ELEC / COMP 177 – Fall 2012

Computer Networking → Ethernet

Some slides from Kurose and Ross, *Computer Networking*, 5th Edition

Schedule - Assignments

- **Project #2** – Due Tonight by midnight
- **Homework #5** – Due Tuesday, Nov 13th
- ***Later this semester:***
  - ***Homework #6 - Presentation on security/privacy***
    - ***Topic selection*** *– Due Tuesday, Nov 20th*
    - **Slides** – Due Monday, Nov 26th
    - **Present!** – Tuesday, Nov 27th (and Thursday)
  - ***Project #3*** *– Due Tuesday, Dec 4th*

Schedule - Topics

- Today – Ethernet
- Thursday – Putting it all together (Review)

- Next Week – No class or Lab! (Traveling)

- Tuesday 20th – Security / Firewalls
- Thursday 22nd - Thanksgiving

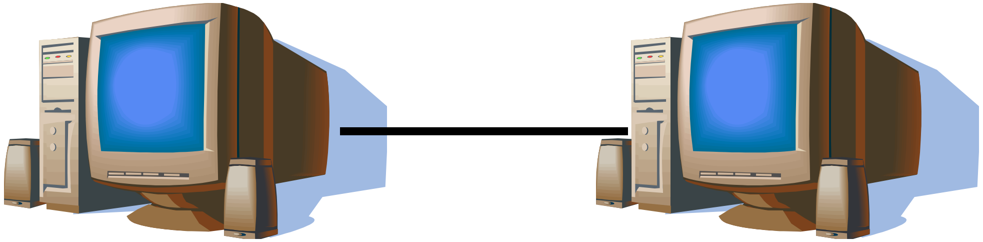
Getting Help

- **This Week**
  - Change in office hours
  - Thur 2-4pm
  - Fri 1-3pm
  - And (always), via email or other times by request
- **Next Week**
  - Via email only (include current project code)

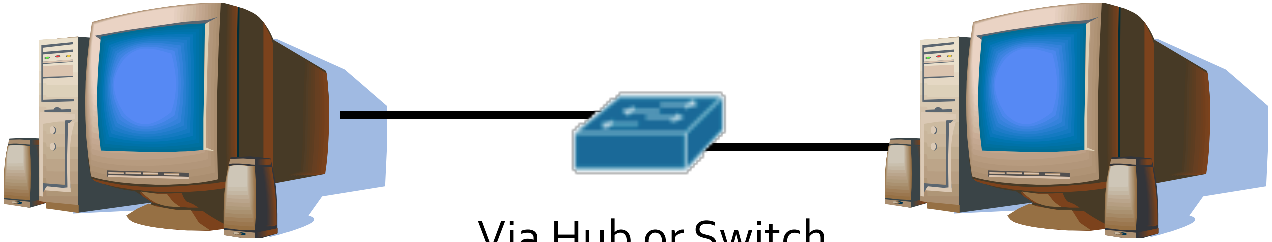
Project #2

- <u>**Peer evaluations**</u>

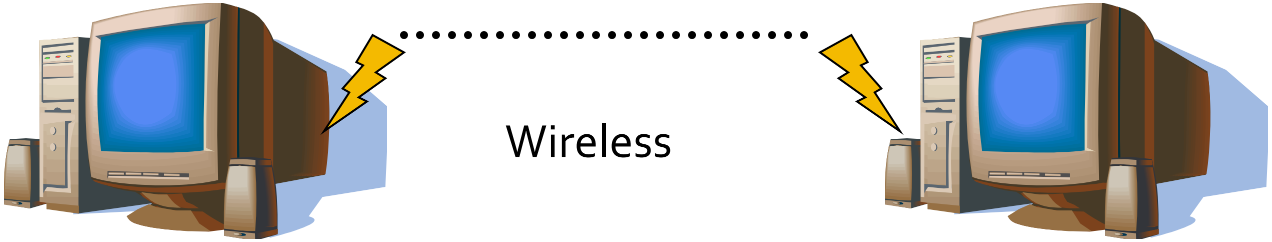
How to Physically Connect Computers?



Direct Connection (USB, Firewire, etc...)



Via Hub or Switch



Wireless

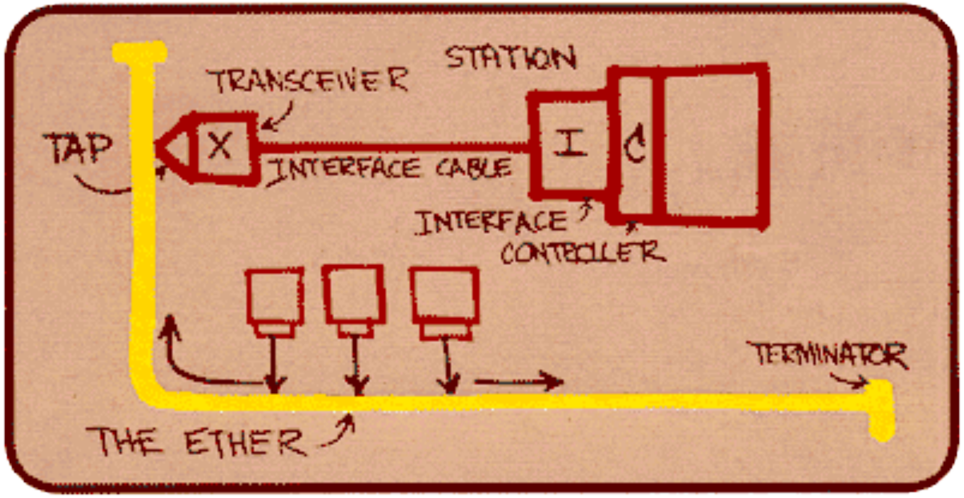
How to Build a Network?

- Four challenges
  - *Encoding* – How to format bits on wire
    - Different solutions for different media (copper, optical, wireless)
  - *Framing* – How to separate sequences of bits into independent message
  - *Error Detection* – How to detect corrupted messages (and possibly repair them)
  - *Media Access Control* – How to share a single wire or frequency among multiple hosts
    - Goals: Fair between users, high efficiency, low delay, fault tolerant

Standards that Solve Challenges

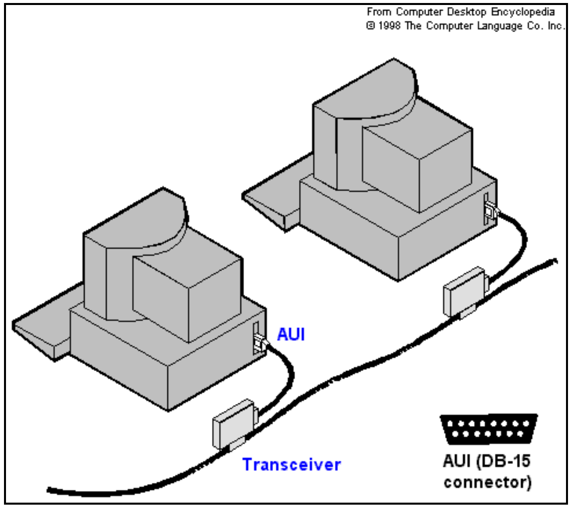
- Many competing standards
  with varying levels of complexity
  (for both wired and wireless networks)
  - Token Ring (IEEE 802.5)
  - Ethernet (IEEE 802.3)
  - Wi-Fi (IEEE 802.11 a/b/g/n)
- We focus on Ethernet networks in this course
  - Different standards made different choices, but design principles are similar

The Original Ethernet



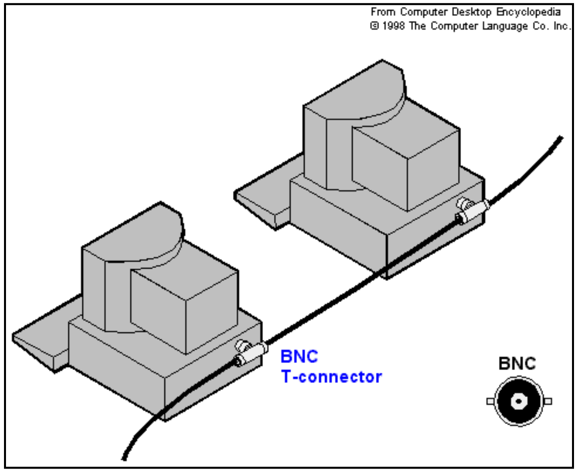
Original picture drawn by **Bob Metcalfe**, inventor of Ethernet (1972 – Xerox PARC)

Ether – 19th century name for media enabling the propagation of light


From Computer Desktop Encyclopedia
© 1998 The Computer Language Co. Inc.
AUI
Transceiver
AUI (DB-15 connector)

The 10Mb/s Ethernet Standard

- IEEE 802.3
- Common MAC protocol and frame format
- Multiple physical layers to choose from
  - Bus architecture (shared)
    - 10Base-5 : Original Ethernet
      - Thick coaxial cable with taps every 2.5 meters to clamp on network devices
    - 10Base-2 : Thin coaxial cable version with BNC "T" connectors
  - Star architecture (point-to-point)
    - 10Base-F / 10-Base-T standards – Introduced later!



From Computer Desktop Encyclopedia
© 1998 The Computer Language Co. Inc.

Challenge - Encoding

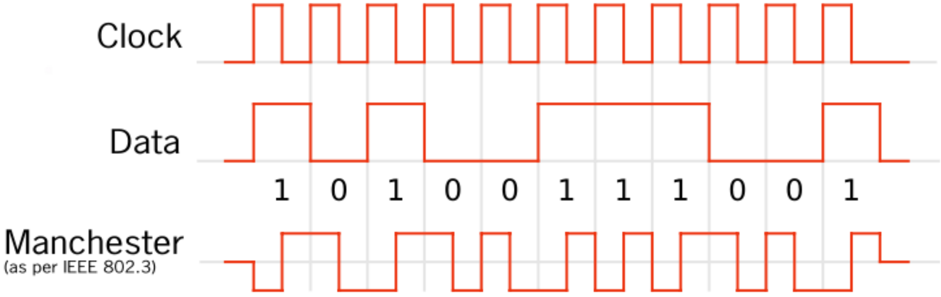
- How to turn bits into physical signals to send across wire?
- Data transmission across media always distorts data
    - Attenuation (amplitude reduction)
    - Distortion (change in shape)
- Encoding will make transmitted data resilient to these effects

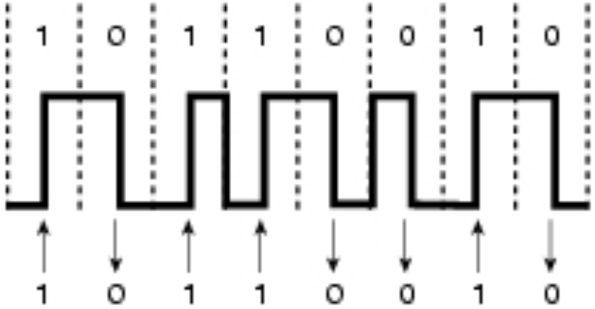
Challenge - Encoding

- Common challenges to recovering data
  (*ignoring any modulation issues, and only using discrete high and low signals*)
  - Baseline Wander – Receiver distinguishes between high and low by keeping an average.
    (Above average = 1, Below average = 0)
    - If data stream contains long periods of 1's or 0's, average can shift!
  - Clock Recovery
    - Too expensive to have dedicated clock wire
    - Must examine incoming data and derive clock
    - Derived clock can skew during long periods of all 1's or 0's

Manchester Encoding (10Mb/s Ethernet)

- Exclusive-OR of input bit and clock
- Pulse decoded by direction of the midpulse transition rather than by its sampled level value
- No long periods without clock transition (prevents receiver clock skew and baseline wander)
- Drawback: Only 50% efficient!
  - *Bit rate* (actual data transfer) is half of *baud rate* (raw channel capacity)





Images from Wikipedia and http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/ethernet.htm

Challenge – Ethernet Frame Format

Bytes:	7	1	6	6	2	0-1500	0-46	4	
	Preamble	SFD	DA	SA	Type	Data	Pad	CRC	*Gap…*

- **Preamble**
  - Alternating 1's and 0's provide data transitions for synchronization
  - Used to train receiver clock-recovery circuit (critical since different transmitters will be using different clocks)
- **SFD** (Start of Frame Delimiter)
  - Indicates start of frame data. Always 0xAB
- **DA** (Destination Address) / **SA** (Source Address)
- **Type**: Indicates data type or length
- **Pad**: Zeroes used to ensure minimum frame length of 64 bytes
- **CRC** (Cyclic Redundancy Check)
- **Interframe Gap**: Allow time for receiver to recover before next packet
  - Length: 96 times the length of time to transmit 1 bit
  - 9.6 µs for 10 Mbit/s Ethernet, 960 ns for 100 Mbit/s Ethernet, and 96 ns for 1 Gbit/s Ethernet

Ethernet - Addressing

- All Ethernet devices have globally unique 48-bit address assigned by manufacturer
  - IEEE assigns OUI (Organizationally Unique Identifier) prefix to each manufacturer
  - Remaining bits are unique per device and chosen by manufacturer
- Example: `0x 00-07-E9-CB-79-4F`
  - `0x 00-07-E9` = Intel Corp (assigned by IEEE)
    - Bit also indicates device is unicast, not multicast
  - `0x CB-79-4F` = Unique address per NIC (picked by Intel)
- Special destination address to broadcast to all devices
  - `0x FF-FF-FF-FF-FF-FF`
- NIC is responsible for filtering packets
  - Address matches (or broadcast)? Send up to host
  - Otherwise, discard

Challenge – Error Detection

- How to detect errors in transmission across copper wire / fiber?
- Ethernet solution
  - 32-bit CRC (Cyclic Redundancy Check) stored in frame header
- n-bit CRC detects all error bursts not longer than n bits, and a 1-2^{-n} fraction of all longer error bursts
  - Very useful since most transmission errors on a wire are bursty in nature!

Ethernet CRC

- Limitations of Ethernet CRC
  - No protection against deliberate corruption or alteration of message in transit – Not security!
  - No protection against corruption when packet is transferred through host systems, but only across wire
    - Can still have failures in NIC, memory, data bus (PCI) at either end of the network
  - Insufficient information to recover from error
- Design decision - More efficient to retransmit upon error than to always send enough redundant bits to repair errors
  - Receiver discards invalid packet
  - Higher-level protocols might trigger retransmit by sender

Ethernet CRC

- Why do I need a CRC? If Ethernet is unreliable, shouldn't I just let a higher-level protocol detect any errors?

Challenge – Media Access Control

- **CSMA / CD** Protocol for Ethernet
  - *Carrier Sense Multiple Access with Collision Detection*
  - Developed for use with single shared coaxial cable of original Ethernet
  - Decentralized technique – No central arbitration, access tokens, or assigned time slots are needed to manage transmission

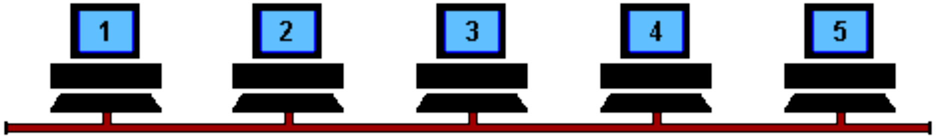
Ethernet – Media Access Control

- How to Transmit
  - Prepare frame for transmission
  - Check if media is idle ("Carrier Sense"). If not, wait until idle (plus interframe gap)
  - Transmit frame. Listen for any collisions and enter recovery mode
  - If no collision, finish transmitting
- Max frame size of 1500 bytes prevents one device from monopolizing network

Animation from http://www.datacottage.com/nch/eoperation.htm

Ethernet – Media Access Control

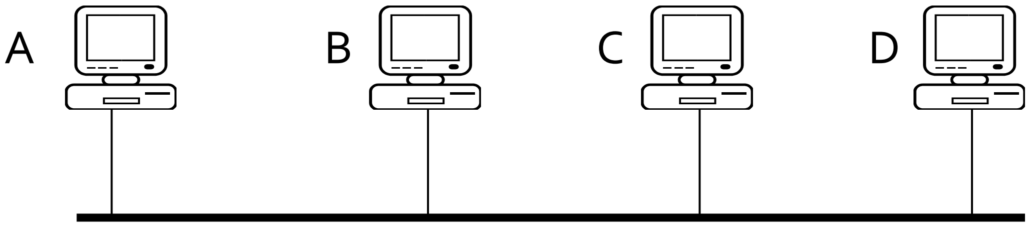
- How to Handle Collisions on Shared "Ether"
  - Continue transmission until minimum packet time is reached to ensure that all receivers detect the collision.
  - Wait random backoff time based on number of collisions
    - Backoff time exponentially increases if >1 collision per frame
  - Restart frame transmission again



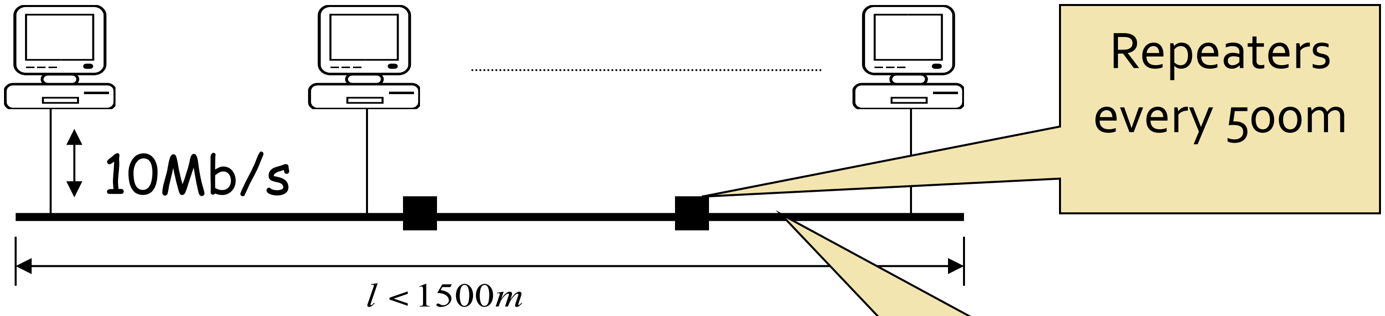
Animation from http://www.datacottage.com/nch/eoperation.htm

Ethernet – Media Access Control

- Example of worst case collision
  - Two most-distant devices send a frame (A and D)
  - D doesn't start transmitting until frame from A has almost arrived
  - D detects collision almost immediately
  - A doesn't detect collision until data propagates all the way down the wire
- Maximum time for collision detection
  - Twice the signal propagation time across entire network (for signal from A to reach D and return with collision)



Minimum Packet Size (10Mb Ethernet)



Repeaters every 500m

Thick copper coaxial cable

$$PROP_{\max} = l / c = 1500 / 2.5 \times 10^{8} = 6 \mu s$$

$$TRANSP > 2PROP \Rightarrow TRANSP > 12 \mu s$$

$$\therefore Packetsize \geq (12 \mu s) \times 10 Mb / s = 120 bits$$

- In practice, minimum packet size is 512 bits
  - Allows for extra time to detect collisions
  - Allows for "repeaters" that can boost signal

Ethernet – Media Access Control

- Ethernet device receives frames meeting any of the following conditions:
  - Frames addressed to its own MAC address
  - Frames addressed to the broadcast address
  - Frames addressed to the multicast address (if configured for this device)
  - All frames (in *promiscuous* mode)

Ethernet – MAC Goals

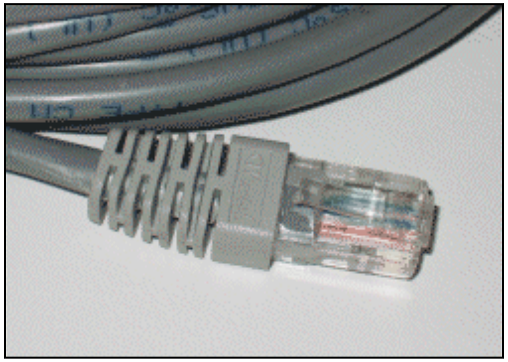
- Previously mentioned design goals – **Were they accomplished?**
  - Fair between users? (What if users cheat?)
  - High efficiency?
  - Low delay?
  - Fault tolerant?

New physical layers

Scaling Ethernet

New Technology Needed

- No more single wire shared by all devices!
  - Too hard to increase to higher speeds
- Point-to-point networking
  - Still use MAC protocol and frame format
  - New network device: Ethernet repeater (“hub”)
  - New physical layer
    - Straight-through cable (device ↔ hub) or crossover cable (device ↔ device)



New Physical Layers

- 100 Mb/s
  - 100Base-T4 (4 pairs copper, 100 meters max)
  - 100Base-TX (2 pairs high-quality copper, 100 meters max)
  - 100Base-FX (2 optical fibers)
  - … and others
- 1000 Mb/s
  - 1000Base-T (4 pairs high-quality copper, 100 meters max)
  - 1000Base-FX (2 optical fibers)
  - … and others
- Different physical layers (and encoding standards)
- Same frame format, error correction, and MAC protocol

Full Duplex Ethernet

- Simultaneous two-way transmission (send and receive)
- No more collisions or retransmissions! (at least due to Ethernet)
- Only useful over point-to-point links, not shared bus (or hub topology)
  - Design enabled by pervasive deployment of switches

Gigabit Ethernet – Same 4 Challenges

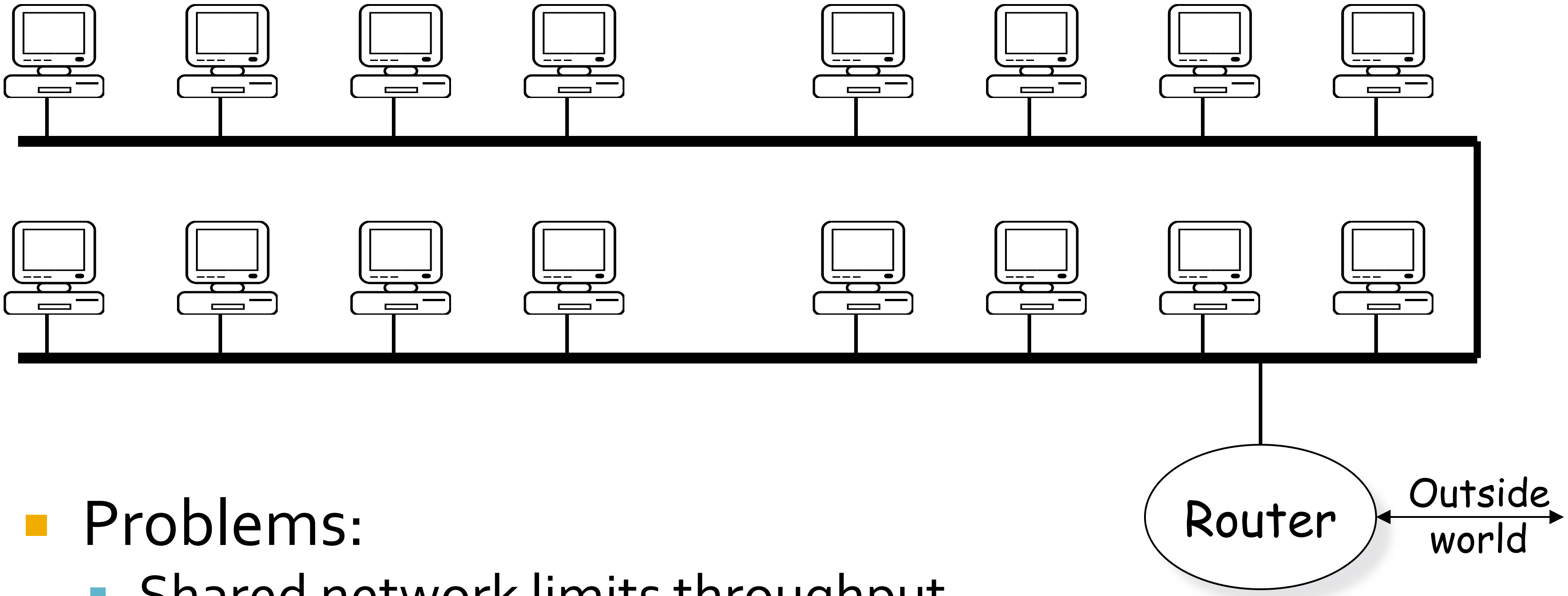
- Encoding
  - Encoding formats grow in sophistication as clock rate increases and stresses physical limits of copper/fiber media
  - 5-level Pulse Amplitude Modulation
  - 4-D 8-State Trellis Forward Error Correction Encoding
- Framing – **Same format**
- Error Detection
  - CRC still used at high frame level
  - Encoding method has reserved illegal symbols that automatically indicate error (noise / corruption) if received
- Media Access Control
  - Point-to-point links remove need for CSMA / CD protocol (but it remains for backwards compatibility)

New network topologies

Scaling Ethernet

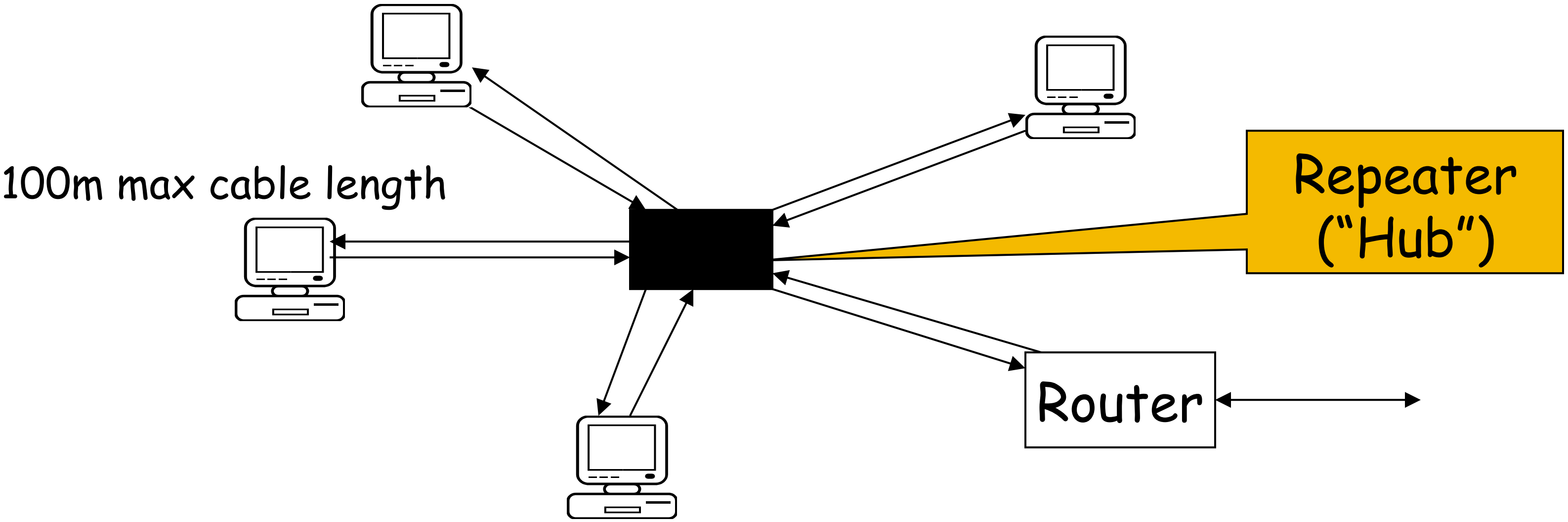
Original Ethernet Network

(10Base-5 or 10Base-2 – Shared Bus Architecture)



- Problems:
  - Shared network limits throughput
  - Frequent collisions reduce efficiency
  - Poor Reliability – Failure at one node can break shared link

Ethernet Star Topology

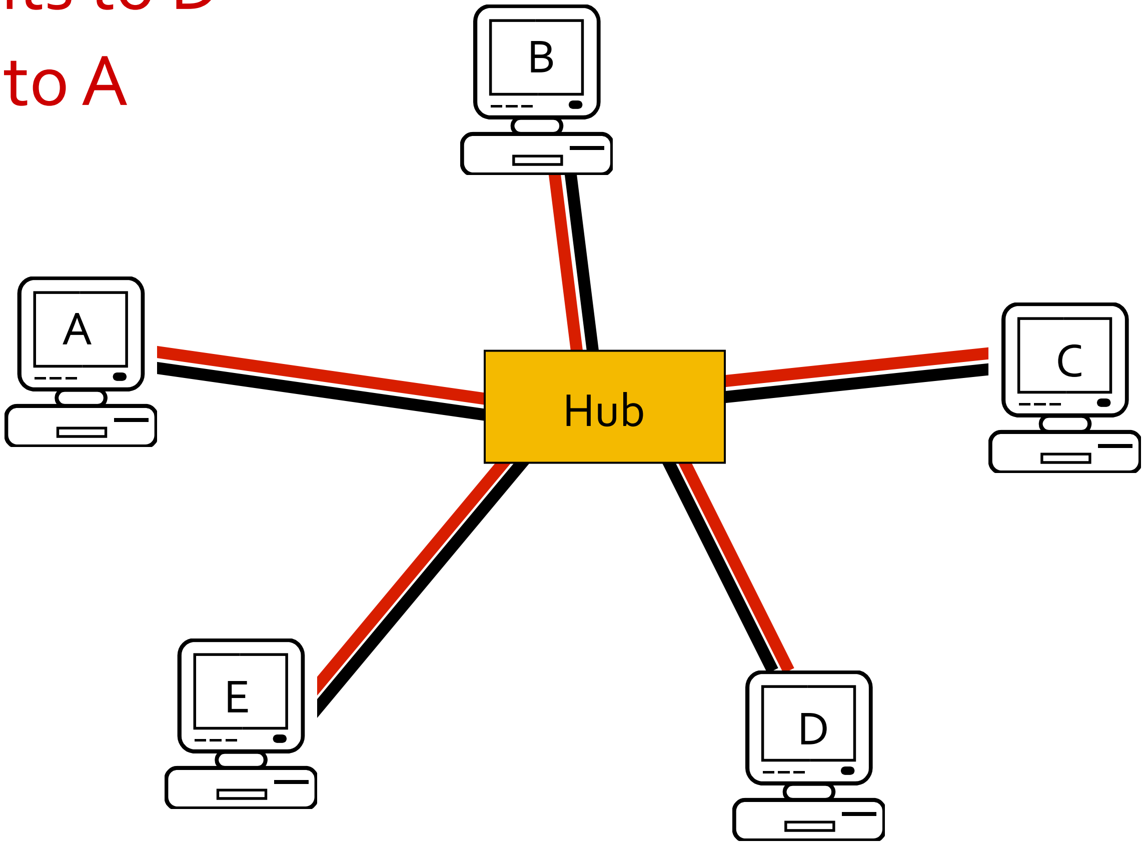


- Direct links instead of shared bus
- MAC protocol still operates as if Ethernet was a single wire
  - Collisions still possible
  - Network still shared
- Increase reliability from wire failure

Ethernet Hub - Operation

A transmits to D

D replies to A



Problems with Ethernet Hub

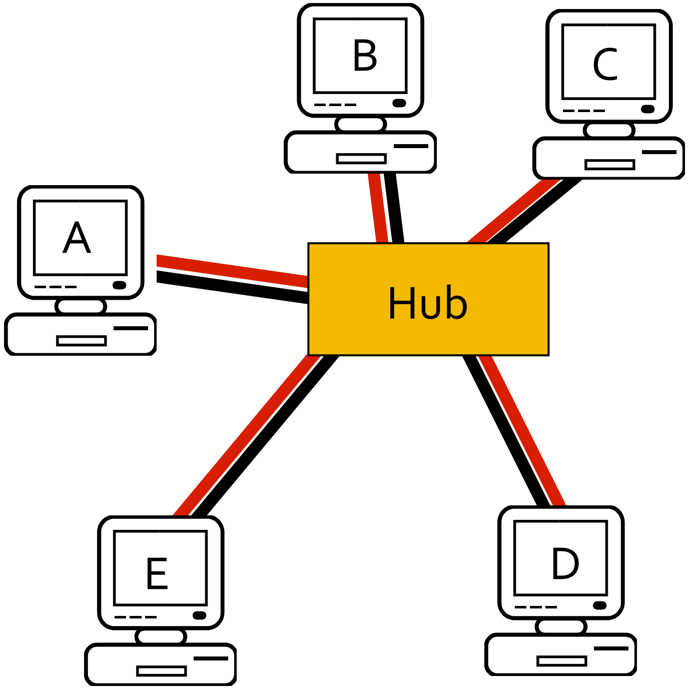
- Security concerns with broadcasting
- Performance
  - Unnecessary broadcasts waste network capacity and cause congestion
  - Communication is serialized – Independent connections between independent devices cannot occur in parallel
- Shared bus architecture limits maximum length of network
  - Due to MAC CSMA algorithm and signal propagation across entire network

Ethernet Switch

- New solution – Bridges (aka Ethernet **switches**)
  - Allow multiple hub-based networks to be partitioned and interconnected
  - Reduces collisions
  - Allow parallel communication between independent devices
  - Allow full duplex communication between multiple pairs of devices

Ethernet Hub vs Switch

Ethernet Hub

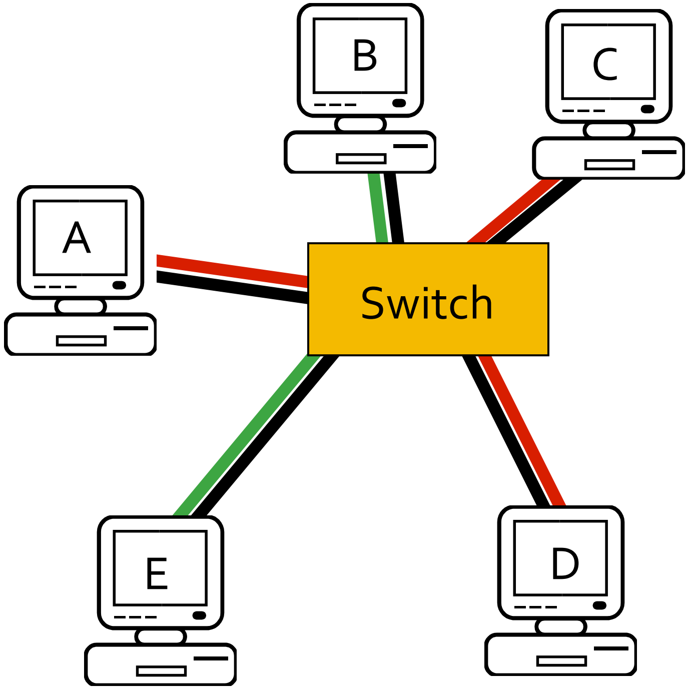


A transmits to D

D replies to A

Ethernet Switch

(assume learning already occurred)



A transmits to D

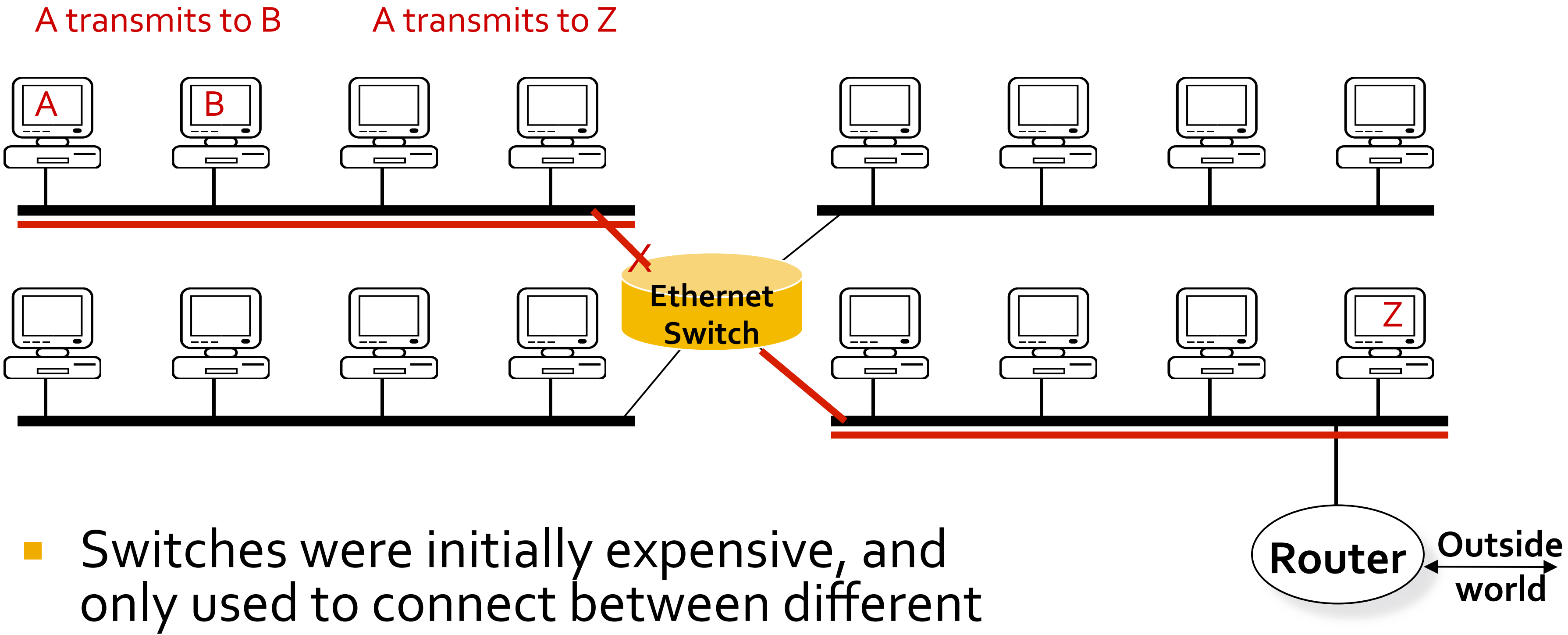
D replies to A

E transmits to B,

and A to C

Animation from http://www.datacottage.com/nch/eoperation.htm

Combining Hubs and Switches



- Switches were initially expensive, and only used to connect between different hub-based networks
- As cost decreased, hubs have been removed entirely
  - Gigabit+ networks are always switched
  - No more collisions!

Switch Design

- Internal FIFOs on each port buffer incoming packet
- Forwarding options
    - *Store-and-Forward*
        - Buffer entire packet before sending it to output port
        - Can verify packet CRC
    - *Cut-Through*
        - Buffer only long enough to examine destination address and then immediately stream data through to output port
            - Will fall back to store-and-forward if output port is busy
        - Cannot validate packet – By the time error is detected, it is too late!

Ethernet Details

Six Design Challenges

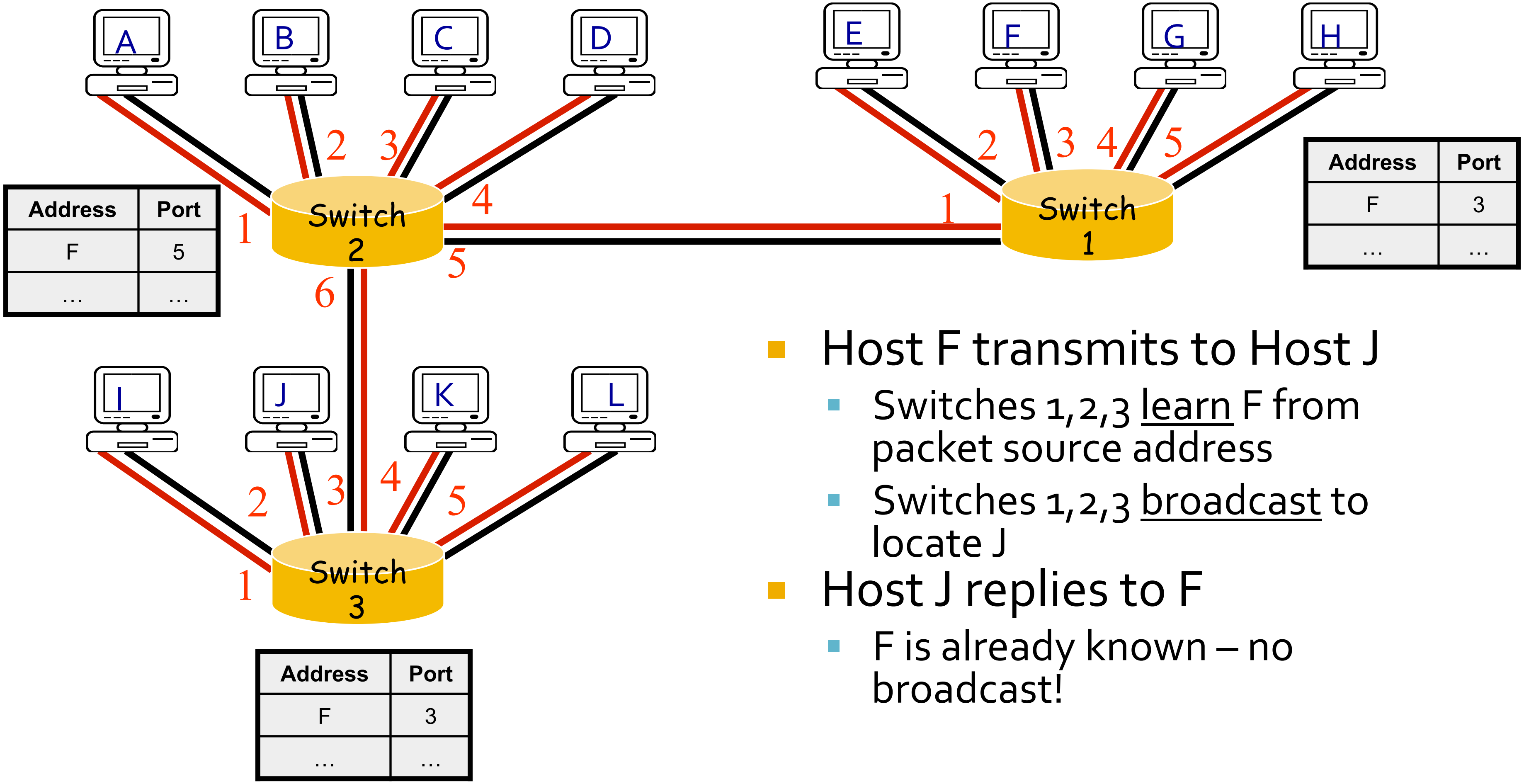
Challenges for Ethernet Switch

1. *Forwarding* – Where does the next packet go?
2. *Migration* – What if devices move on the network?
3. *Congestion* – What if too much traffic is received?
4. *Preventing Loops* – How to avoid forwarding packets in a big loop?
5. *Configuration* – How to determine speed of every device connected to switch
6. *Isolation* – How to isolate devices from each other (i.e. student computers from faculty computers)

Challenge 1 – Forwarding Packets

- Basic operation of Ethernet Switch
  - Examines header of each arriving frame
  - Learn that Ethernet SA is accessible from arriving port and update forwarding table
  - Examine Ethernet DA and search *Forwarding Table* on the switch
    - If in table, forward frame to the correct output port(s)
    - If not in table, broadcasts frame to all ports (except the one through which it arrived)

Switches - Learning Addresses



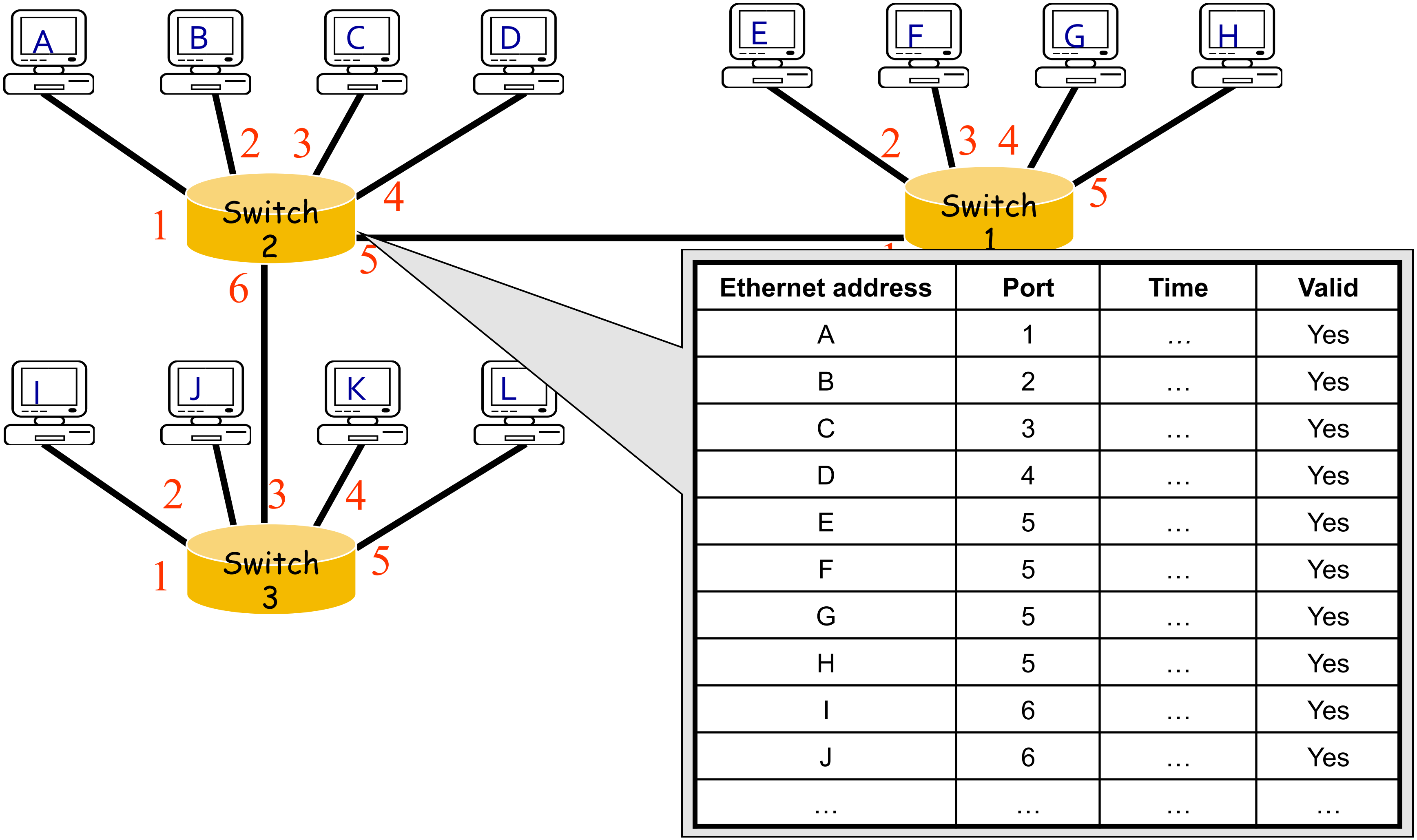
Address	Port
F	5
…	…

Address	Port
F	3
…	…

Address	Port
F	3
…	…

- Host F transmits to Host J
  - Switches 1,2,3 learn F from packet source address
  - Switches 1,2,3 broadcast to locate J
- Host J replies to F
  - F is already known – no broadcast!

Switches – Forwarding Table



Ethernet address	Port	Time	Valid
A	1	…	Yes
B	2	…	Yes
C	3	…	Yes
D	4	…	Yes
E	5	…	Yes
F	5	…	Yes
G	5	…	Yes
H	5	…	Yes
I	6	…	Yes
J	6	…	Yes
…	…	…	…

Forwarding Table Capacity

- At NewEgg in 2012:
  - $25 Rosewill gigabit switch – 8192 devices
  - Switches from better vendors: 16384 devices
  - Table capacity is rarely advertised anymore (all devices are "sufficient")
- Capacity is not infinite, but 16k+ devices is a very large network without routing
  - Except, perhaps, for a large cluster computer…

Forwarding Table Maintenance

- How to remove stale entries from the table? (e.g. device leaves the network)
  - Entries expire if no communication from device within last epoch
  - 5 minute timer is default on Cisco switches

Forwarding Table Maintenance

- What if the table is full? What entry do we remove to make room for a new one?
  - Round-robin (oldest device)
    - Pros: Simple!
    - Cons: Oldest entry might be very active device
  - Least-Recently Used
    (e.g. device that last transmitted a packet a long time ago)
    - Pros: High effectiveness (device not likely to transmit again soon)
    - Cons: Complicated – Switch must count # of packets per device, and sort/search the table to determine LRU device
  - None – Don't learn that device until a table entry expires normally. Until then, broadcast any packets destined to it
    - Pros: Simple. Ensures old (but active) devices are not evicted
    - Cons: If new devices is high traffic, entire network will suffer (due to broadcasts) until there is space in forwarding table
    - Used by Cisco switches

Challenge 2 - Migration

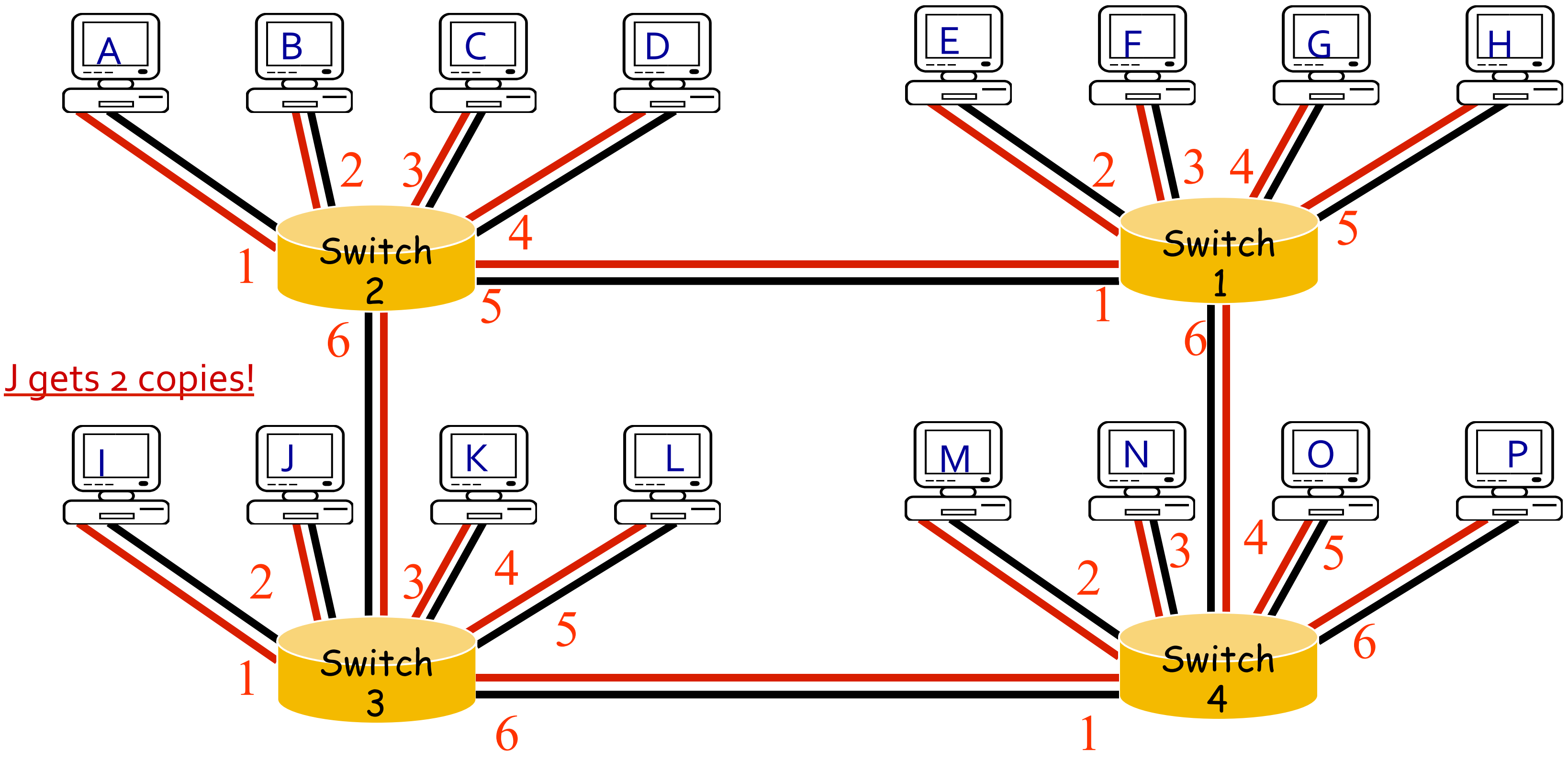
- What if a network device (e.g. laptop computer) moves from one port to another? (on same switch)
    - Data is forwarded to wrong port until either:
        - Forwarding table entry expires
        - Device transmits a packet, and switch learns new port
- What if the device moves from one switch to another?
    - Have to wait for entry on old switch to expire (unless device happens to send a packet through that switch)

Challenge 3 - Switch Congestion

- What happens if the switch is too busy?
  - Example: Traffic from 10 input ports all heading out single output port
- Easiest solution
  - Switch drops traffic as internal buffers overflow
  - Devices don't know and keep transmitting!
    - A higher level protocol such as TCP might eventually notice and throttle back…
- Can we do better?
  - *Ethernet flow control could throttle sender (only works across 1 wire, <u>not</u> end-to-end across Internet!)*

Broadcasting on Topology with Redundant Paths

- Host F sends message to Host J



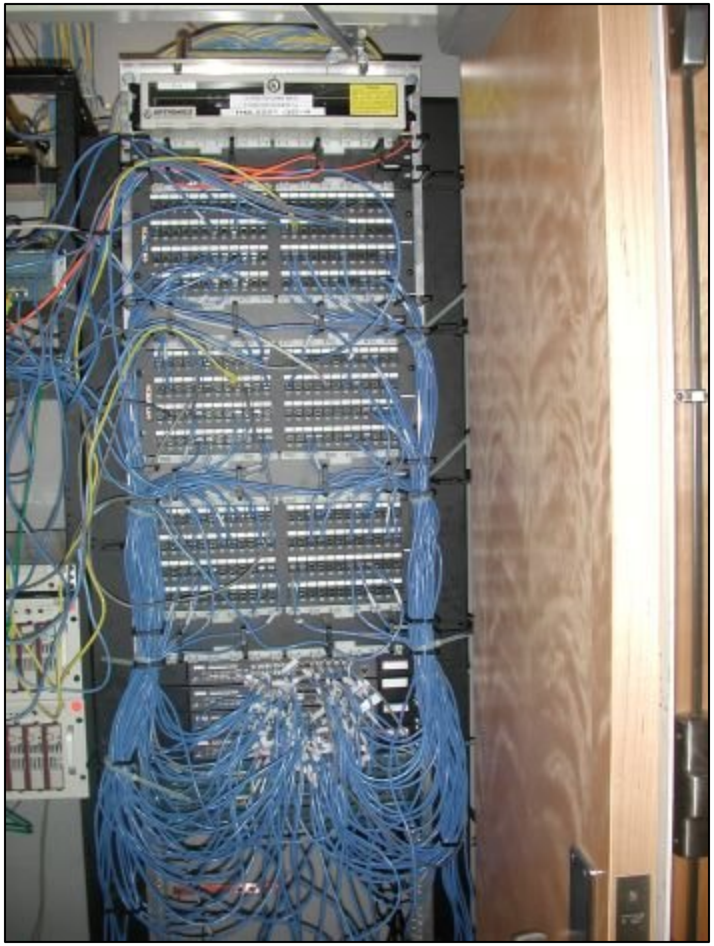
Challenge 4: Problems with Loop Topology

- Broadcast Storm
  - Packets are forwarded forever
  - Ethernet has no time-to-live field
- Forwarding Table Oscillation
  - Packets from host are received via multiple ports. Table is constantly updated

Photo from http://www.extremeinstability.com/

Topology Challenge – Loops!

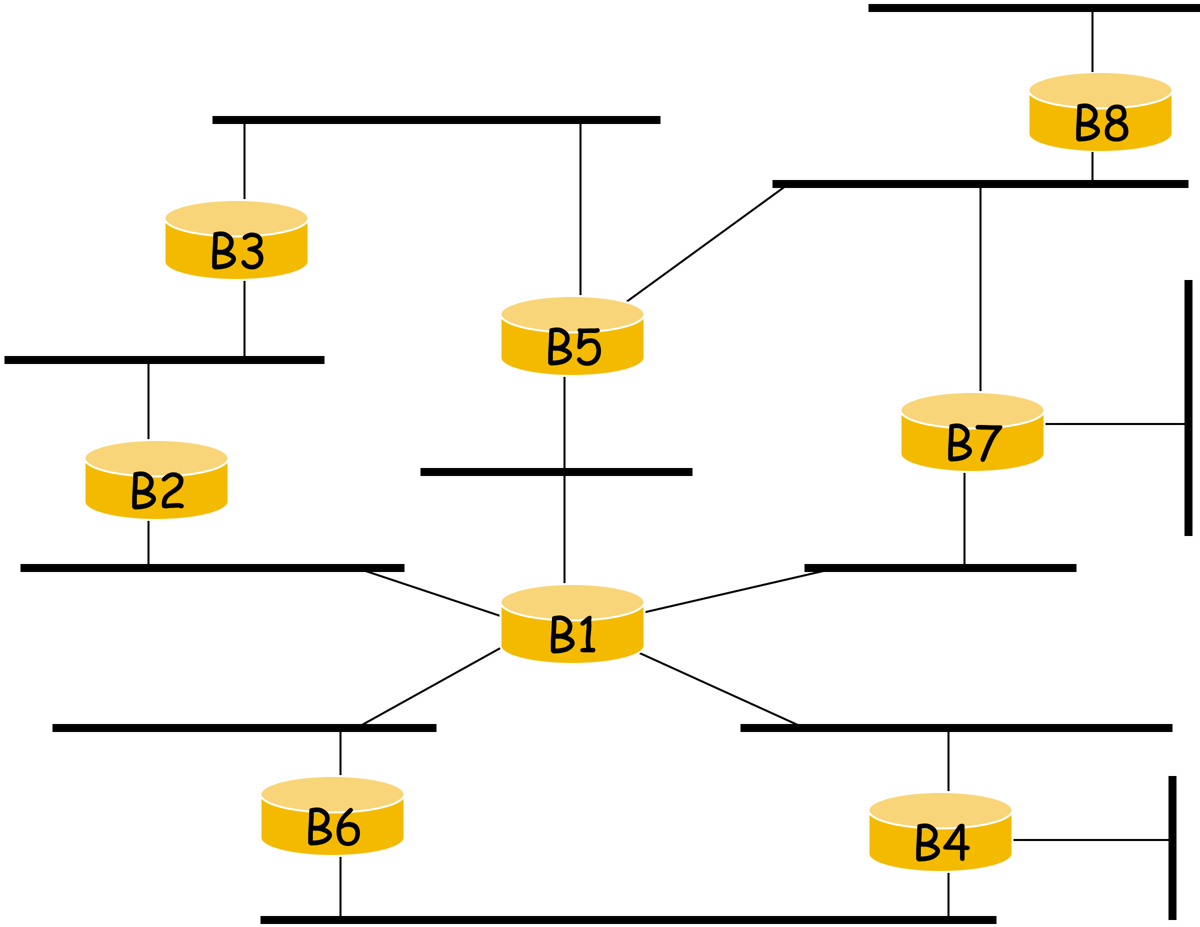
- Can't we just avoid creating loops?
  - Redundant paths are useful for reliability
  - What if a loop is accidentally created? (Have you *seen* some of these wiring closets?)



Spanning Tree Protocol (IEEE 802.1D)

- Principles
  - Raw network is a mesh / graph
  - Create a tree from this mesh
    - Tree is a subgraph that spans all the vertices (switches) without loops
  - Disable all links not part of the tree - prevents loops!
- Features
  - Decentralized – Switches communicate among themselves via Bridge Protocol Data Units
  - Automatic – No user configuration required
  - Fault tolerant – Spanning tree will adapt if links fail (and can automatically use redundant links that were previously disabled)

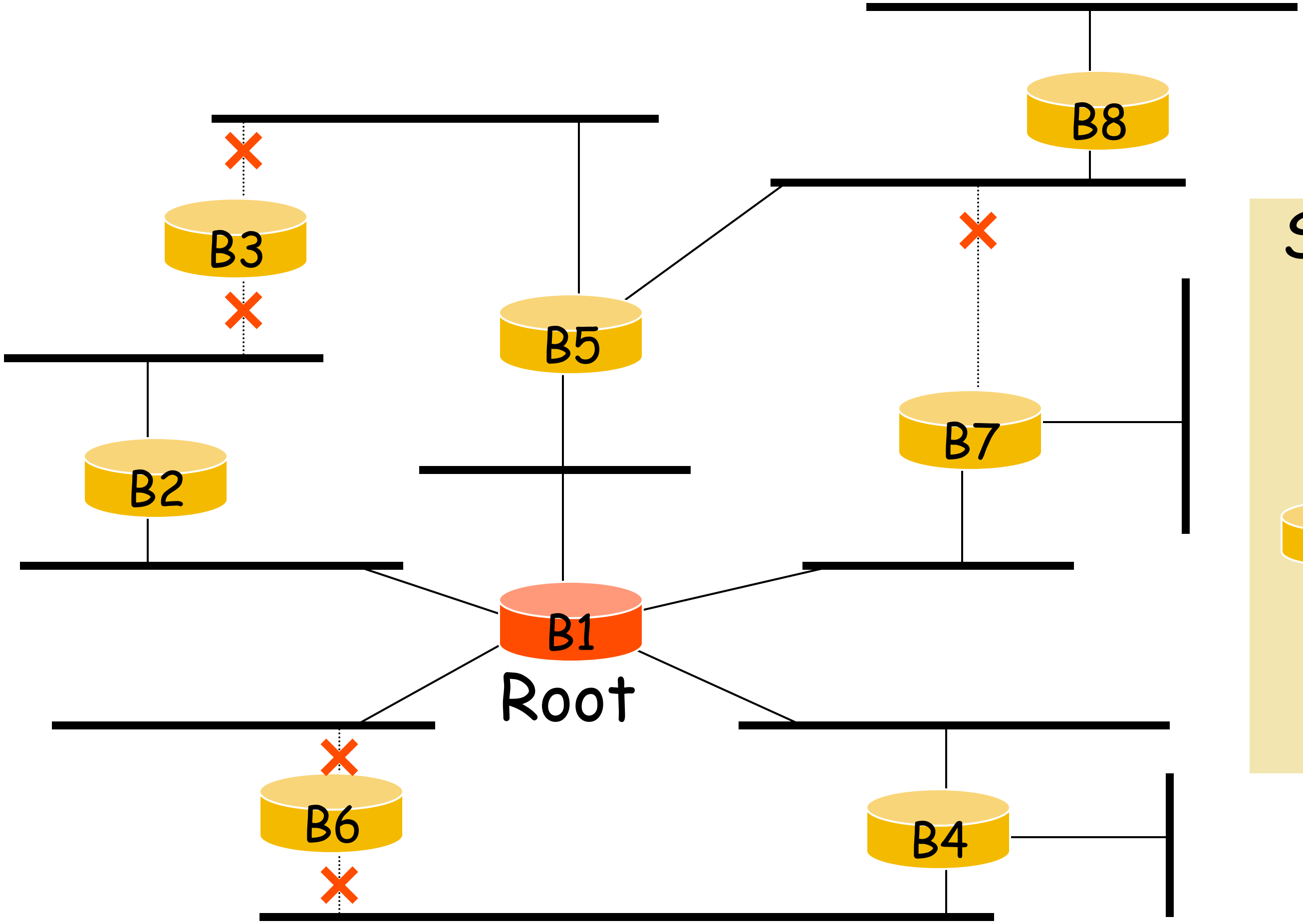
Example Spanning Tree



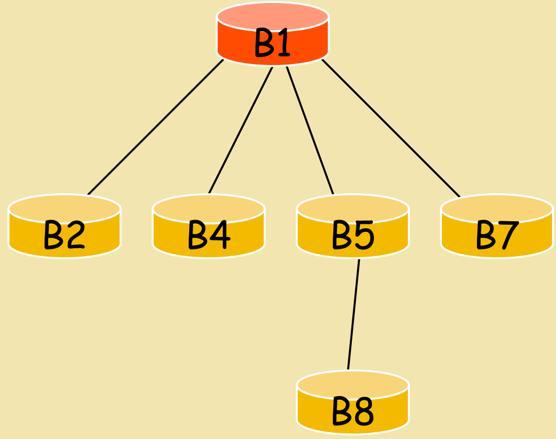
Protocol operation:

1. Pick a root. The root forwards over all its ports.
2. For each segment, pick a designated switch that is closest to the root.
3. All switches on a segment send packets towards the root via the designated switch.

Example Spanning Tree



Spanning Tree:



Spanning Tree Issues

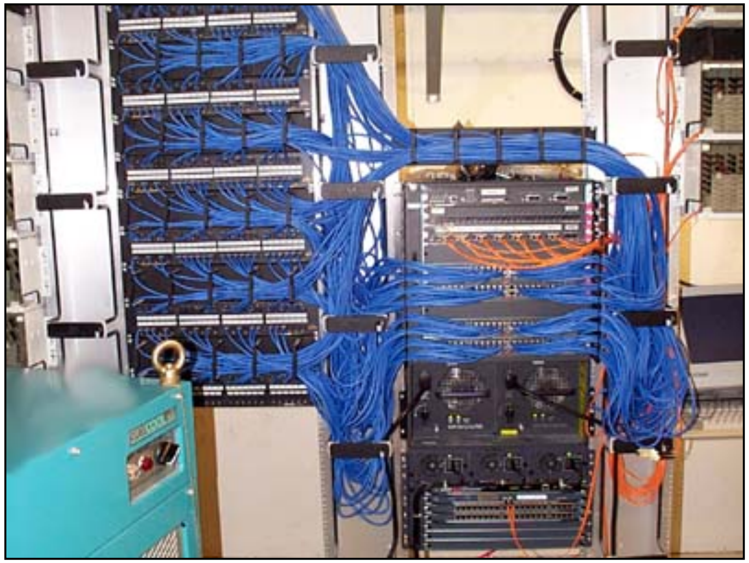
- Spanning Tree is not guaranteed to be a minimum spanning tree
  - Packets might take a longer path than necessary
  - Root switch might not be anywhere near “center” of network
- Solution?
  - Manual tweaking – Administrators can adjust device IDs to force different root

Challenge 5 – Switch Configuration

- Problem – Each port on the switch might connect to a device running at different speed (10, 100, 1000Mbps) or duplex setting
- Do we want to configure each device manually?
  - Of course not.
- Solution: **Auto-Negotiation**
  - Upon power-up, each network device sends custom signals across link to other end announcing its capabilities
  - Each device listens and picks the highest mutually supported transmission mode
  - Format is backwards compatible down to 10Base-T, half duplex
- Modern switches have internal FIFOs that can buffer data between devices with varying performance capabilities
  - 1Gbps device → 100Mbps device – flow control useful!

Challenge 6 – Device Isolation

- Imagine I have a campus network, and want to isolate a few devices on a "private network". How do I do it?
  - Buy more switches?
  - *Could get expensive…*
  - *Imagine the mess in the wiring closet…*



Challenge – Device Isolation

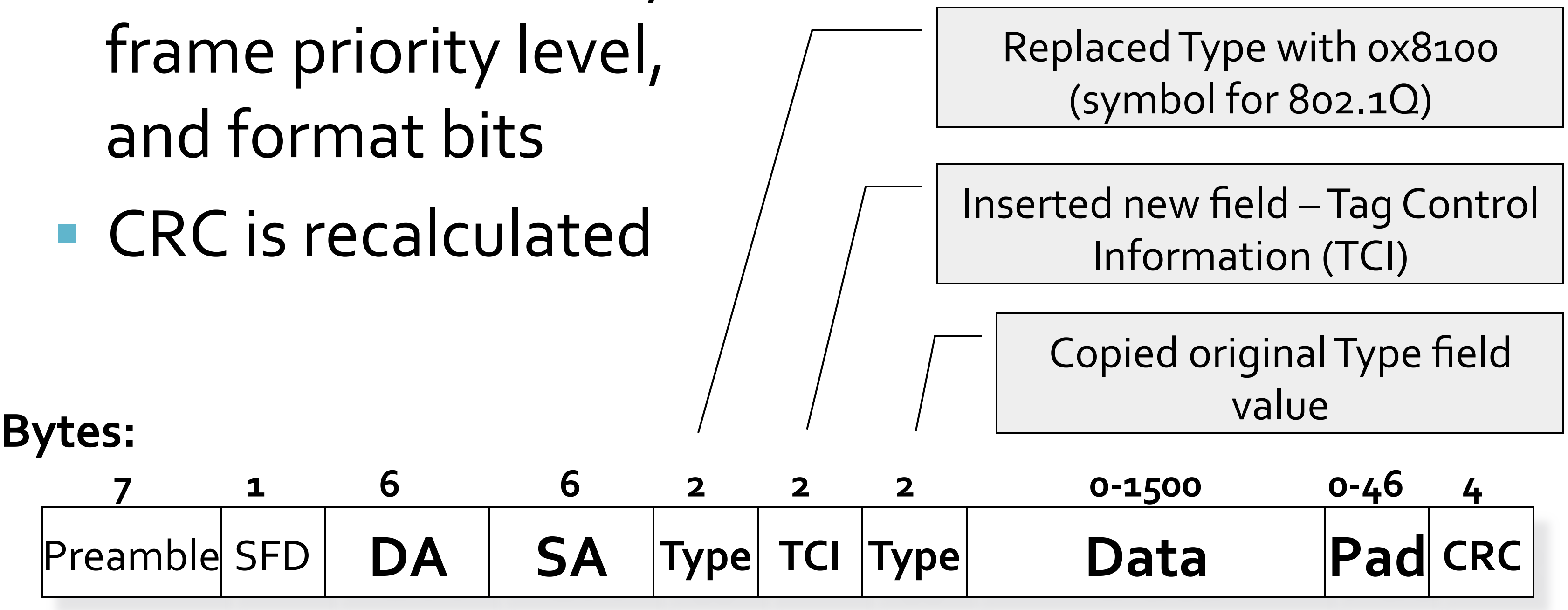
- Better idea – Make the switch more intelligent and have it provide device isolation
- Virtual LAN (VLAN) technology
  - Virtualizes the network – Each network device on a VLAN communicates as if they were connected to the same physical network, even if they are not
  - Can create a virtual LAN composed of machines from around the world

VLAN Overview

- Controlled by network switch
  - Each port is mapped to a VLAN
  - Forwarding / broadcast is only allowed to other ports on the same VLAN (provides isolation)
  - Spanning Tree Protocol can be run independently over each VLAN
    - Might even have different topology!
- Joining VLAN – How to assign devices?
  - Static – Port is permanently mapped to VLAN
  - Dynamic – Based on MAC address or user authentication (e.g. Cisco CleanAccess)

VLAN Operation

- Standardized format: IEEE 802.1Q
  - TCI stores VLAN ID, frame priority level, and format bits
  - CRC is recalculated



Replaced Type with 0x8100 (symbol for 802.1Q)

Inserted new field – Tag Control Information (TCI)

Copied original Type field value

Bytes:

7	1	6	6	2	2	2	0-1500	0-46	4
Preamble	SFD	DA	SA	Type	TCI	Type	Data	Pad	CRC