special case of a bridge: each bridge port is connected to a single host.**
  - Can make the link full duplex (really simple protocol!)
  - Simplifies the protocol and hardware used (only two stations on the link) – no longer full CSMA/CD
  - Can have different port speeds on the same switch
    - Unlike in a hub, packets can be stored
    - An alternative is to use cut through switching

Ethernet – Anything but Name and Framing

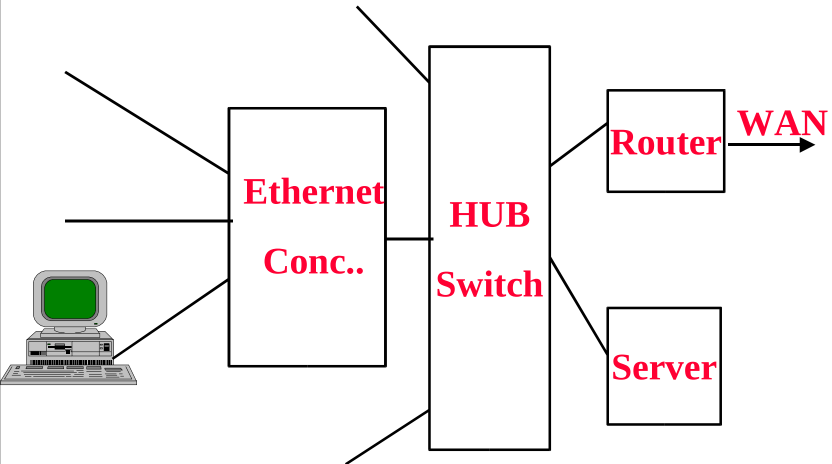
Ethernet or 802.3

Early Implementations



- A Local Area Network
- MAC addressing, non-routable
- BUS or Logical Bus topology
- Collision Domain, CSMA/CD
- Bridges and Repeaters for distance/capacity extension
- 1-10Mbps: coax, twisted pair (10BaseT)

Current Implementations



- Switched solution
- Little use for collision domains
- 80% of traffic leaves the LAN
- Servers, routers 10 x station speed
- 10/100/1000 Mbps, 10gig coming: Copper, Fiber
- 95% of new LANs are Ethernet

CSMA - Carrier Sense Multiple Access

CD - Collision Detection

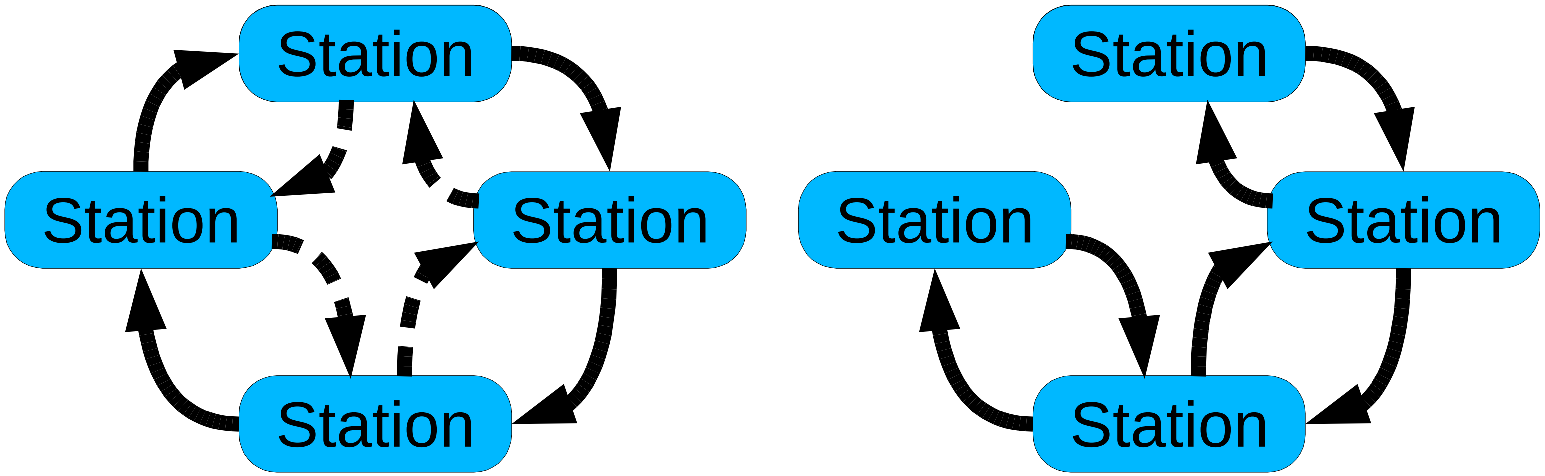
Outline

- **Ethernet**
  - Conceptual history
  - Carrier sense, Collision detection
  - Ethernet history, operation (CSMA/CD)
  - Packet size
  - Ethernet evolution
  - Connecting Ethernets
- ☞ **Not Ethernet**
  - FDDI, wireless, ...

FDDI

- **Fiber Distributed Data Interface**
  - "Token ring grown up"
  - 100 Mbit/s
  - Nodes connected by fiber
    - – Multi-mode fiber driven by LED
    - – Single-mode fiber driven by laser (long distance)
  - Up to 500 nodes in ring, total fiber length 200 km
  - Organized as <u>*dual ring*</u>

FDDI – Fault Recovery



Hui Zhang, Dave Eckhardt

Token Bus

- **Basic idea**
  - Ethernet is cool
    - ...run one cable throughout building
    - ...popular technology, commodity, cheap
  - Factory automation people worry about frame delay
    - ...must bound delay from sensor to controller to robot
  - Token ring is cool - firm bound on transmission delay
- **Virtual network**
  - Run token-ring protocol on Ethernet frames
    - No collisions, delay bound (though generally worse)
  - May be a nested lie: bus atop bridge atop star!

NCR WaveLAN

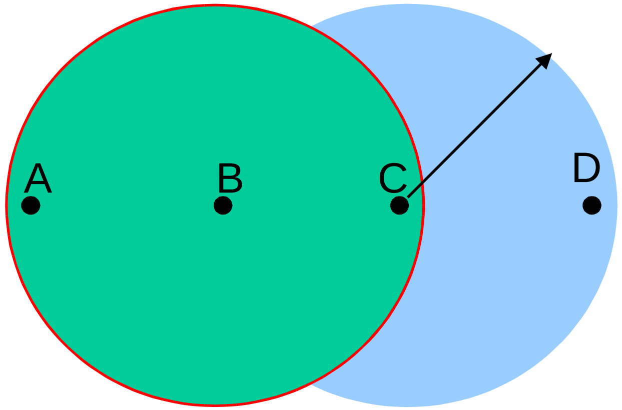
- **Basic idea**
  - Ethernet is cool
    - ..."wireless Ethernet" would be cooler
    - ...re-use addresses, bridging protocols, ...
- **Recall: radio collision detection is hard**
  - Undetected collisions waste a lot of time
  - Hack: collision <u>*inference*</u>
    - Is medium busy when you want to transmit?
      - » Assume true of other stations too
      - » Assume a collision will happen
      - » “Back off” pro-actively

Wireless (802.11)

- **Designed for use in limited geographical area (i.e., couple of hundreds of meters)**
- **Designed for three physical media (run at either 1Mbps or 2 Mbps)**
  - Two based on spread spectrum radio
  - One based on diffused infrared

Collision Avoidance: The Problems

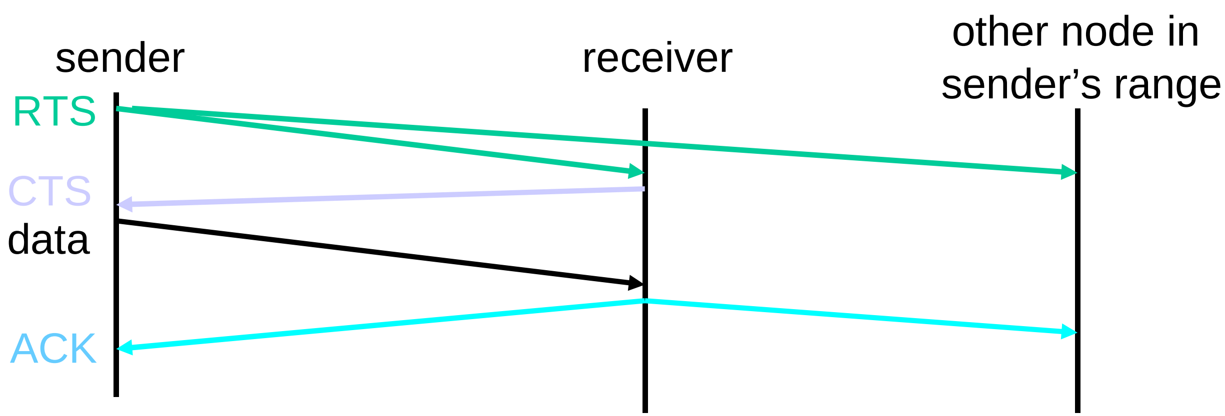
- Reachability is not transitive: if A can reach B, and B can reach C, it doesn't necessary mean that A can reach C



- **Hidden nodes**: A and C send a packet to B; neither A nor C will detect the collision!
- **Exposed node**: B sends a packet to A; C hears this and decides not to send a packet to D (despite the fact that this will not cause interference)!

Hui Zhang, Dave Eckhardt

Multiple Access with Collision Avoidance (MACA)



- **Before every data transmission**
  - Sender sends a Request to Send (RTS) frame containing the length of the transmission
  - Receiver respond with a Clear to Send (CTS) frame
  - Sender sends data
  - Receiver sends an ACK; now another sender can send data
- **When sender doesn't get a CTS back, it assumes collision**

Summary

- **Problem: arbitrate between multiple hosts sharing a common communication media**
- **Wired solution: Ethernet (use CSMA/CD protocol)**
  - Detect collisions
  - Backoff exponentially on collision
- **Wireless solution: 802.11**
  - Use MACA protocol
  - Cannot detect collisions; try to avoid them
  - Distribution system & frame format in discussion sections