

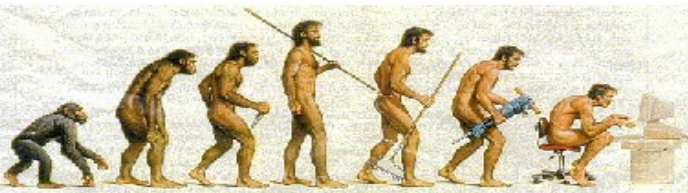
The Ethernets: Evolution from 10 to 10,000 Mbps – How it all Works!

N+I Atlanta 2000

Workshop W924

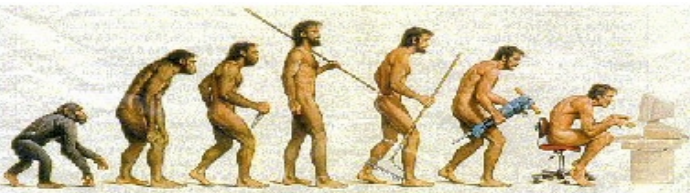
Hadriel Kaplan & Bob Noseworthy

**NETWORLD™
+ INTEROP®**



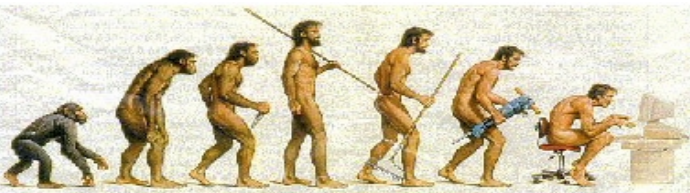
Who Are We?

- Robert Noseworthy
 - Managing Engineer, University of New Hampshire InterOperability Lab (UNH IOL)
 - Active member, 802.3 Working Group
 - Developed early Fast Ethernet and Gigabit Ethernet test devices
 - Part of team that built the first multi-vendor Fast Ethernet network
- Hadriel Kaplan
 - Team Leader, Nortel Networks Advanced Interoperability Lab, Technology Center
 - Active member, 802.3 Working Group
 - Led the team that built the first multi-vendor Gigabit Ethernet and 802.1Q networks



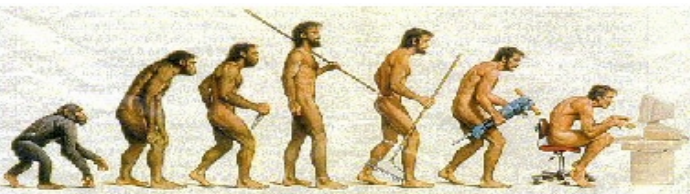
What Will You Learn?

- LOTS. Too much, in fact.
- We'll teach you about what you need to know to understand, troubleshoot, and design Ethernet networks.
- We'll discuss the common problems, work-arounds, and issues in Ethernet hardware.



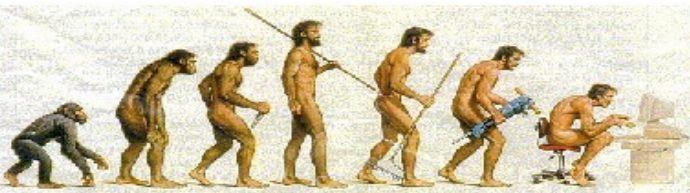
What Won't You Learn?

- Pricing - we don't know, we don't care.
- Specific product features - we're not here to sell.
- The following technologies:
 - ATM
 - QoS
 - IPv6
 - VoIP
 - DWDM
 - OSPFThose are all other workshops...
- How to design a real, complete network. (we'll cover Ethernet and switching, but not routing)



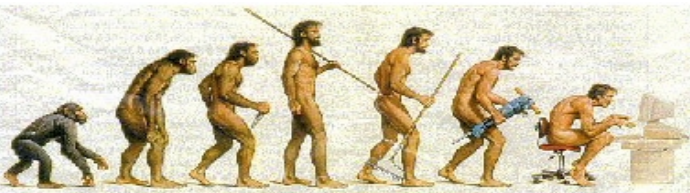
Outline

- Ethernet Essentials
- Media
- Core PHYs
- Auto-Negotiation
- Future Ethernet
 - DTE Power via MDI
 - 10 Gigabit Ethernet
- Switched Network Design
 - Spanning Tree
 - Link Aggregation
 - VLANs
 - And more...



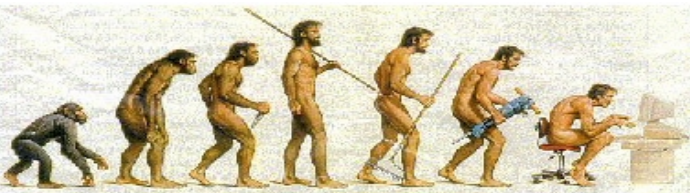
Ethernet Essentials - Outline

- Ethernet History
- Ethernet Standards
- Ethernet Frame
- Half Duplex MAC
- Repeaters
- Full Duplex MAC
- Switches
- Flow Control



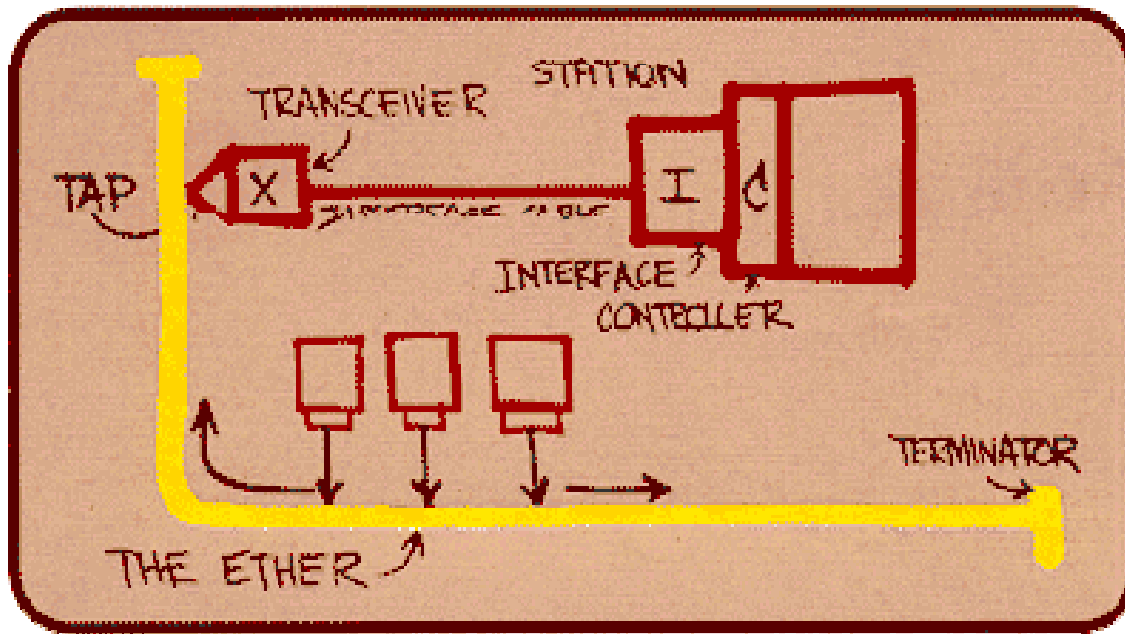
Ethernet History

- Why is it called Ethernet?
 - “In late 1972, Metcalfe and his Xerox PARC colleagues developed the first experimental Ethernet system to interconnect the Xerox Alto, a personal workstation with a graphical user interface. The experimental Ethernet was used to link Altos to one another, and to servers and laser printers. The signal clock for the experimental Ethernet interface was derived from the Alto's system clock, which resulted in a data transmission rate on the experimental Ethernet of 2.94 Mbps.
 - Metcalfe's first experimental network was called the Alto Aloha Network. In 1973 Metcalfe changed the name to "Ethernet," to make it clear that the system could support any computer-not just Altos-and to point out that his new network mechanisms had evolved well beyond the Aloha system. He chose to base the name on the word "ether" as a way of describing an essential feature of the system: the physical medium (i.e., a cable) carries bits to all stations, much the same way that the old "luminiferous ether" was once thought to propagate electromagnetic waves through space. Thus, Ethernet was born.”

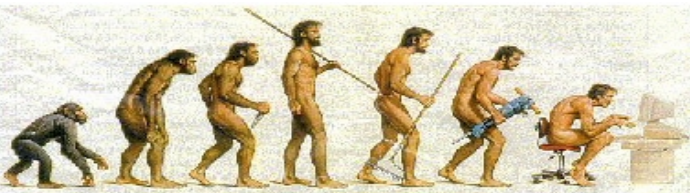


Ethernet History

- Robert Metcalf's Idea

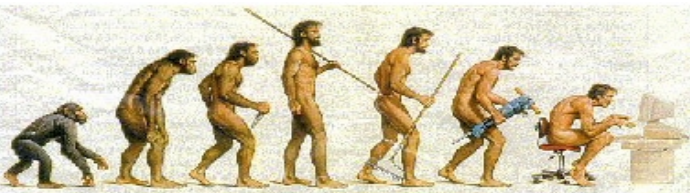
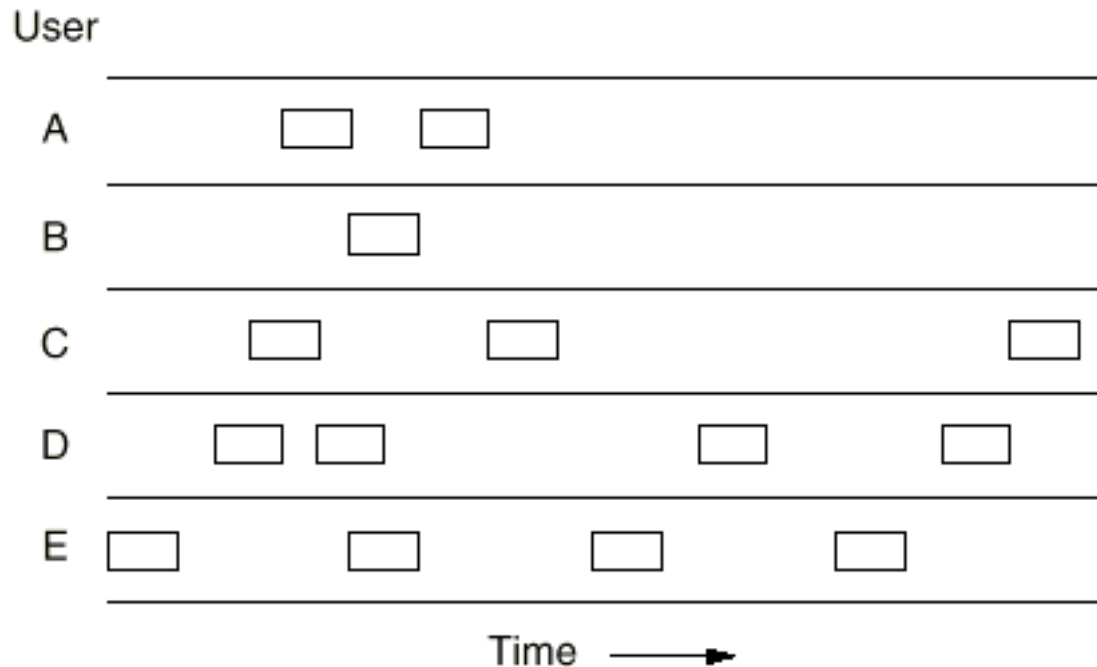


- Invented by Metcalf at Xerox in 1973 and patented in 1976
- Xerox convinced Digital and Intel to join in making products (hence the group called DIX)
- IEEE standard in 1989



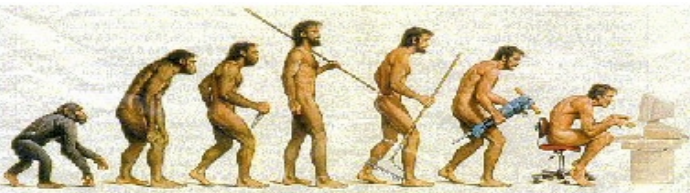
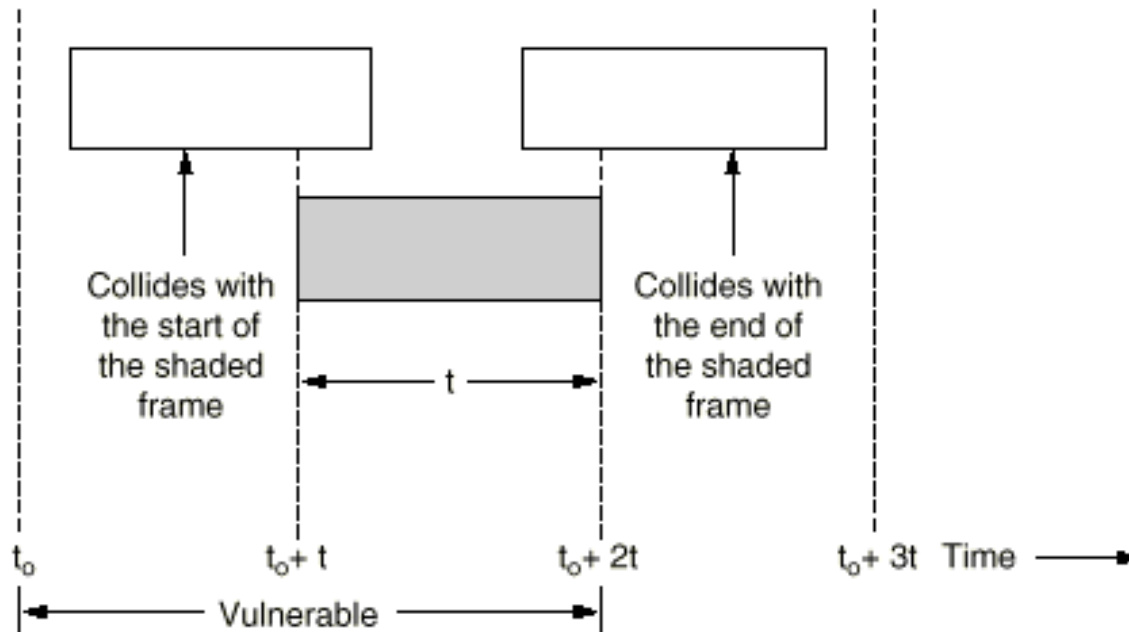
Ethernet History: Pure ALOHA

- Transmit when you want to, regardless of others.



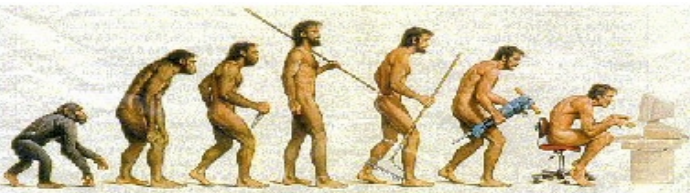
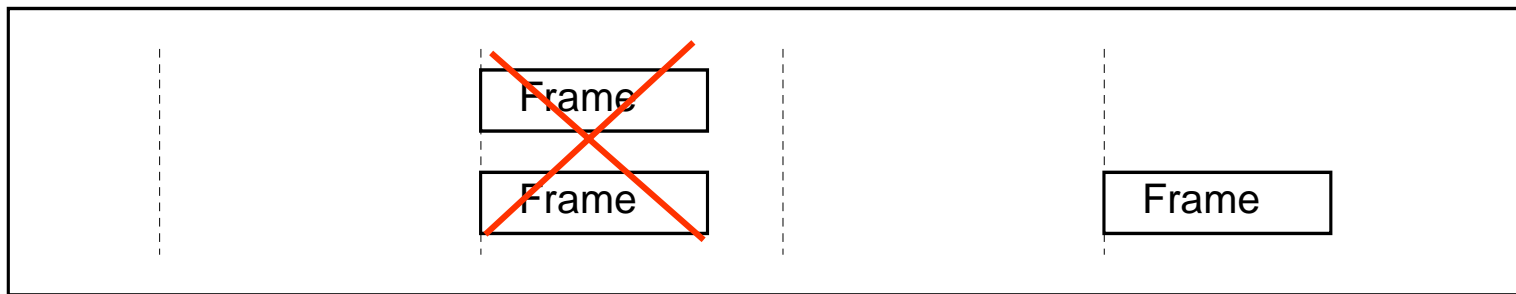
Ethernet History: Pure ALOHA Collisions

- Extremely inefficient, since the worst-case period of vulnerability is the time to transmit two frames.



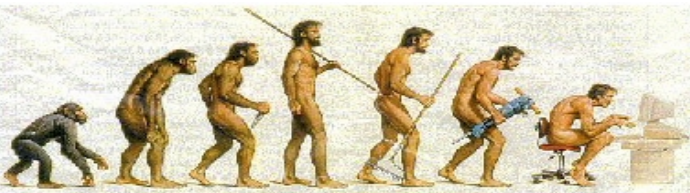
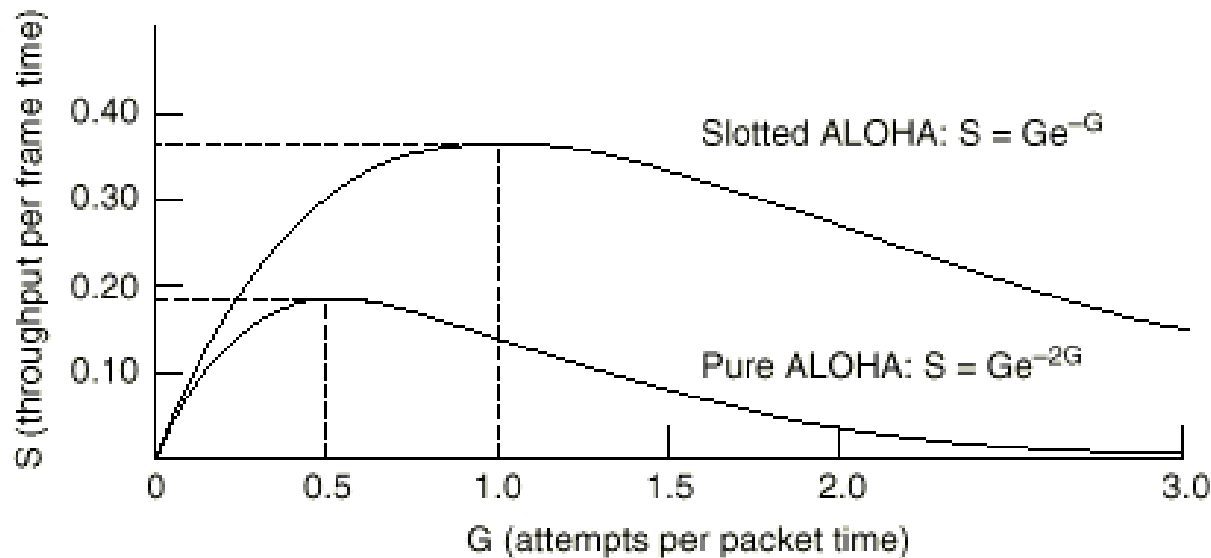
Ethernet History: Slotted ALOHA

- Transmit only at the beginning of synchronized “slot times”
- Collision inefficiency limited to one frame transmission time



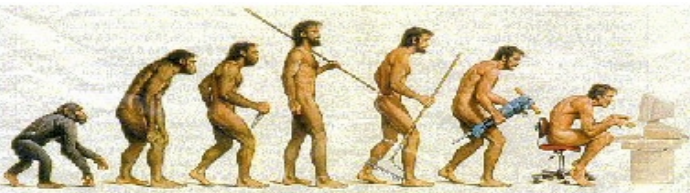
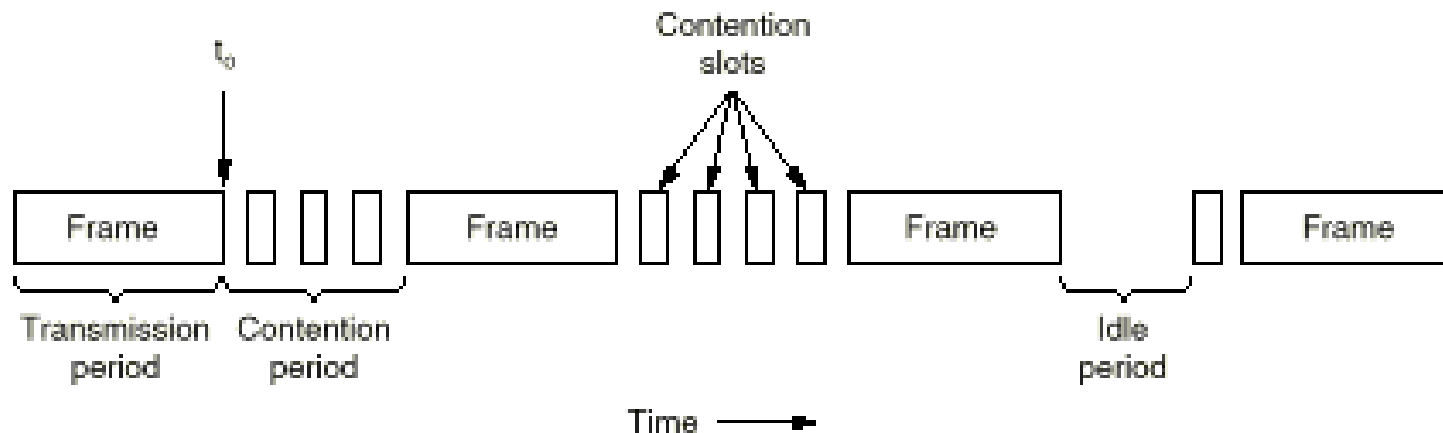
Ethernet History: ALOHA v Slotted ALOHA

- Throughput efficiency increases dramatically for Slotted Aloha.



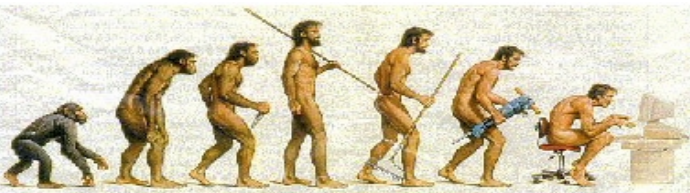
Ethernet History: CSMA/CD

- Take Slotted ALOHA to the next level, use the slots as “contention periods”.
 - If no collision occurs before the end of the period, then complete transmission of the frame.
- CSMA/CD can be in one of three states: contention, transmission, or idle. More on CSMA/CD later...



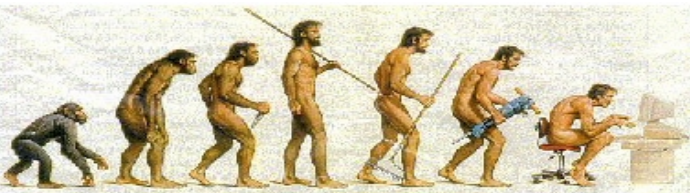
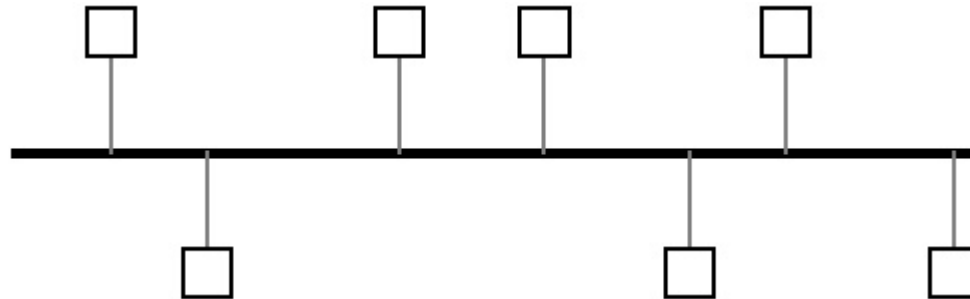
Ethernet History: Collisions

- Collisions
 - Two or more transmissions literally collided with one another on the same medium.
 - Result corrupts the data contents of the transmissions.
- Possible due to the medium used by the original Ethernet



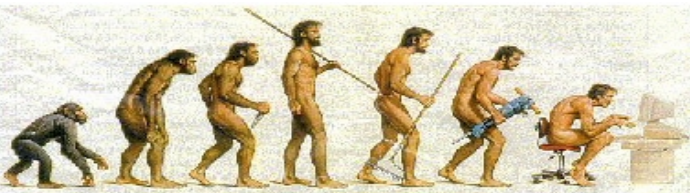
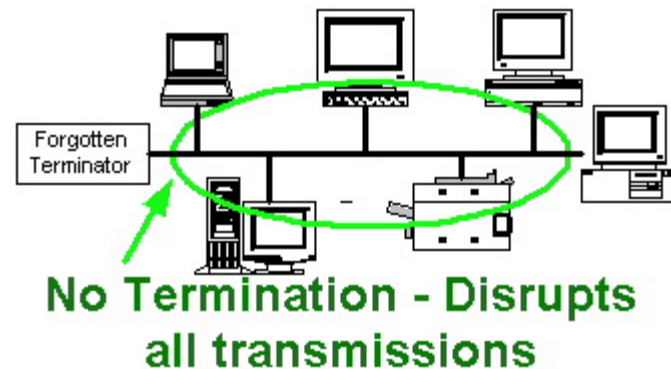
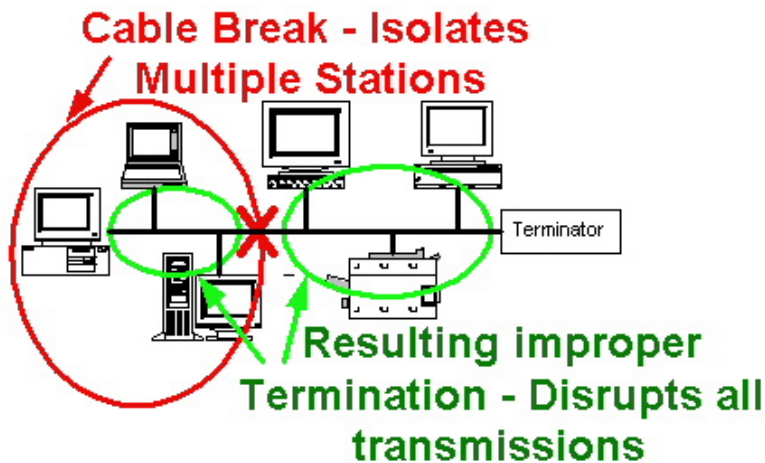
Ethernet History: The Shared Bus Topology

- Coaxial Cabling, 10 Mbps
 - 10BASE-5 “ThickNet”
 - 10BASE-2 “ThinNet”
- Bus Topology = Truly shared media



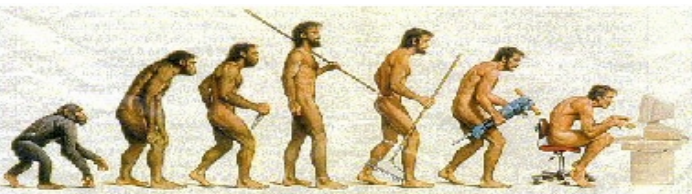
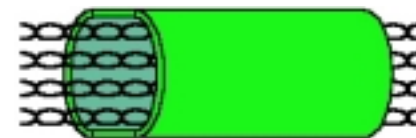
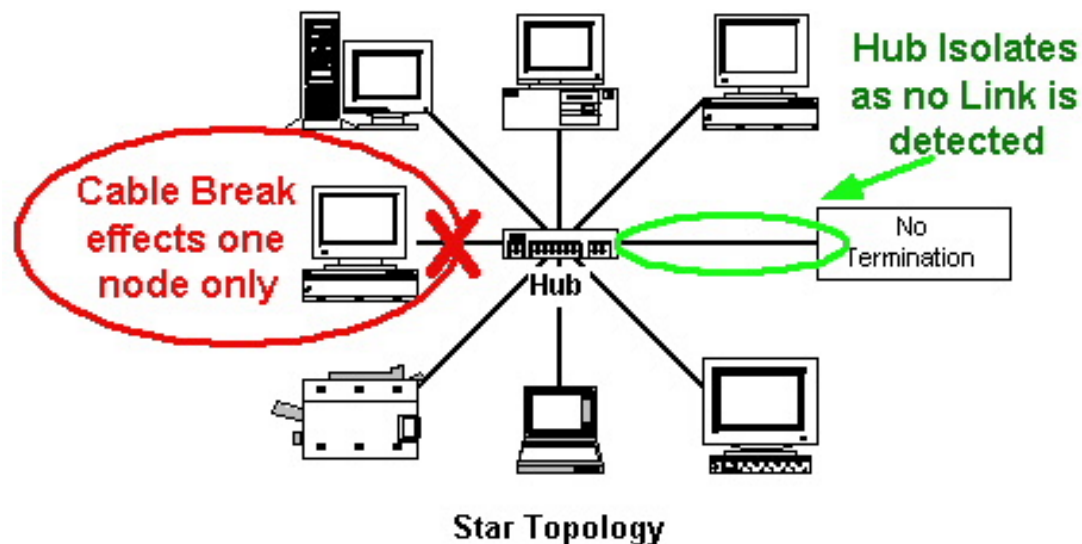
Ethernet History: Bus Topology Extinction

- Problems with Bus Topology
 - Break in Coax cable can sever service to multiple nodes
 - Fault in Coax cable can disrupt service to all nodes
 - » ground fault
 - » Incorrect termination
 - Adding/Removing nodes disrupts network



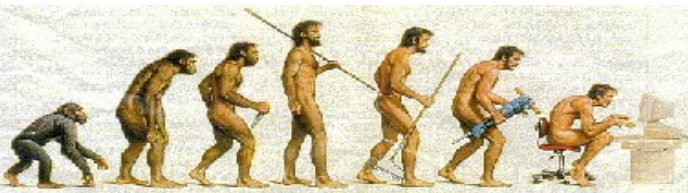
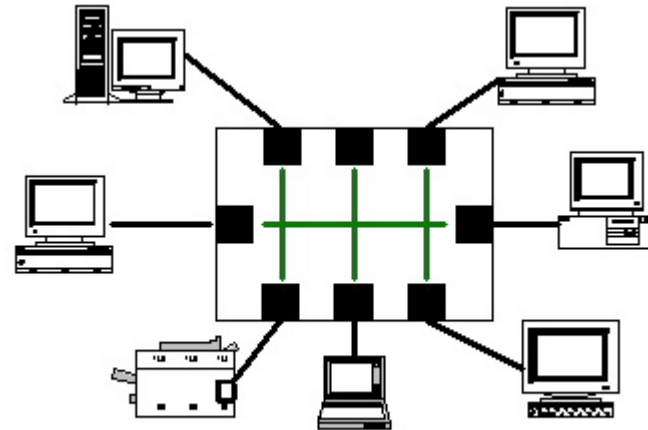
Ethernet History: Star Topology

- Bus Topology Evolved into a Star Topology
 - Driven by cabling issues
 - » Single cable breaks/faults effect only one node
 - » Emergence of cheap unshielded twisted pair (UTP) cable
 - Introduces “Hub” or “Concentrator” that isolates faulty nodes/cables



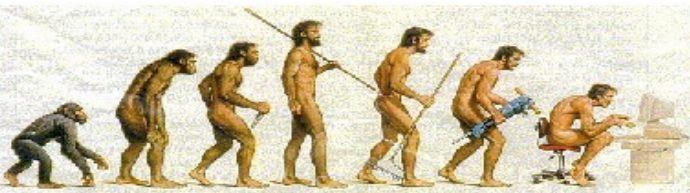
Ethernet History: Hub

- Hub can refer to either:
 - Repeater (“Bus in a Box”)
 - » Star Topology with Logical Bus
 - Switch / Bridge
 - » Still Star Topology:
Allows simultaneous transmissions between different stations



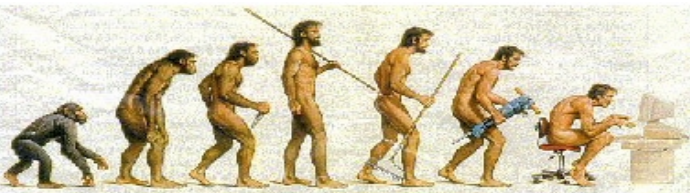
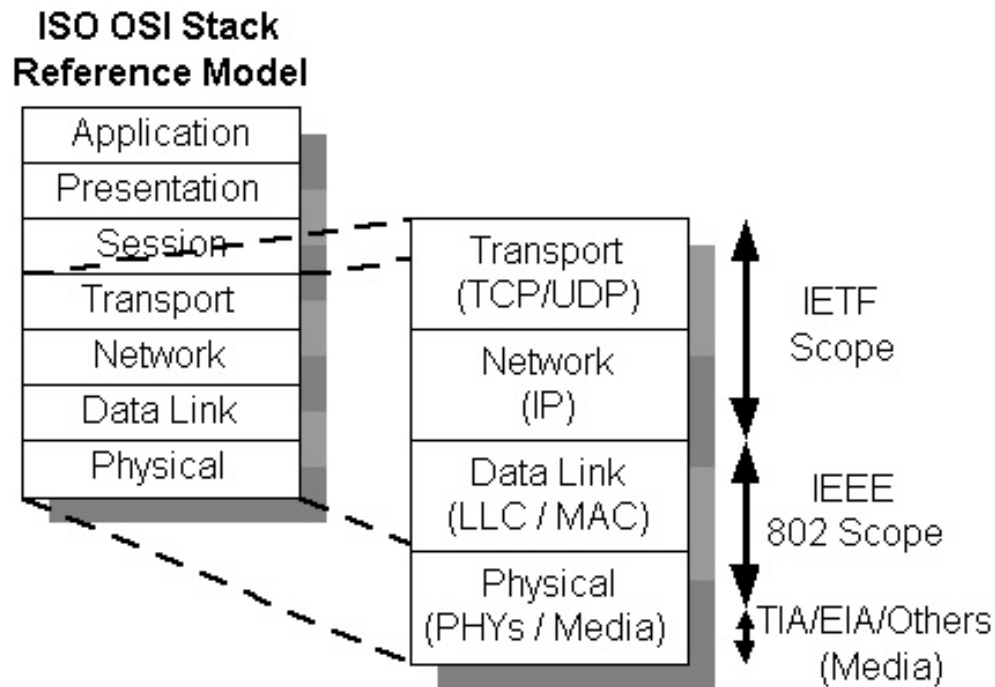
Ethernet History: Standardization

- First IEEE Ethernet Standard in 1985
- Standardization
 - Preferred over competing proprietary solutions
 - Creates a shared, open market for component and systems vendors
 - Defines interfaces and mechanisms to permit interoperable solutions
 - » Permits creation of heterogeneous (multi-vendor) networks

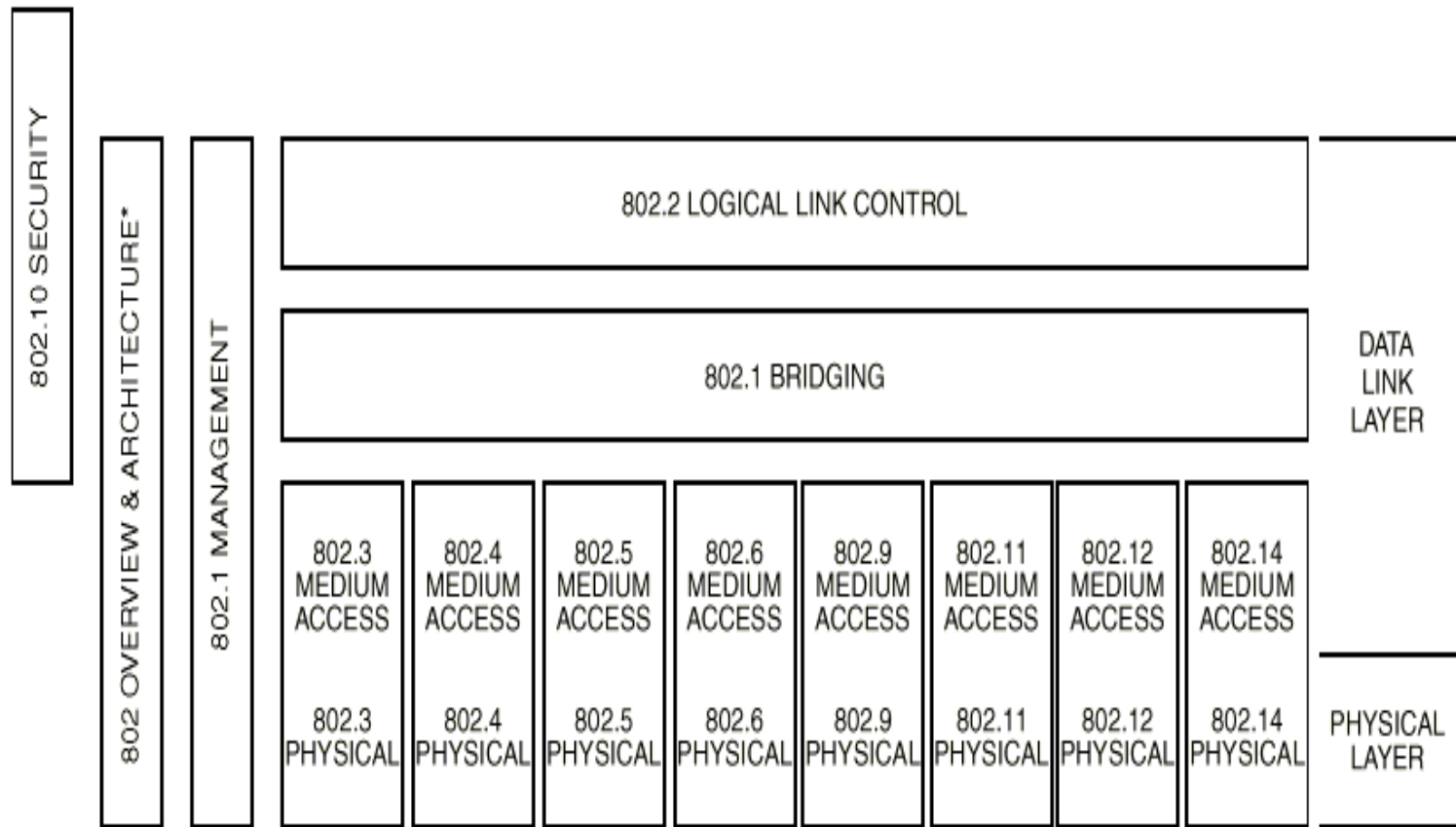


Ethernet Standards

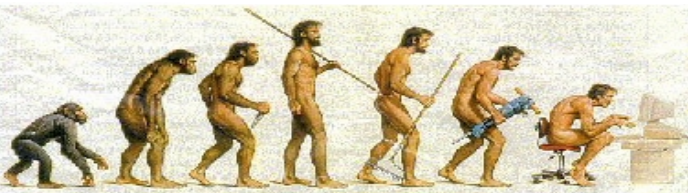
- Ethernet fits in the Open Standards Interface (OSI) model of the International Standards Organization as shown.



Ethernet Standards: IEEE 802 Architecture

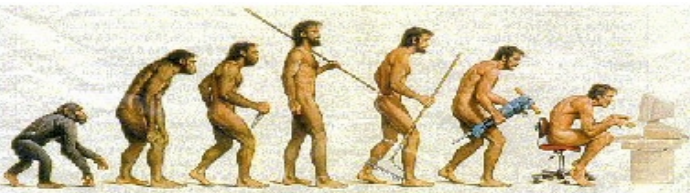


* Formerly IEEE Std 802.1A.



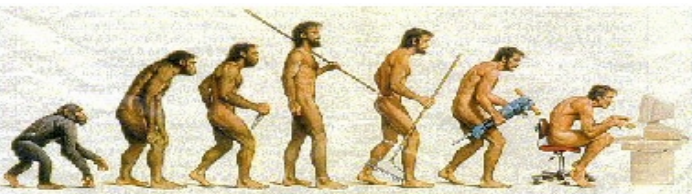
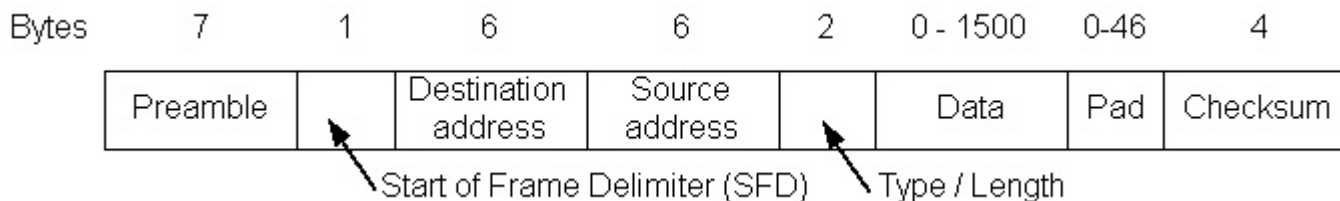
Ethernet Standards: IEEE 802.3

- 802.3 Now encompasses
 - Original 802.3: 10BASE-T 10BASE-5 10BASE-2 10BROAD-36
 - 802.3u Fast Ethernet: 100BASE-TX 100BASE-FX 100BASE-T4
 - 802.3x: Flow Control
 - 802.3z Gigabit Ethernet: 1000BASE-SX / -LX / -CX
- 802.3ab Copper Gigabit Ethernet: 1000BASE-T
- 802.3ac Frame Tagging for VLAN support
- 802.3ad Link Aggregation
- 802.3ae 10 Gigabit Ethernet: Completion by March 2002
- 802.3af DTE Power via MDI: Completion by Sept 2001



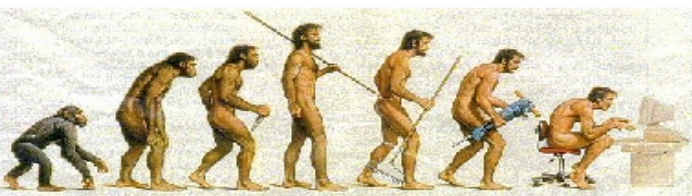
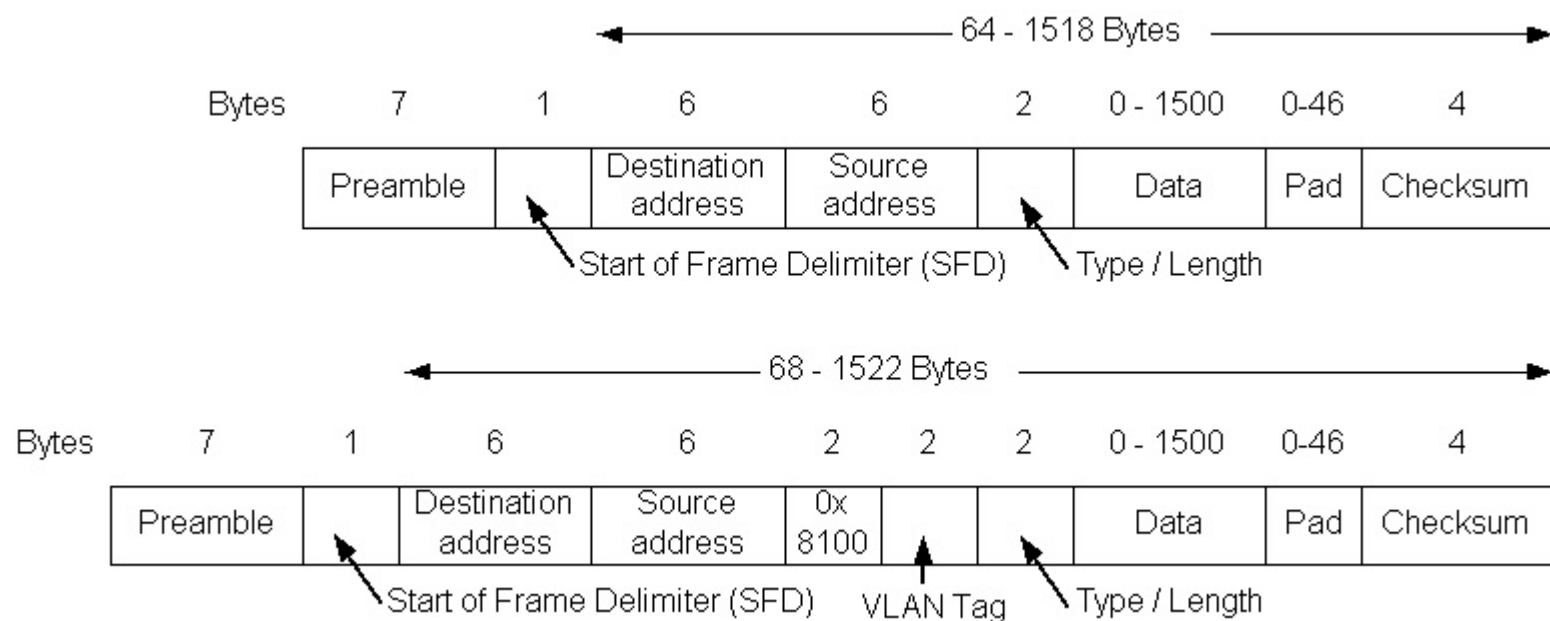
Ethernet Frame

- Transmitted Data is embedded in a container, called a frame
- This Frame Format DEFINES Ethernet
 - Historically, two types of frames existed:
 - » 802.3 Framing used a Length field after the Source Address
 - » Ethernet II (DIX) Framing used a type field after the Source Address
 - Now both frame types are defined and supported within IEEE 802.3



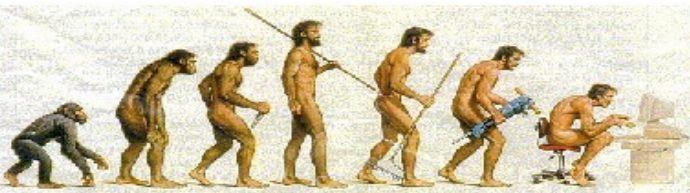
Ethernet Frame

- Frame size varies from 64 to 1518 Bytes except when VLAN tagged (more on that later...)



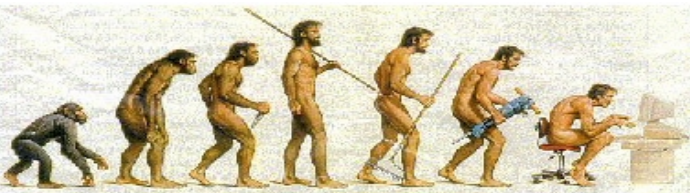
Ethernet Frame: Addresses

- Station Addresses
 - Must be unique on every LAN
 - Unless locally administered (uncommon), each node has a unique address assigned by the manufacturer. First 3bytes of address are assigned by IEEE (Organization Unique Identifier - OUI)
- Source Address: Must always be Station Address
- Destination Address: May be –
 - Unicast Address (Addressed to Station Address of one other station)
 - Multicast Address (Addressed to multiple stations simultaneously)
 - » Broadcast Address (FF FF FF FF FF FF) To All Stations



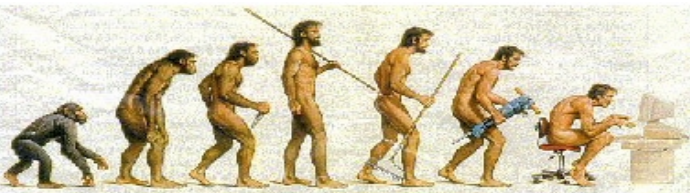
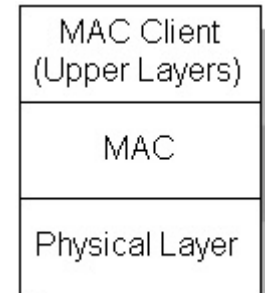
Ethernet Frame: Other Fields

- Preamble: repeating 1010 pattern needed for some PHYs
- Start of Frame Delimiter: mark byte boundary for MAC
- Type / Length: length, or type of frame if >1536 (0x0600)
- Data: protocol data unit from higher layers (ie: IP datagram)
- Pad: Only used when necessary to extend frame to 64 bytes
- Checksum: CRC-32 Detect if frame is received in error
- Idle: Occurs between frames, must be at least 96 bit times.



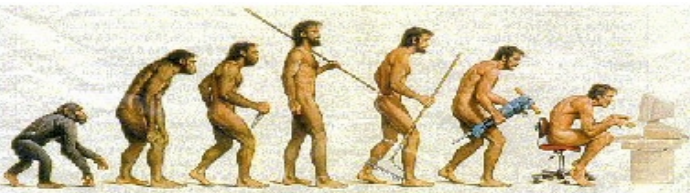
MAC

- Ethernet Frame transmission and reception must be controlled – via the Media Access Control (MAC) layer.
- Ethernet MAC
 - ◆ Operates in either Half or Full Duplex dependent on support from the Physical Layer
 - ◆ Original Ethernet was Half Duplex Only
 - ◆ Handles
 - ◆ Data encapsulation from upper layers
 - ◆ Frame Transmission
 - ◆ Frame Reception
 - ◆ Data decapsulation and pass to upper layers
 - ◆ Does NOT care about the type of Physical layer in use
 - ◆ Does need to know speed of physical layer



MAC: Half Duplex

- Half Duplex: Only one station may transmit at a time
- Requirement on shared mediums (Bus Topologies)
 - 10Base-2, 10Base-5
- Half Duplex Mechanism employed by Ethernet is:
 - **Carrier Sense, Multiple Access with Collision Detect (CSMA/CD)**

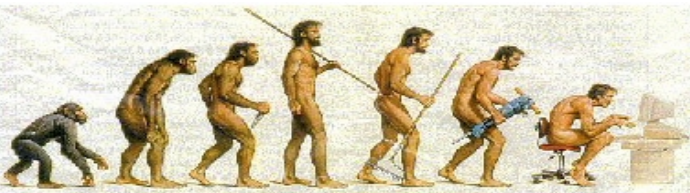


MAC: CSMA/CD

- CS - Carrier Sense (Is someone already talking?)
- MA - Multiple Access (I hear what you hear!)
- CD - Collision Detection (Hey, we're both talking!)

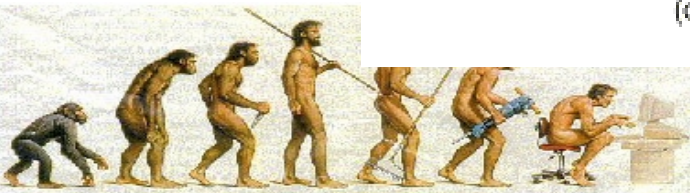
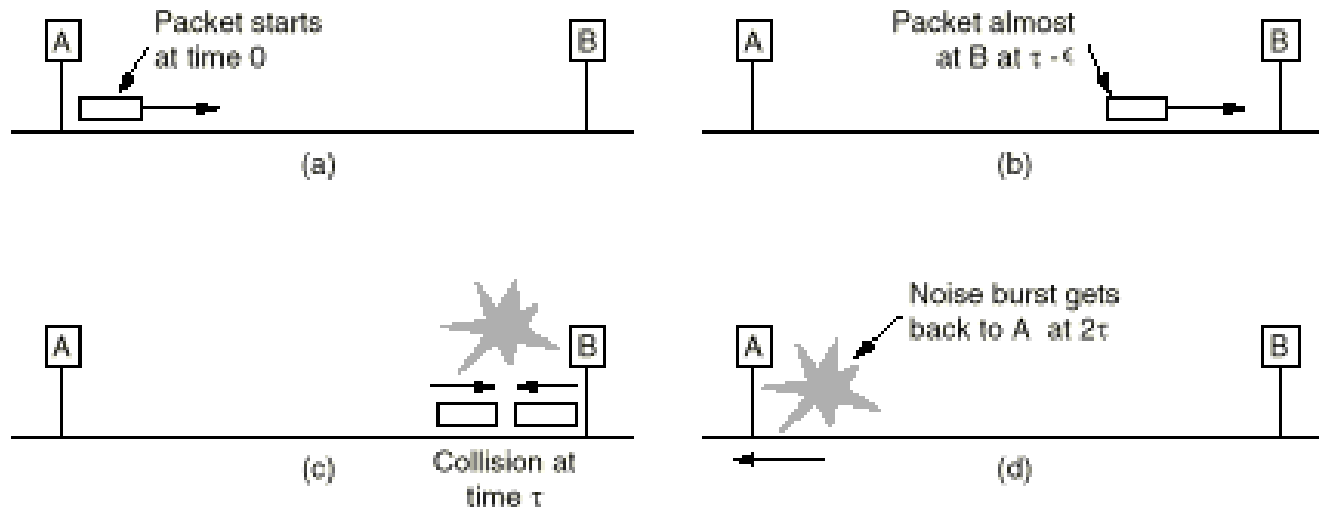
1. If the medium is idle, transmit anytime.
2. If the medium is busy, wait and transmit right after.
3. If a collision occurs, send 4 bytes Jam, backoff for a random period, then go back to 1.

- We use CSMA/CD in normal group conversation.



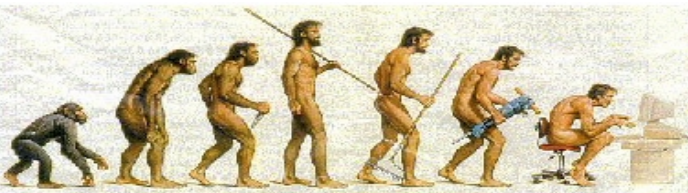
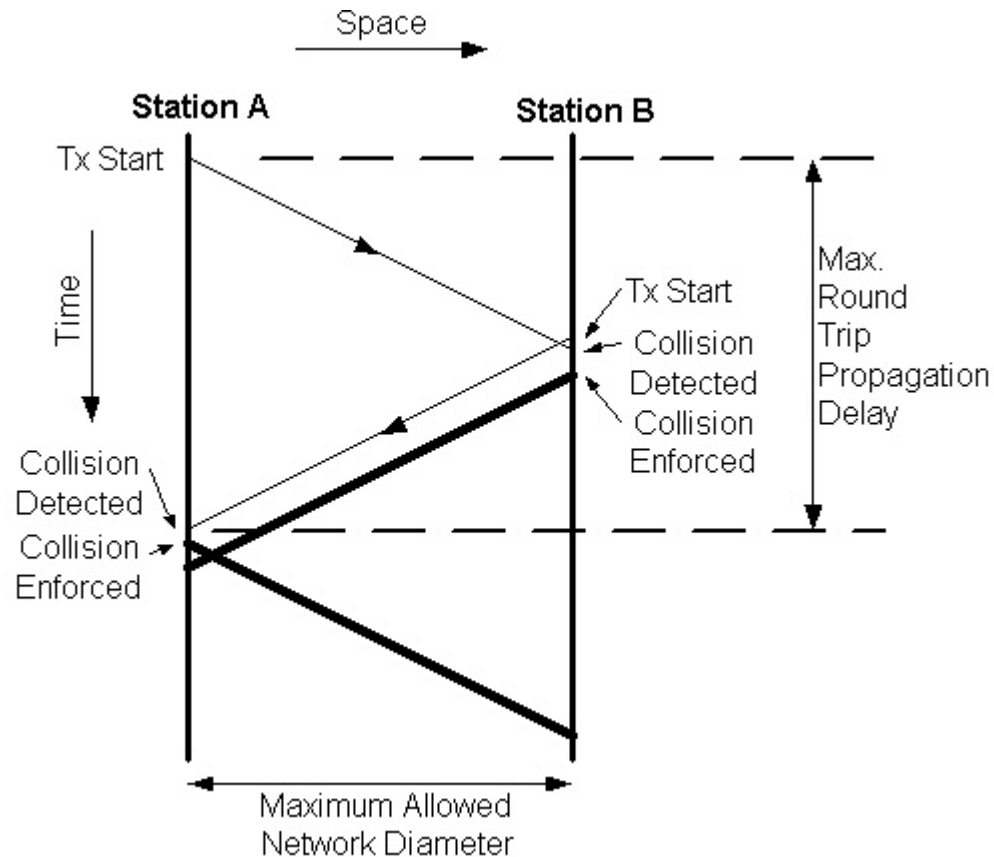
MAC: Collision Domain

- Collision detection can take as long as twice the maximum network end to end propagation delay, worst case.
- This “round-trip” delay defines the max Ethernet network diameter, or collision domain.
- Round-trip delay = 512 bit times for all Ethernets up to this point.



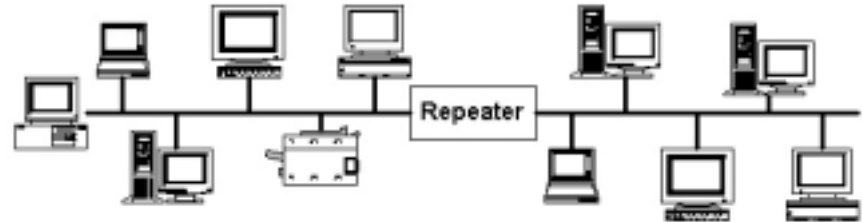
MAC: Collision Domain

- Space-Time depiction of Collision Domain

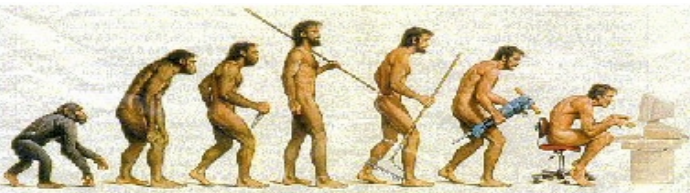
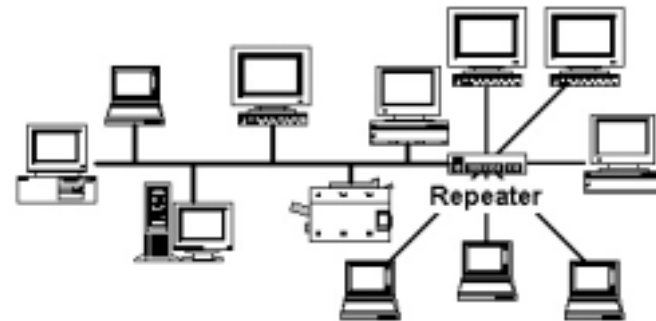
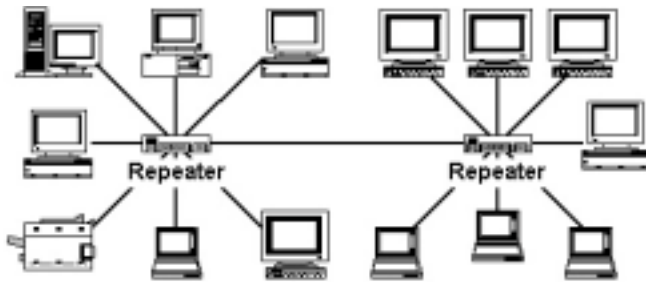


Repeaters

- Works at layer 1 (PHY layer) ONLY
 - **thus it doesn't understand frame formats**
- Repeat incoming signal from a port to all other ports with:
 - **restored timing**
 - **restored waveform shape**
 - **very little delay**

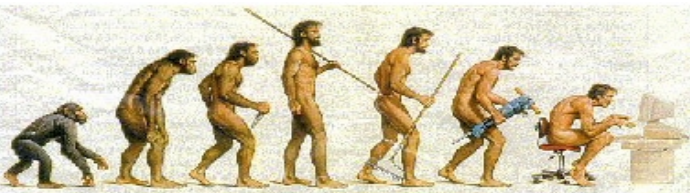
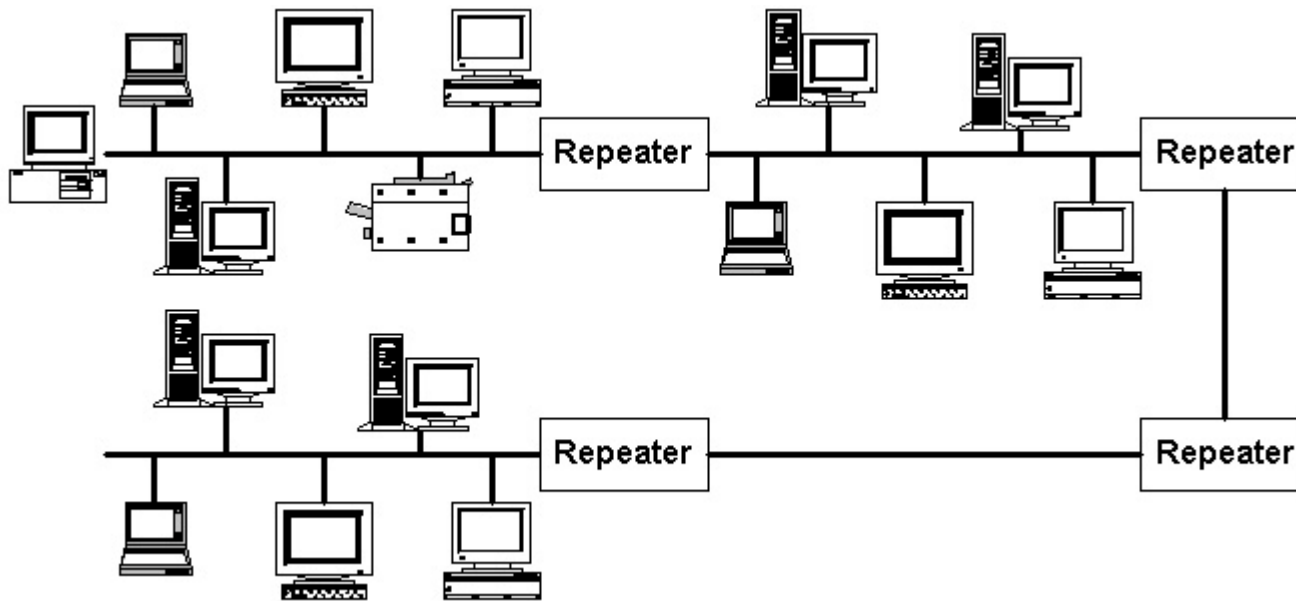


- Half Duplex ONLY: If 2 or more receptions, transmit jam
- Can connect dissimilar media/PHY types (e.g., 10base-T and 10base-2)



Repeaters: 10Mbps

- 5-4-3 Rule
 - 5 Segments
 - 4 Repeaters
 - 3 Populated segments (in the case of 10Base-5 or 10Base-2)



Repeaters: 100Mbps

- 512 bit times isn't much for F.E., because the bit time is 1/10 what it was for 10mbps
 - Even on fiber, the max diameter is 412 meters, and that's purely because of the round-trip time.

Table 29-2—Maximum Model 1 collision domain diameter

Model	Balance d cable (copper)	Fiber
DTE-DTE (see figure 29-3)	100	412
One Class I repeater (see figure 29-4)	200	272
One Class II repeater (see figure 29-4)	200	320
Two Class II repeaters (see figure 29-5)	205	228

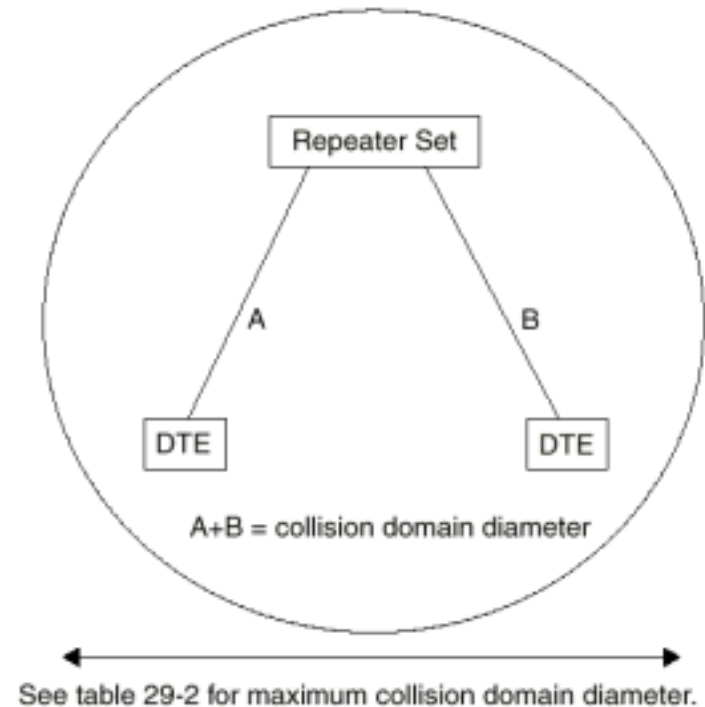
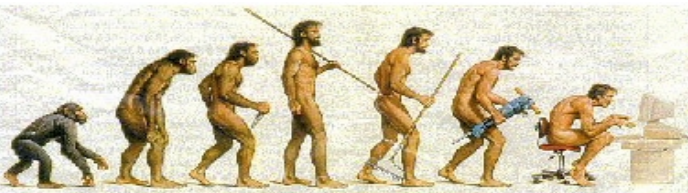
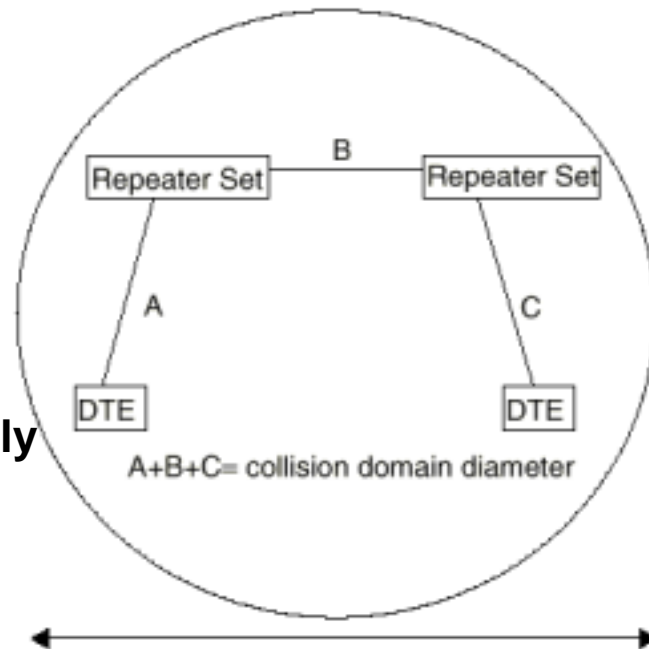


Figure 29-4—Model 1: Single repeater



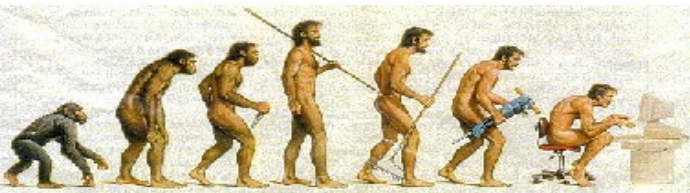
Repeaters: 100Mbps

- Repeater delay is VERY significant. So much so, they defined two types or speeds of repeaters:
 - Type I are slower
 - Type II are faster
 - Even using a Type II, you can only have 2 of them in a network!



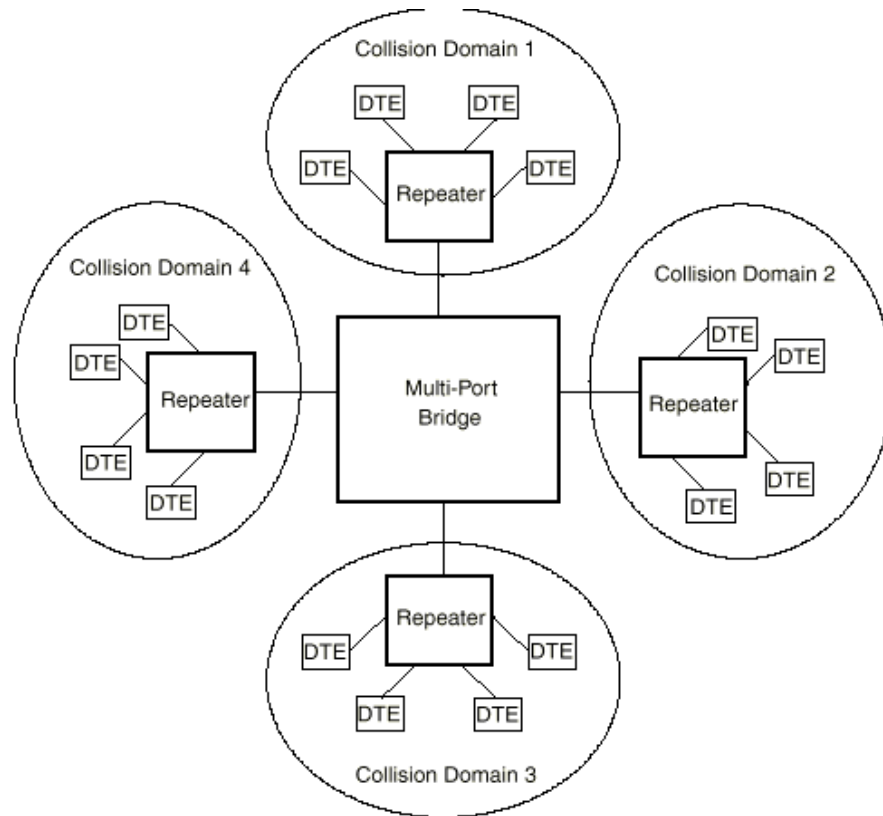
See table 29-2 for maximum collision domain diameter.

Figure 29-5—System Model 1: Two Class II repeaters

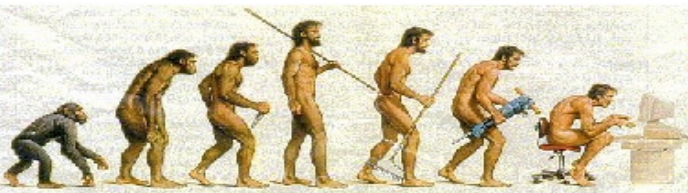


The Ethernet Bridge

- How do you make a half duplex Ethernet network bigger than the Collision Domain allows?? Use a Bridge.



- Repeaters are inside the collision domain, since they propagate collisions
- Bridges/Switches break up the domains, since they operate at layer 2 and buffer packets before sending them



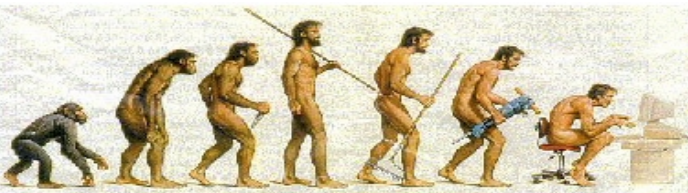
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

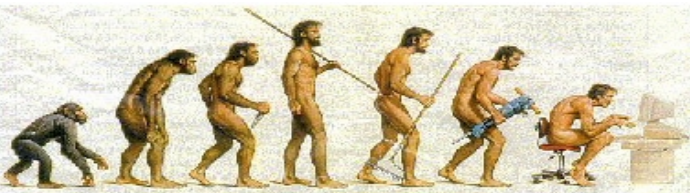
Evolution...

- 10Base-T became dominant in early 90s
 - Half Duplex / CSMA/CD is simple
 - Repeaters are cheap (not complex / low-speed)
 - Growth of Star Topology in building infrastructure
 - Unshielded Twisted Pair (UTP) cheaper than coax
- Predominant use of UTP allows for the creation of Full Duplex MAC.
 - Media is no longer SHARED
 - » 10BASE-T devices transmit on one pair of UTP, and receive on an entirely separate pair.
 - » Unlike coax – simultaneous reception and transmission on the media does NOT corrupt the data transmission.



MAC: Full Duplex

- Remember CSMA/CD? Now forget it.
- As long as we don't need to share a network (like with a repeater or coax), why bother “colliding”?
- New MAC: transmitting while receiving is OK
- Still maintain IPG, frame sizes, and physical layer



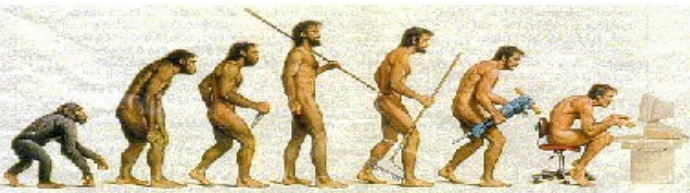
MAC: Full Duplex Pros and Cons

- Pros:

- aggregate throughput = 200mbps
- no collision efficiency penalty
- no need to defer to incoming transmission
- no collision domain - Distance is media dependent and not affected by the protocol.

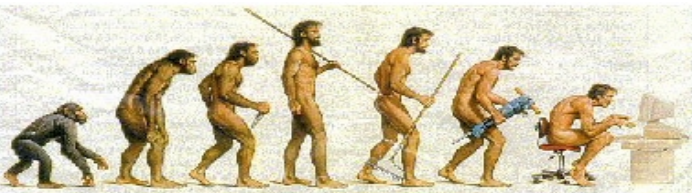
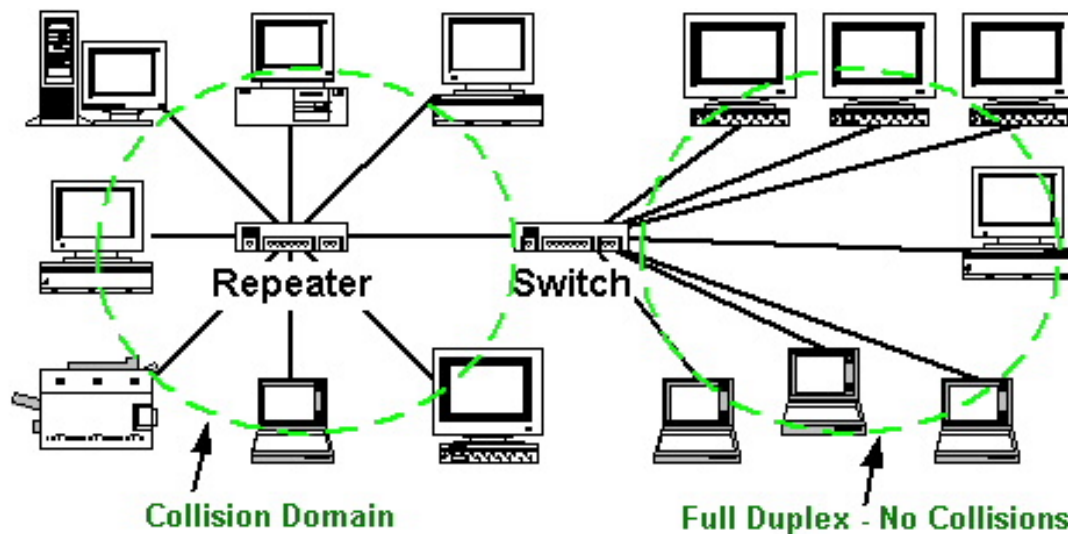
- Cons:

- Must be point-to-point link (i.e., no repeaters)
- Higher throughput means higher speed equipment which means higher cost.
- no built in back pressure mechanism (see MAC Control / Flow Control)
- Need for point to point links requires new hub device to interconnect multiple station – a full duplex capable switch.



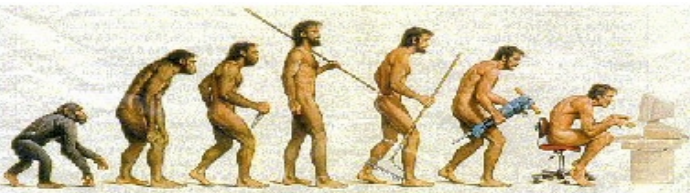
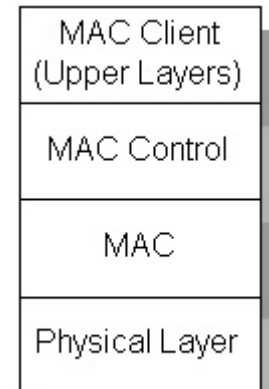
The Ethernet Switch

- Each port of Switch has its own MAC
- Typically can support either Half or Full Duplex
- Can switch between different network speeds (i.e., 10Mbps and 100Mbps)
- More on this to come...



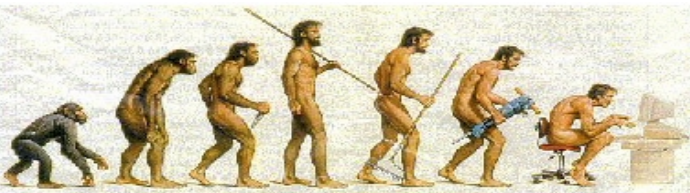
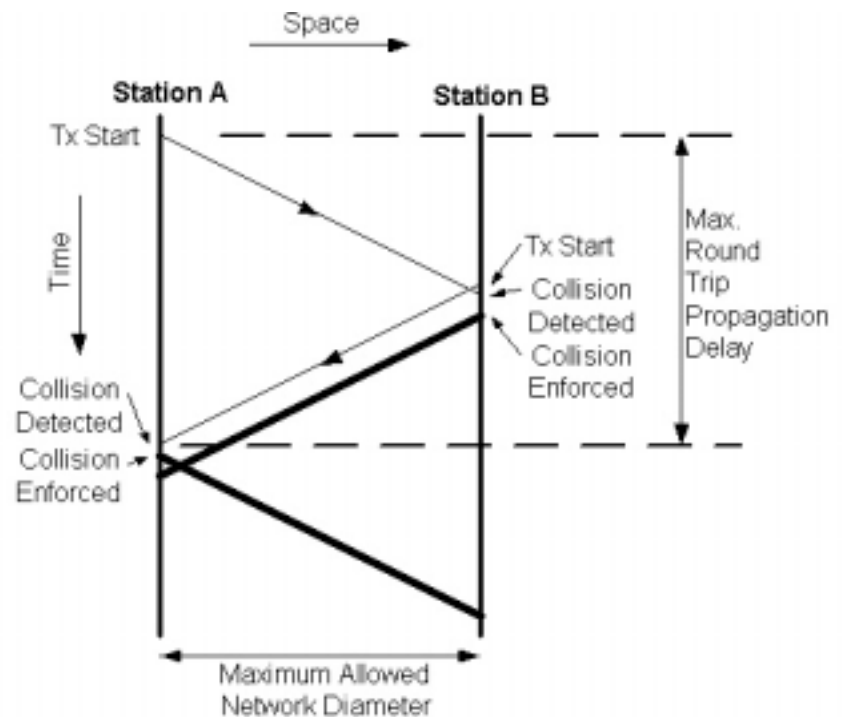
Flow Control

- Supported on Full Duplex links only.
- Sends MAC Control Frames called Pause Frames
- Pause Frames
 - Destination address is a special multicast address that is never forwarded by bridges/switches.
 - » Thus, Link Level Flow Control ONLY
 - Tells MAC Control to pause frame transmission to the MAC for a period of time
- Useful for input constrained devices such as network interface cards (NICs) and buffered distributors (more on those in a bit).



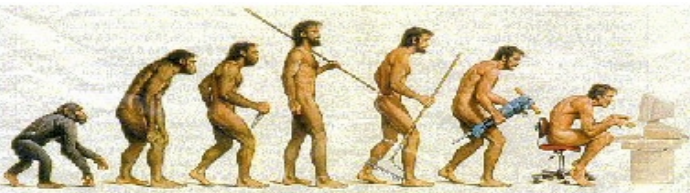
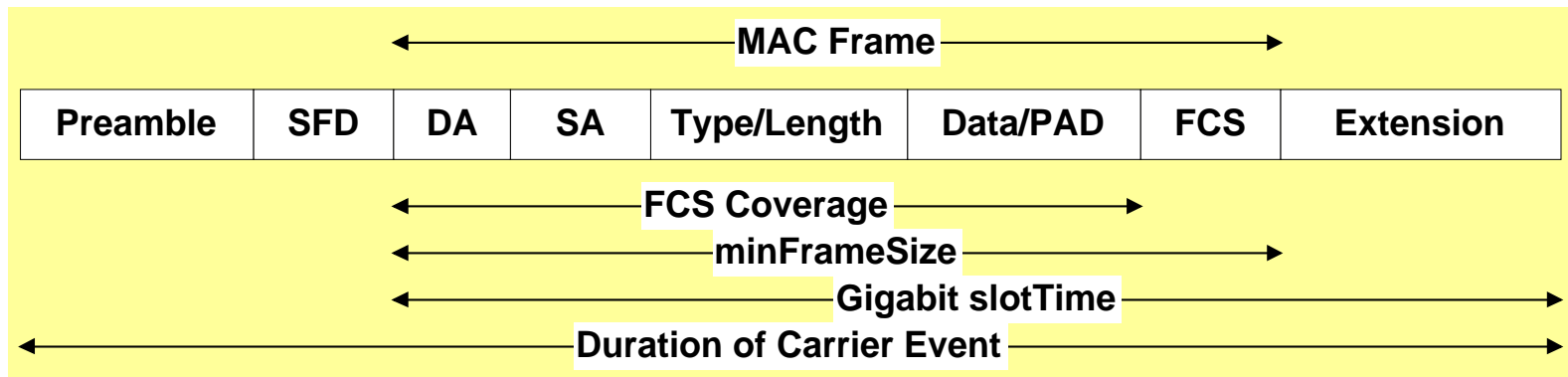
MAC: Half Duplex at 1 Gbps

- Recall the maximum network diameter from 100Mbps is only ~200m.
 - Without modification to the Ethernet MAC, the allowed diameter at 1Gbps (10x faster) would be ~20m (10x smaller)
 - To fix this, the minimum transmission size must be increased such that an acceptable distance (~200m) is covered. Recall the diagram at right:



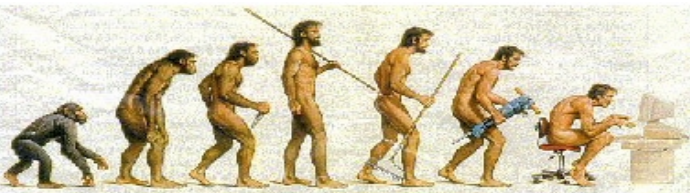
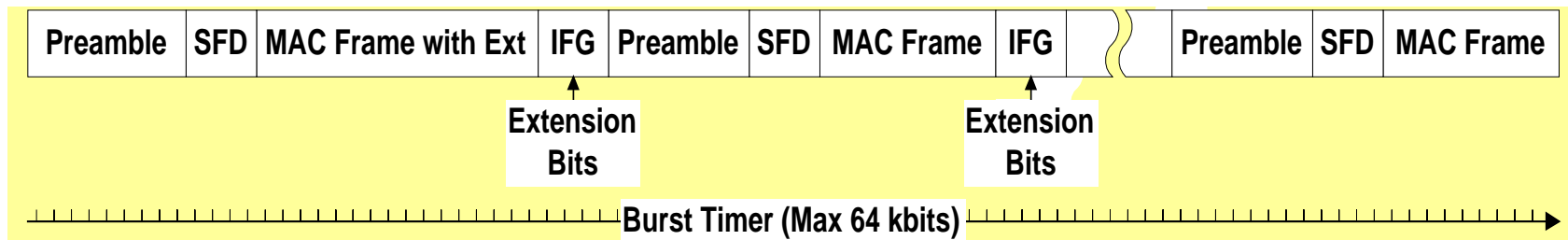
MAC: Half Duplex at 1Gbps

- Solution: MAC adds Extension (non-data) bits after the end of the frame until transmission is at least 4096bits
- Benefits of the extension system
 - Extends collision diameter
 - Maintains compatibility
 - Bigger frame size would add greater multi-rate switch complexity



MAC: Half Duplex at 1Gbps

- Extension fields create waste in small packets
 - This waste cannot be eliminated but can be reduced by frame bursting
- Frame Bursting
 - First frame, if less than 512 Bytes must be extended to 512 Bytes, but may be followed minimum size inter frame gaps (IFGs) filled with extension bits and frames of any size
 - Maximum size transmitted burst ~64Kbits
 - MAC must end burst if no frame is ready to be sent
 - Useful for Servers/Switches transmitting high loads on a half duplex network

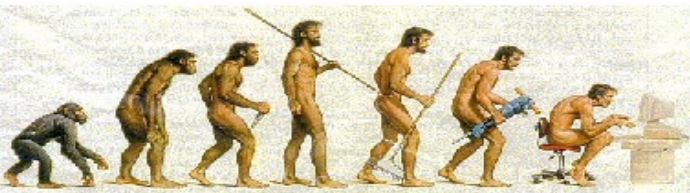


Buffered Distributors

- Very rare devices - That said, they are far more common than half-duplex 1 Gbps devices (which are essentially non-existent) Why?
- Where a repeater is a “bus in a box”, a buffered distributor (BD) is “CSMA/CD in a box”.
 - Each attached device is connected to an input buffer via a full duplex link. If the input buffer is full, the BD uses PAUSE frames to stop traffic from the attached device. When a frame is removed from the input buffer, it is transmitted out all other ports – hence BDs are sometimes called “Full duplex repeaters”
 - Thus the collision domain is the maximum delay within the box, in fact, the medium access method used in the box need not be CSMA/CD, so long as it is fair.
- Benefit: Silicon cheaper than a switch (max speed 1Gbps) and less complex than frame extension and bursting

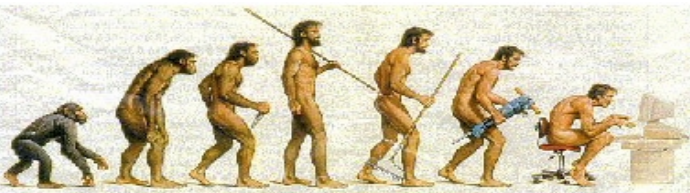


Physical Media Overview



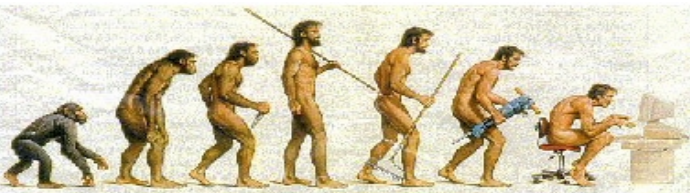
Media Outline

- Copper and Copper and Copper
- Fiber and Fiber



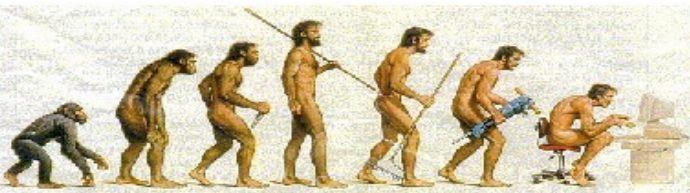
Structured Cabling

- Defines a generic telecommunications cabling system for commercial buildings.
- Specifies the performance of the cable and connecting hardware used in the cabling system.
- Why?
 - The installation of a cabling system is simpler and cheaper during building construction than after the building is occupied.
 - Such a cabling system must have the flexibility to allow the deployment of current and future network technologies.
 - A structured cabling standard provides a design target for the developers of new network technologies (like 1000BASE-T).

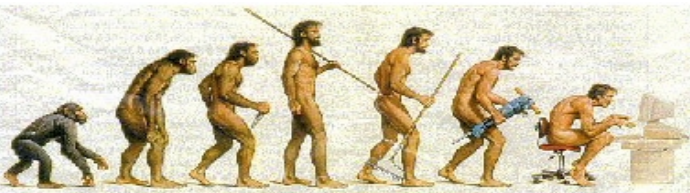
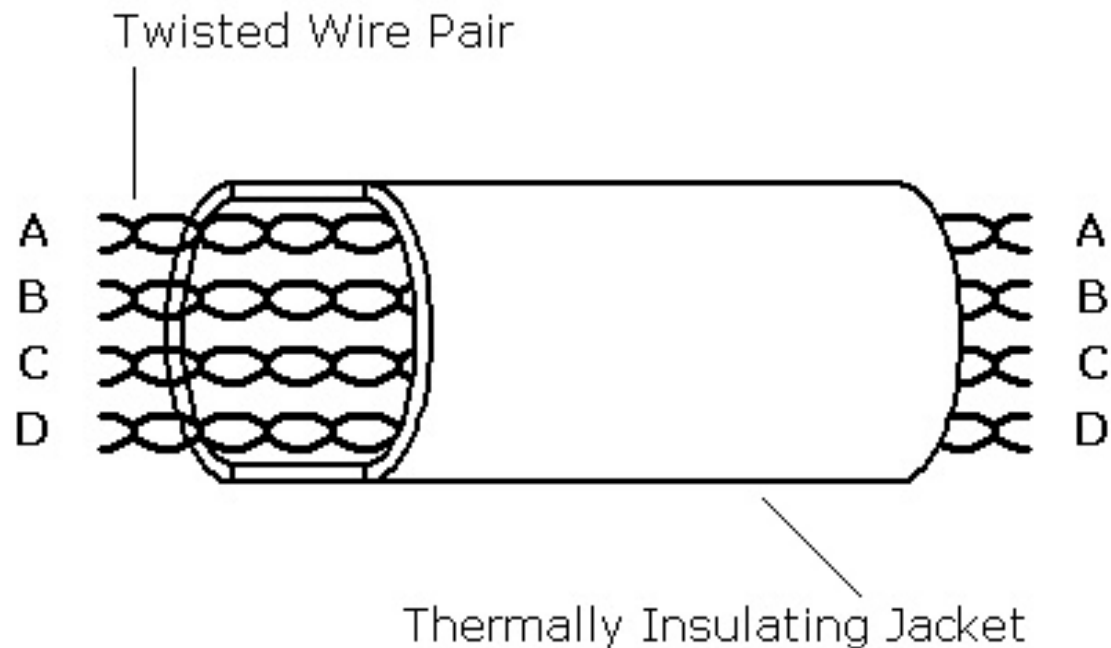


Structured Cabling Standards

- TIA/EIA-568-A, North America
- ISO 11801, International
- Scope:
 - define performance of unshielded twisted pair (UTP), shielded twisted pair (STP), and fiber optic cables and connecting hardware.
 - define how these cables will be used in a generic cabling system.
- Both standards define similar distribution systems and performance requirements.
 - developers do not have to hit two separate targets

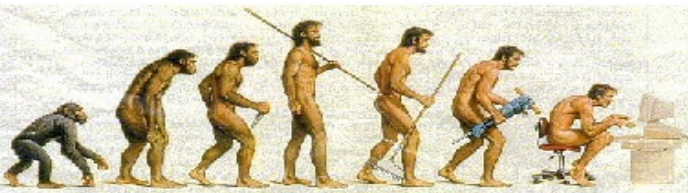


Unshielded twisted pair (UTP) cable



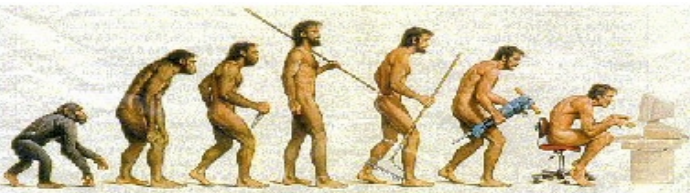
The Category System

- TIA/EIA-568-A defines a performance rating system for UTP cable and connecting hardware:
 - Category 3 performance is defined up to 16MHz.
 - Category 4 performance is defined up to 20MHz.
 - Category 5 performance is defined up to 100MHz.
 - Category 6 performance is defined up to 200MHz



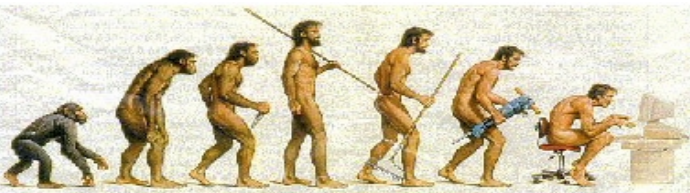
Performance Parameters for UTP Cable

- DC resistance
- characteristic impedance and structural return loss
- attenuation
- near-end crosstalk (NEXT) loss
- propagation delay



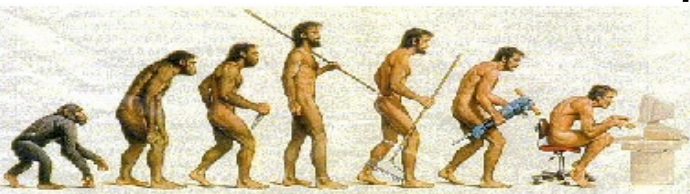
Parameters for UTP Connecting Hardware

- DC resistance
- attenuation
- NEXT loss
- return loss

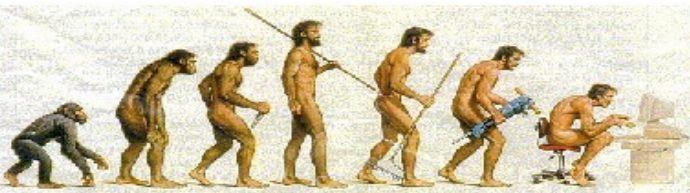
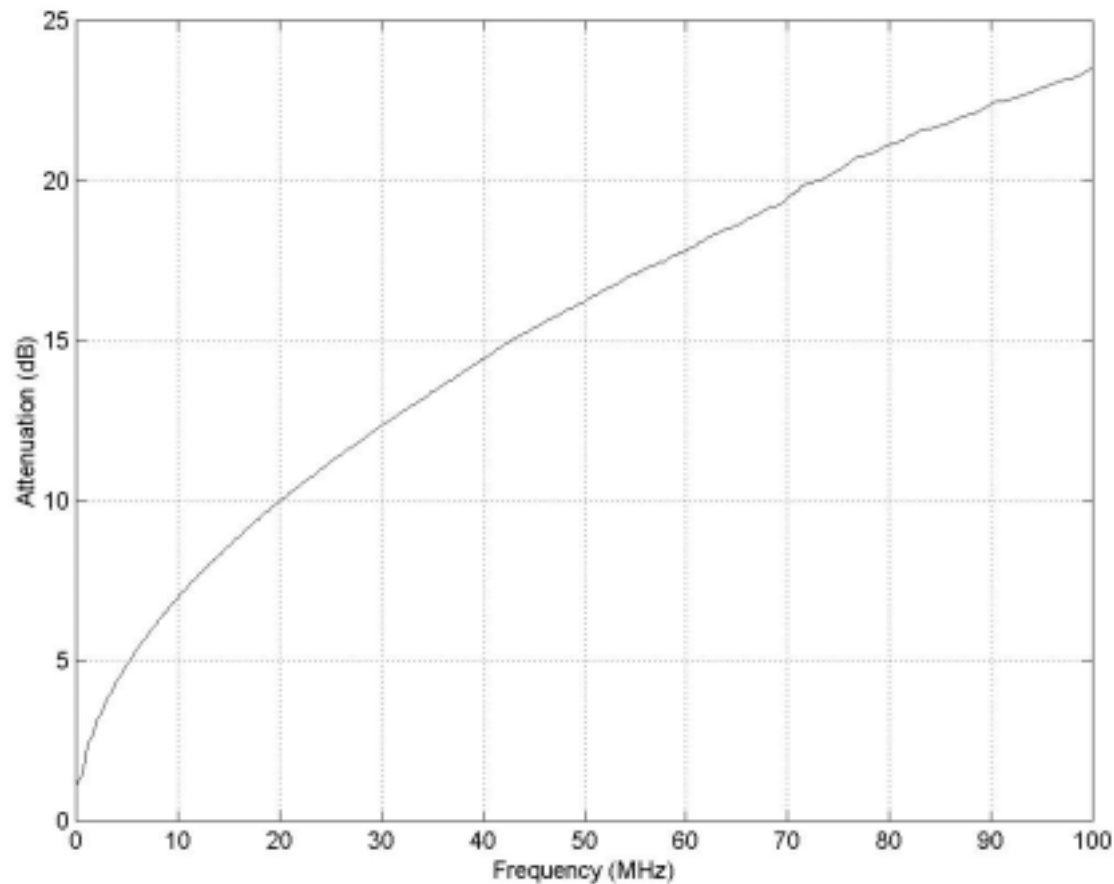


Attenuation

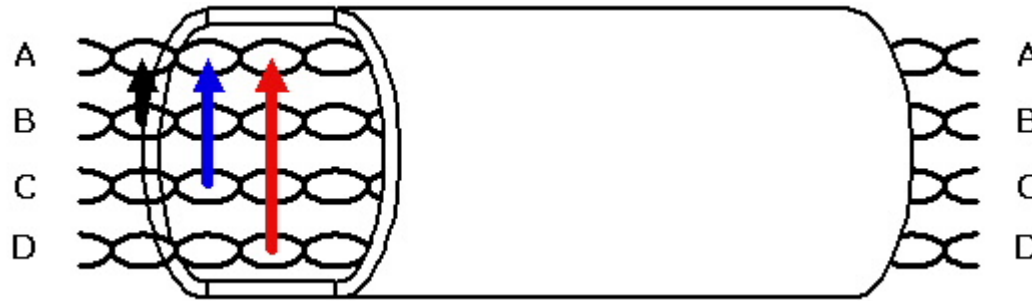
- Electrical signals lose power while traveling along imperfect conductors.
- This loss, or attenuation, is a function of conductor length and frequency.
- The frequency dependence is attributed to the skin effect.
- Skin Effect:
 - AC currents tends to ride along the skin of a conductor.
 - This skin becomes thinner with increasing frequency.
 - A thinner skin results in a higher loss.
- Attenuation increases up to 0.4% per degree Celsius above room temperature (20°C).



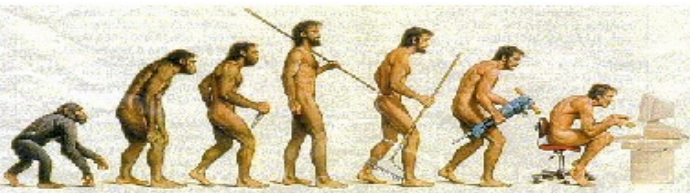
Attenuation vs. frequency



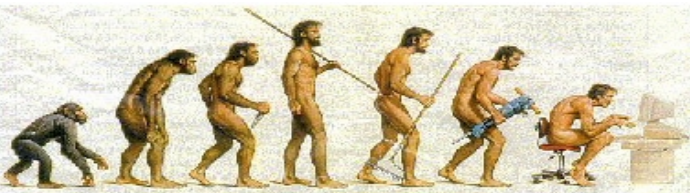
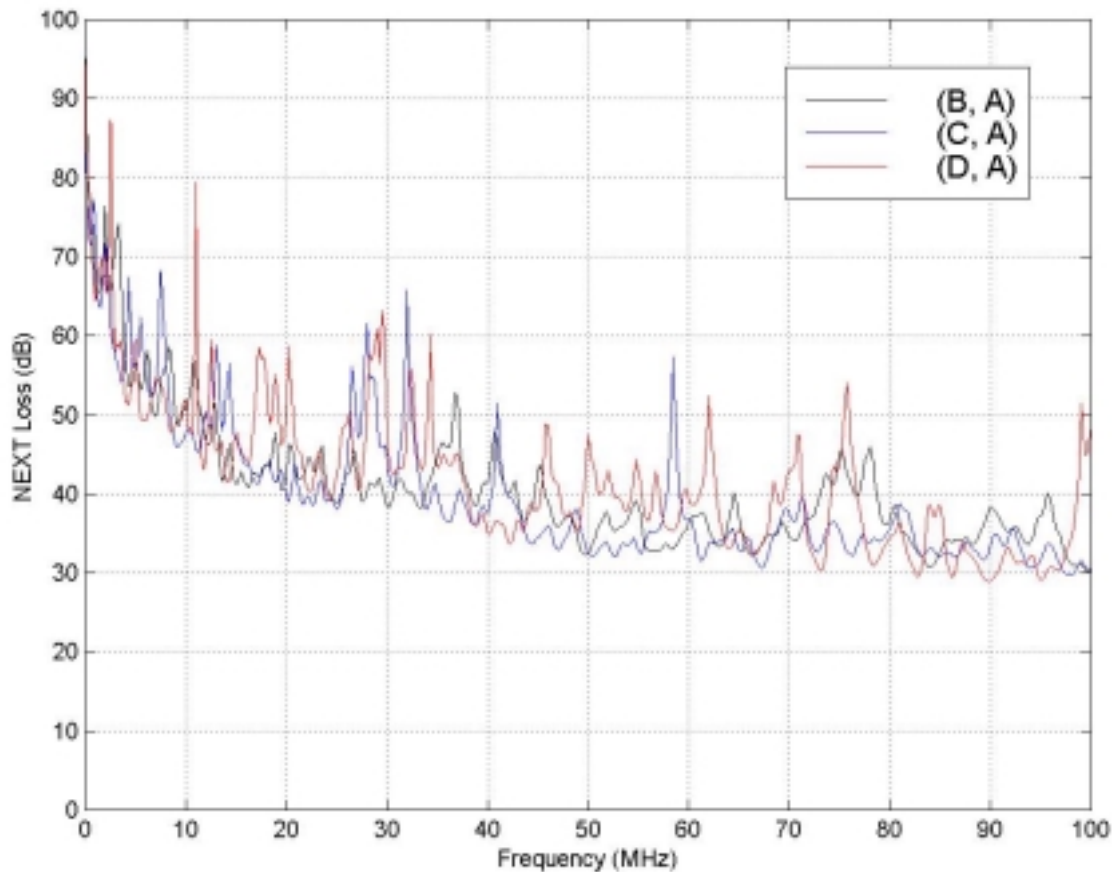
Near-end crosstalk (NEXT) loss



- Crosstalk:
 - **Time-varying currents in one wire tend to induce time-varying currents in nearby wires.**
- When the coupling is between a local transmitter and a local receiver, it is referred to as NEXT.
- NEXT increases the additive noise at the receiver and degrades the signal-to-noise ratio (SNR).

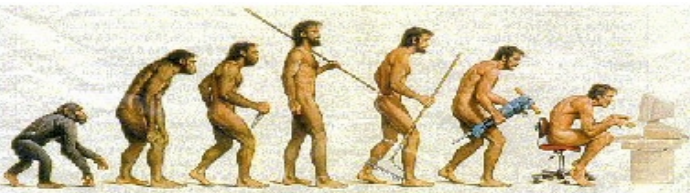


NEXT loss vs. frequency (pair A)



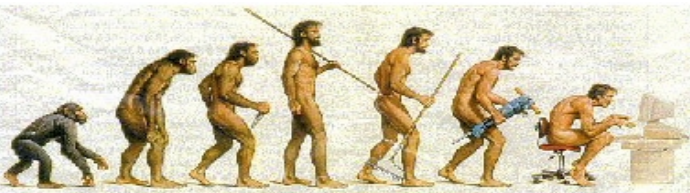
Return loss

- The reflection coefficient is the ratio of the reflected voltage to the incident voltage.
- The return loss is the magnitude of the reflection coefficient expressed in decibels.



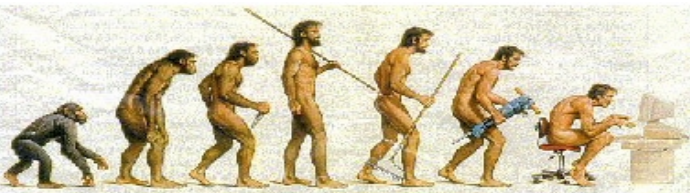
Structured cabling overview I

- Work area
 - for example, an office
- Telecommunications closet
 - focal point of horizontal cabling
 - access to backbone cabling and network equipment
- Equipment Room
 - can perform any of the functions of a telecommunications closet
 - generally understood to contain network resources (for example, a file server)
- Entrance Facility
 - the point at which the network enters the building, usually in the basement



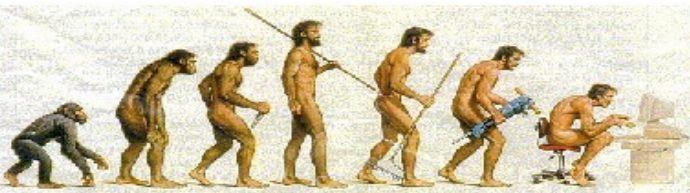
Structured cabling overview II

- **Horizontal Cabling**
 - from the work area to the telecommunications closet.
 - up to 90m of 4-pair unshielded twisted pair (UTP) cable.
- **Backbone Cabling**
 - between telecommunications closets, equipment rooms, and entrance facilities.
 - up to 90m of 4- or 25-pair UTP cable.
- **Flexible Patch Cords**
 - cables use solid conductors making them inflexible and difficult to work with
 - cords use stranded conductors for greater flexibility at the expense of up to 20% more loss than the same length of cable.
 - cords are used at points where the network configuration will change frequently

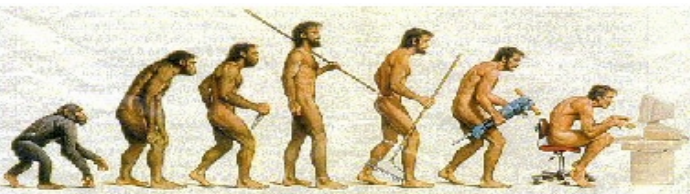
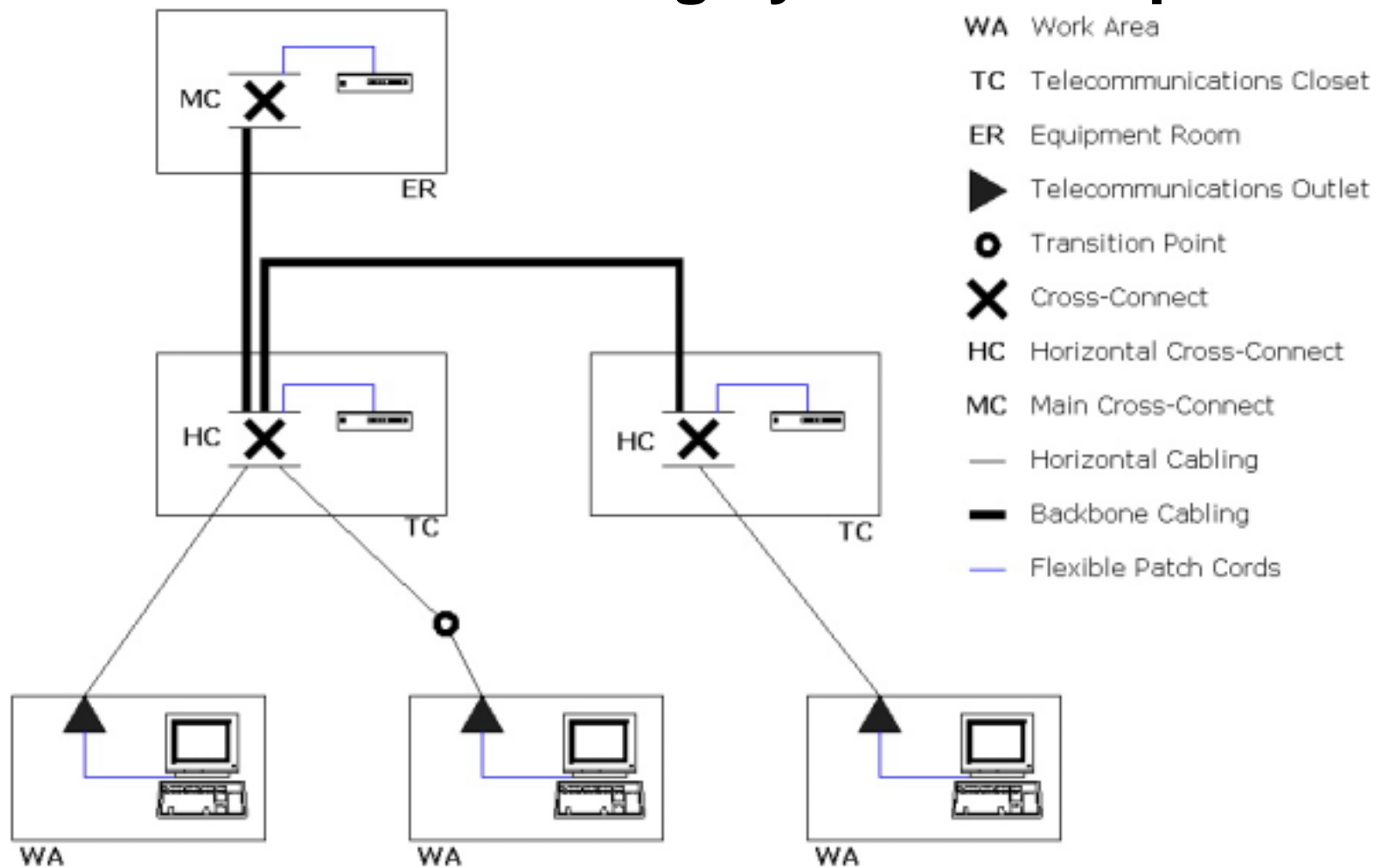


Structured cabling overview III

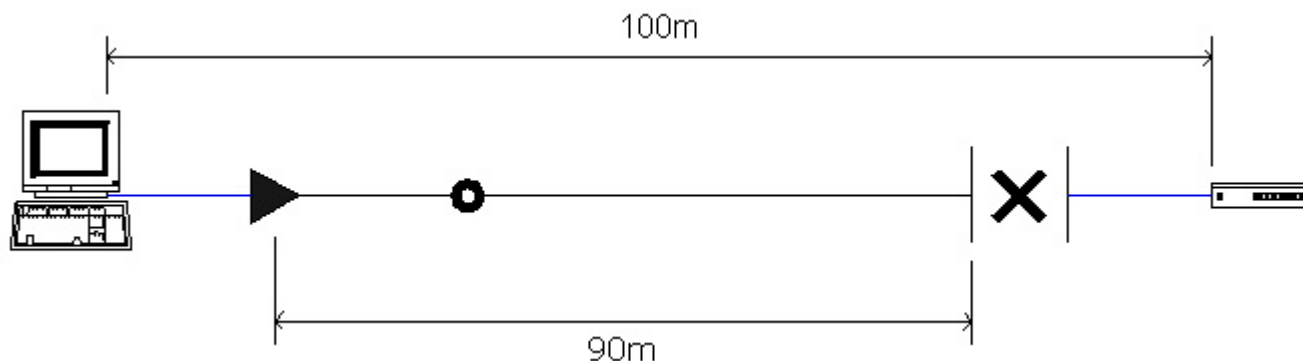
- Transition point
 - connects standard horizontal cable to special flat cable designed to run under carpets.
- Cross-connect
 - a patch between two interconnects
 - horizontal and backbone cabling runs end at interconnects
 - network equipment may use an interconnect
- For UTP cabling systems, horizontal and backbone runs are always terminated in the telecommunications closet and equipment room
 - for example, you cannot cross-connect a horizontal run to a backbone run.



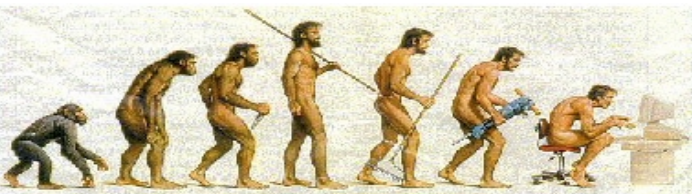
Structured cabling system example



TIA/EIA-568-A channel definition

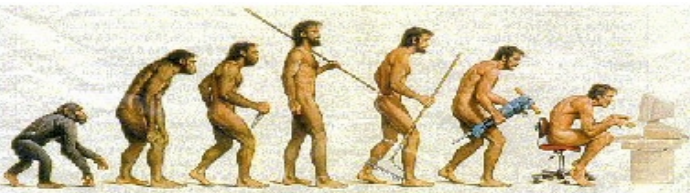


- 90m of horizontal cable
- 10m of flexible cords
- 4 connectors
- ISO 11801 channel definition does not include a transition point (3 connectors).
- The channel definition is the developer's design target.



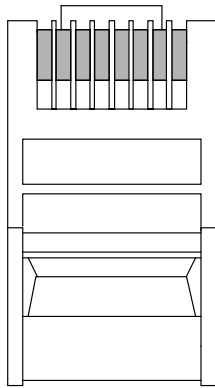
UTP Copper Info

- Two types of UTP – Solid Core (for Horizontal runs) and Stranded Core (for patching)
- RJ-45 connectors – when making cable:
 - **solid:** use 8-pin modular (RJ-45) plug with three prongs on the metal contacts
 - **stranded:** two prongs on metal contacts
- **Wiremap** – there are four pairs in a normal UTP cable: orange, green, blue, and brown. In each pair there is a white wire with a colored stripe and a colored wire with a white stripe. By convention, the white wire carries the signal and the colored wire carries the inverted signal.



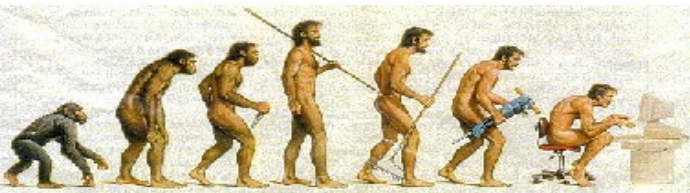
More UTP Copper Info

- looking down at the top of the RJ-45 (locking tab facing down)



1	2	3	4	5	6	7	8
Solid White, Orange Stripe	Solid Orange, White Stripe	Solid White, Green Stripe	Solid Blue, White Stripe	Solid White, Blue Stripe	Solid Green, White Stripe	Solid White, Brown Stripe	Solid Brown, White Stripe

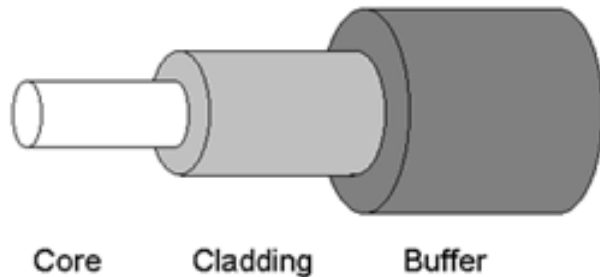
- A good UTP cable will have:
 - All 8 wires visible at the very end of the RJ45 cable
 - The Jacket of the cable fully inserted into the RJ45 (so when you pull on the cable, you pull the jacket and the RJ45, and not the wires and contacts)



Fiber Types

- Fiber's form and 3 basic types

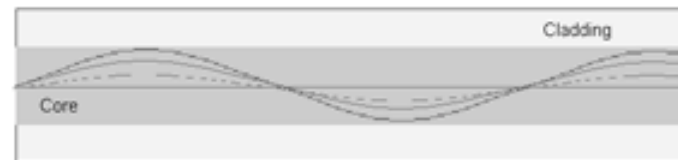
Fiber Structure



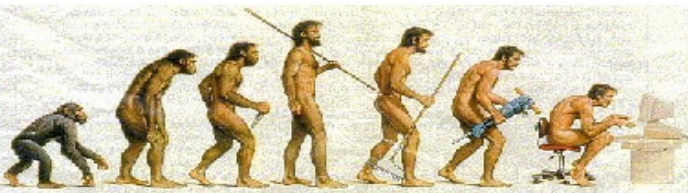
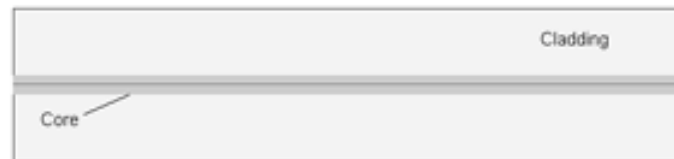
Multimode Step Index



Multimode Graded Index

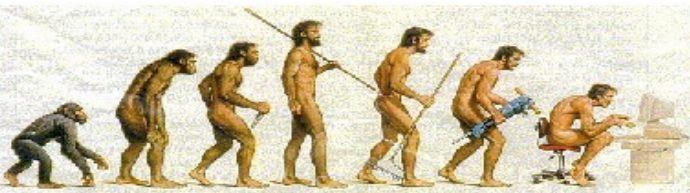


Singlemode



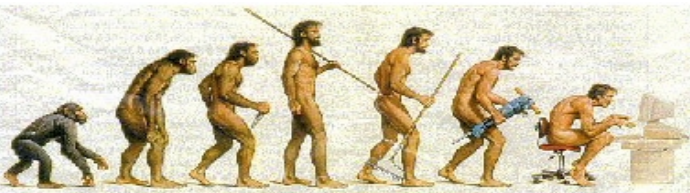
Fiber Core Sizes

- Multimode: Two most common are 62.5 and 50 micron
- Singlemode – To be singlemode, diameter must be no more than ~6 times the wavelength, thus for 1330 – 1510 nm light, the diameter must be 7 to 9 micron
- Mismatching core sizes can be done, but in general is bad!



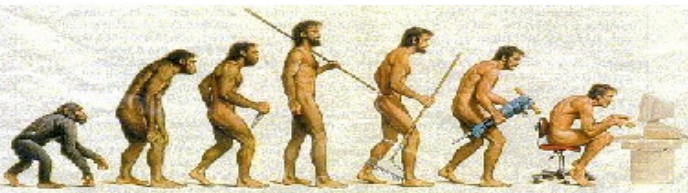
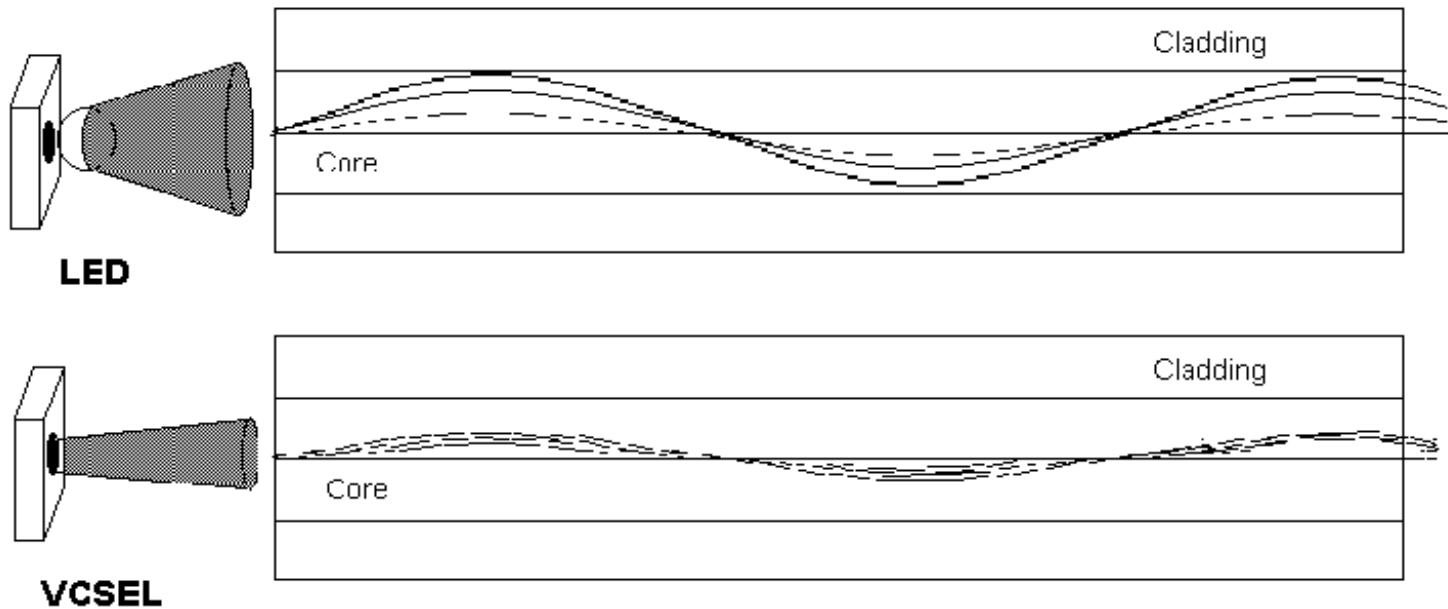
Modal Dispersion

- A flashy name to describe how in a multi-mode fiber, the multiple paths arrive at the end of the fiber at different times.
- This multi-path delay causes a “small” input pulse to smear into a “wide” output pulse, degrading the rate at which those data pulses can be sent down the fiber (degrading the bandwidth)
- Note that a multimode step index fiber has high modal dispersion, as a result, such fiber is not typically used today.
- Graded index fiber varies the refraction index of the fiber material from the center of the fiber out to its edge. The result is that the modes traveling the longer sinusoidal paths actually propagate faster than the modes traveling the shorter, straighter paths – thus, in a perfect graded index fiber, all modes would arrive at the same time at the output of the fiber, and no smearing would occur. (of course perfection is impossible to achieve)



Fiber Launches

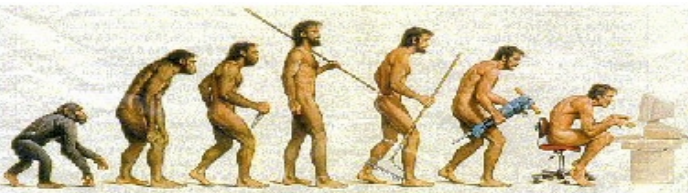
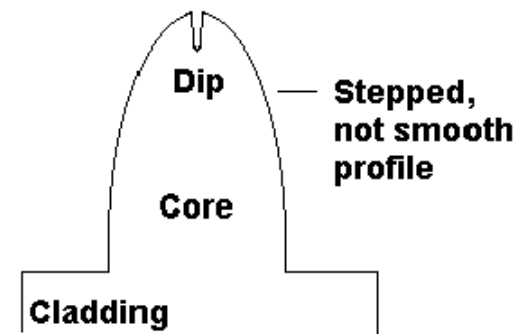
Mode Fill Variations by LED and VCSEL Sources



Recent Modal Dispersion Issues

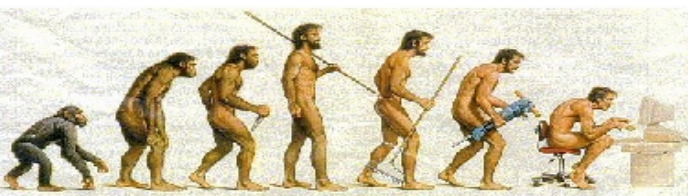
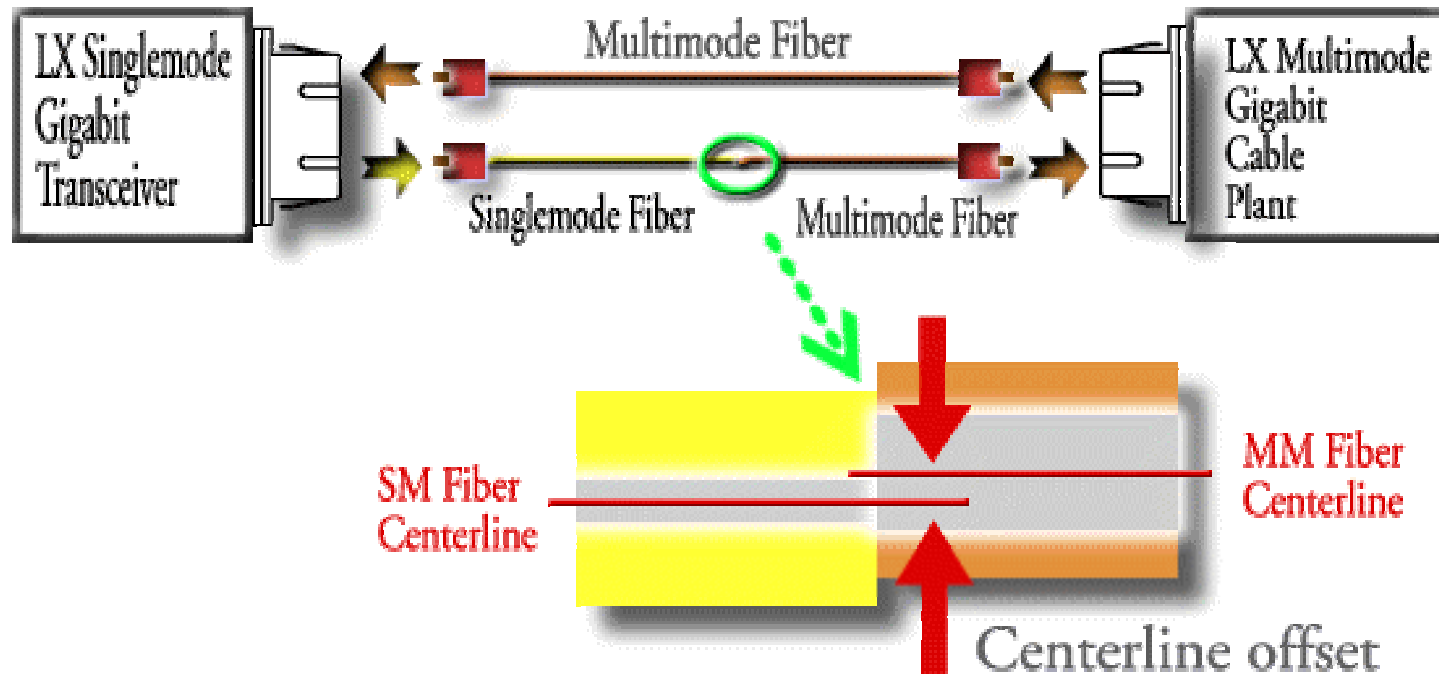
- As a side-note: During the development of Gigabit Ethernet, it was discovered that a sizeable percentage of installed (old) multimode graded index fiber had a deviation in the center of the core, as depicted at left. This was determined to cause increased modal dispersion (DMD, differential modal delay as the gig folk called it)
- The result of this discovery of a flaw in the installed media was the subsequent reduction in supported link lengths for GbE on installed multimode fiber.
- Which leads to the following new term...

Graded-Index Multimode Fiber Index Profile



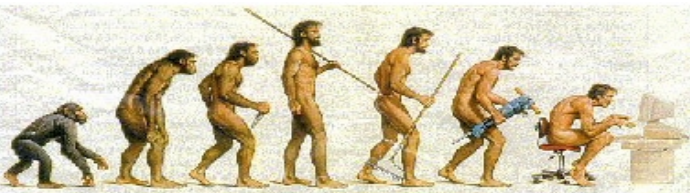
Launch Conditioning

- Many options exist to minimize the DMD effect, one of which is to use a single-mode “pig-tail” to condition/reduce the modes launched into the multi-mode segment, by launching them off-center (and thus avoiding the index “notch” in the center of the installed fibers)



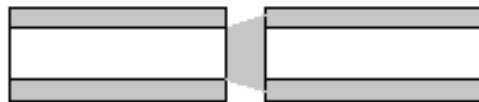
Chromatic Dispersion

- Effects all fiber types, but is one of the primary limiting factors of long-haul single-mode connections.
- As mentioned earlier, EM waves propagate at different speeds in different media. Well, simply put, EM waves at different frequencies propagate at different speeds, even in the same media! In general, this is referred to as group delay, but in Fiber Optics, its more commonly called chromatic dispersion.
- This spreading of different colors smears the transmitted pulses just as with modal dispersion

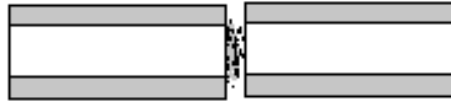


Bad Fiber Connections

Connector Loss Factors



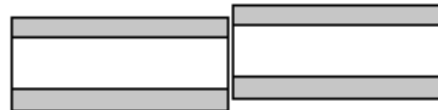
End Gap



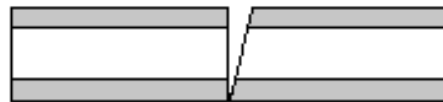
Dirt & Finish



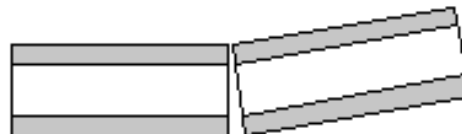
Concentricity



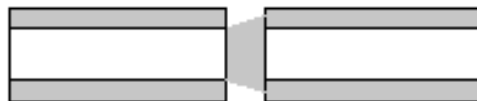
Coaxiality



End Angle



Axial Runout

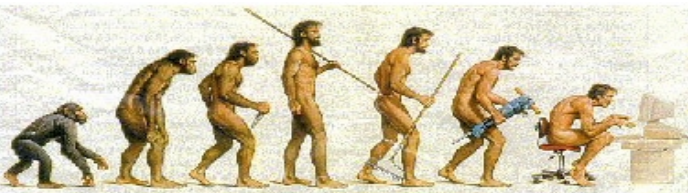


NA Mismatch

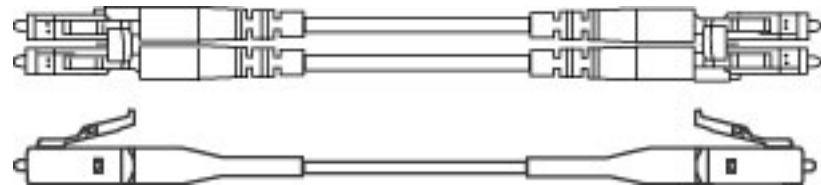
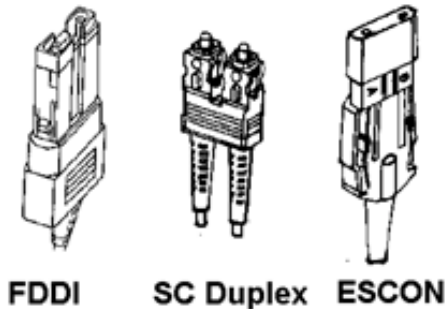
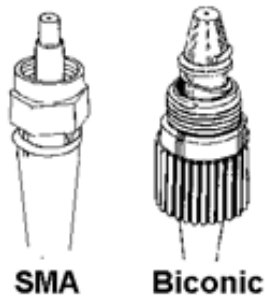
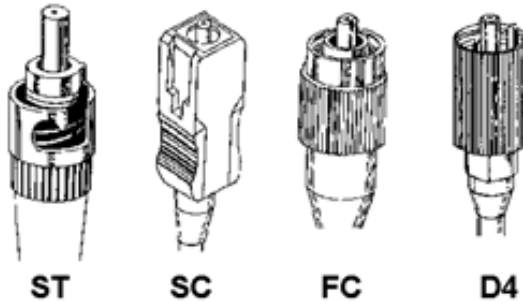


Core Mismatch

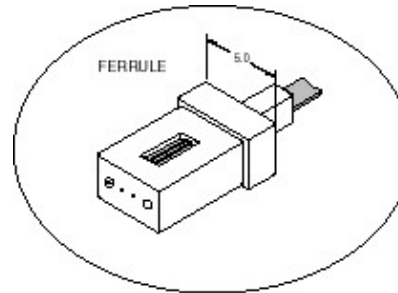
- Always keep your Fiber **CLEAN!** And Capped when not in use (both patch and ports)



Some Types of Connectors

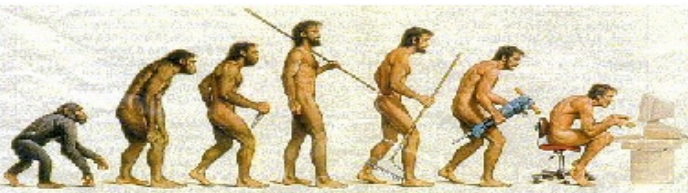


• LC



• MTRJ

• Especially note
SC, ST, and MTRJ

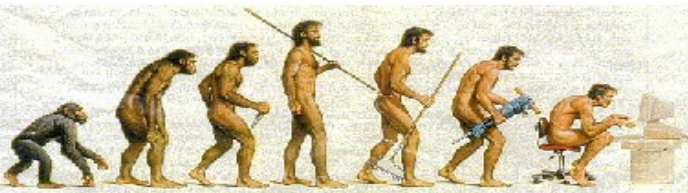
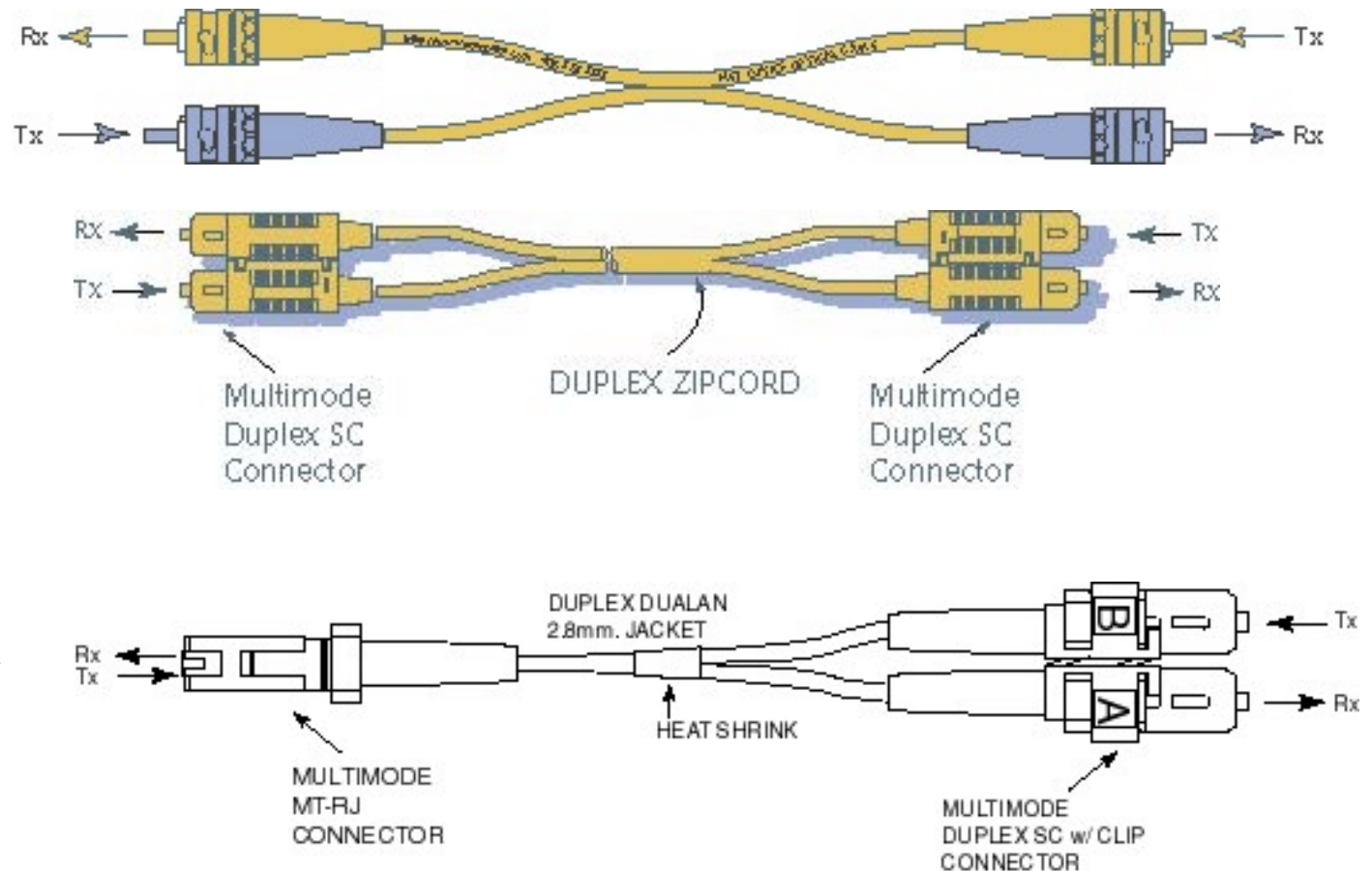


Most Common Connectors

- ST

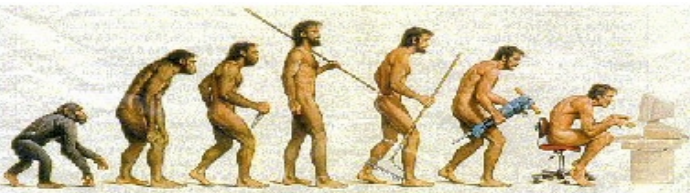
- SC

- MTRJ
to SC



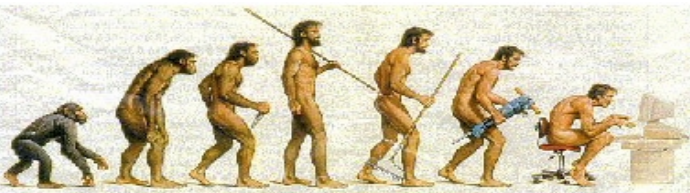
Fiber Info

- Not only must the fiber be kept clean, but its end should be polished (to eliminate scratches) and rounded.
- The rounding of the ends allows the two fiber ends to touch without an airgap forming between them.
- DO Keep your fiber clean. If in doubt, use an airgun or alcohol swab to clean the ends.
- DO Keep your fiber capped when not in use (to prevent dust and scratches)
- DO Keep fiber ports on devices capped when not in use – dust collected on a transmitter or receiver can only – at best, be blown at with an airgun.

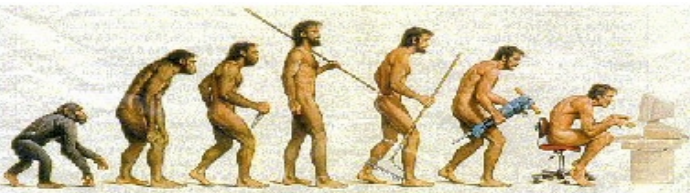


More Fiber Info

- DON'T EVER allow the fiber to bend more than the diameter of your closed fist. Fiber is glass, bent glass breaks and/or creates microfractures.
- DON'T touch the tip of the fiber with your finger/body – The core may be protruding and slice you open (it is glass!)
- DON'T look down a fiber! Its YOUR EYESIGHT at stake!! Use a power meter to check your cable!
- Quick way to test fiber / find the right fiber –Hold one fiber at the far end up to a light, look in the other end to find the white dot! – If you don't see it, it's the wrong fiber, or its EXTREMELY broken.

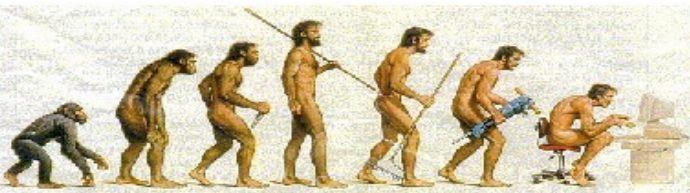


Core PHYs



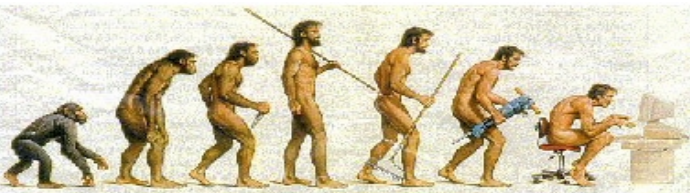
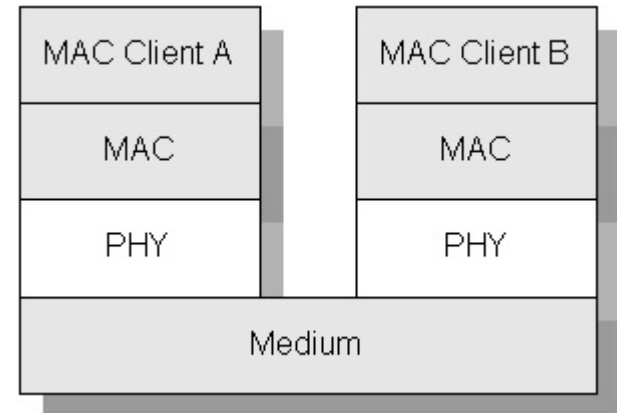
Core PHYs Outline

- Review of the Physical layers role
- UTP Copper Phys
- Fiber Phys
- Primary focus on 10Base-T, 100Base-TX, and 1000Base-T



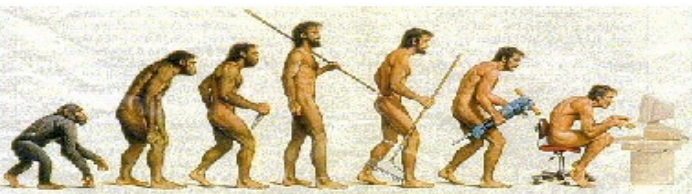
Review

- **MAC**
 - Sends/Receives data to/from higher layer client
 - Handles addressing of data frames for the LAN
 - Appends a checksum to ensure frame validity on reception
- **Medium**
 - Available channel in infrastructure for data transmission
 - May be copper or fiber
 - Medium selected is typically a matter of cost and installed base
- **PHY – Physical Layer**
 - Prepares MAC frame for Medium
 - Drive signal across Medium (Tx to Rx)



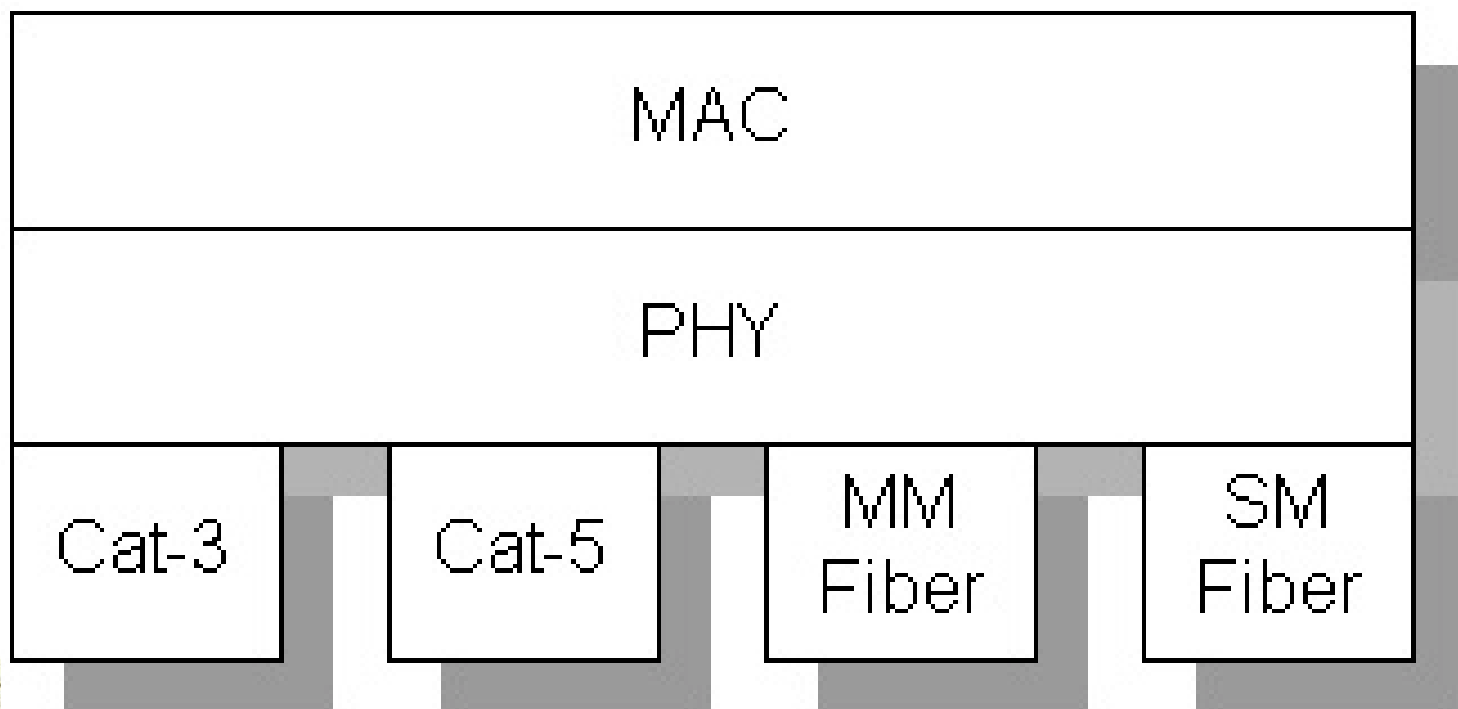
PHY Objectives

- Balance of the following desirables:
 - **Low Cost**
 - **Long Distance**
 - **High Data Rate capability**
 - **Low Bit Error Rate**
 - **Support the current installed media when possible**



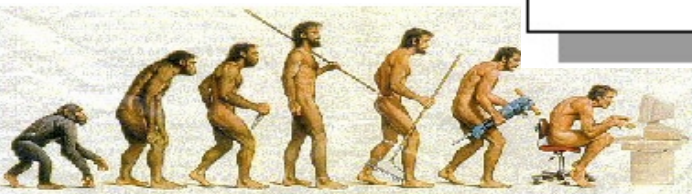
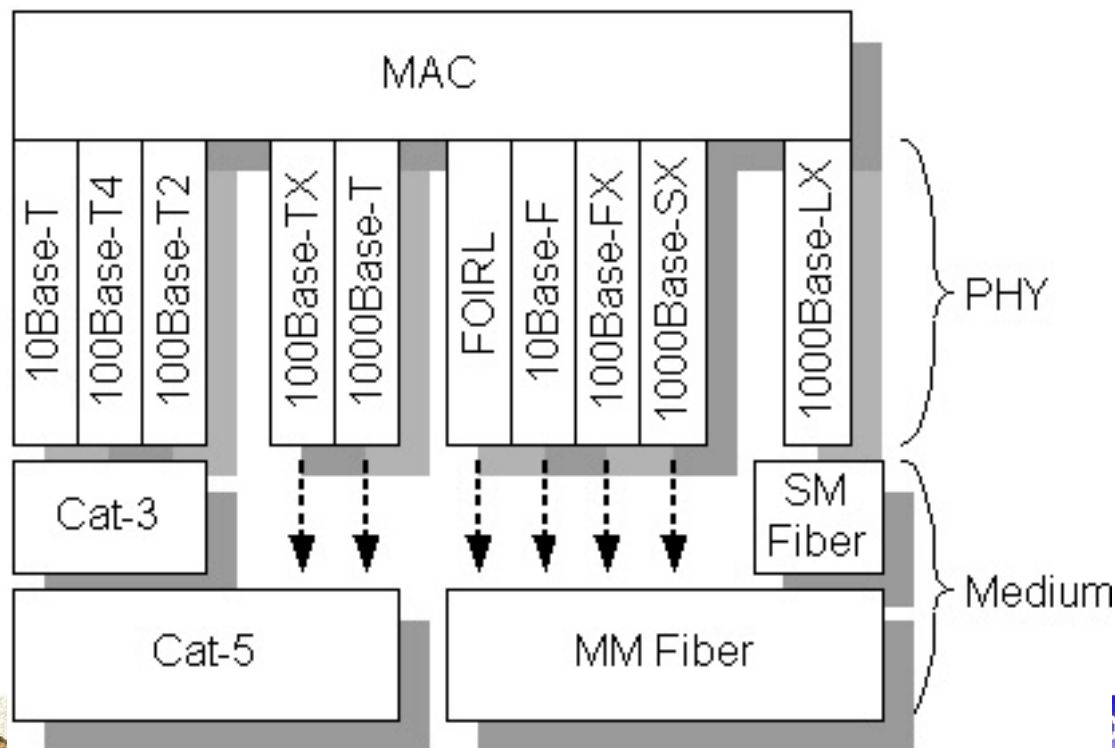
PHY Stack Model - Slide 1

- Notice the PHY connects the SAME MAC layer to different types of Media, the four most common are shown:



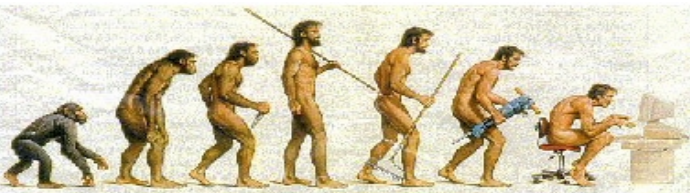
PHY Stack Model – Slide 2

- To support different media, different PHYs are required.
- Each PHY balances the objectives (cost, speed, etc) differently



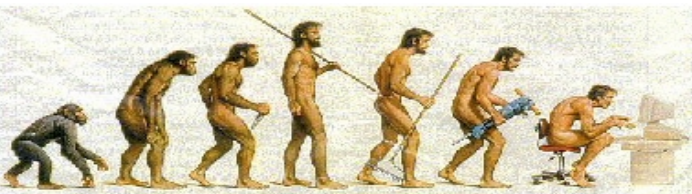
PHY Evolution

- 1985 – IEEE 802.3 – 10Base-5 & 10Base-2
- 1987 – IEEE 802.3d – FOIRL
- 1990 – IEEE 802.3i – 10Base-T
- 1993 – IEEE 802.3j – 10Base-F
- 1995 – IEEE 802.3u – 100Base-T4 / TX / FX
- 1997 – IEEE 802.3y – 100Base-T2
- 1998 – IEEE 802.3z – 1000Base-SX / LX / CX
- 1999 – IEEE 802.3ab – 1000Base-T



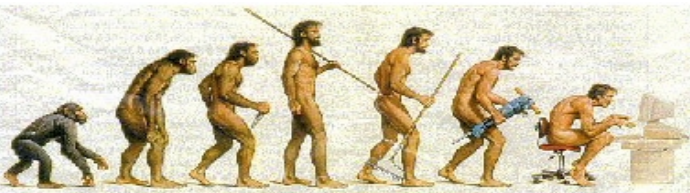
UTP Copper Evolution Summary

- 1989 – 10Base-T:
 - Cat-3 Cabling dominant (pre Cat-5)
- 1995 – 100Base-T4:
 - Support Cat-3, using all 4 pairs, only half duplex capable
- 1995 – 100Base-TX:
 - Capitalizing on CDDI (FDDI) work
 - Requires Cat-5 (low installed base at time), full duplex capable
- 1997 – 100Base-T2:
 - Cat-3 or better, requires only 2 pair, full duplex capable
- 1999 – 1000Base-T:
 - Cat-5, requires all 4 pair, full duplex capable



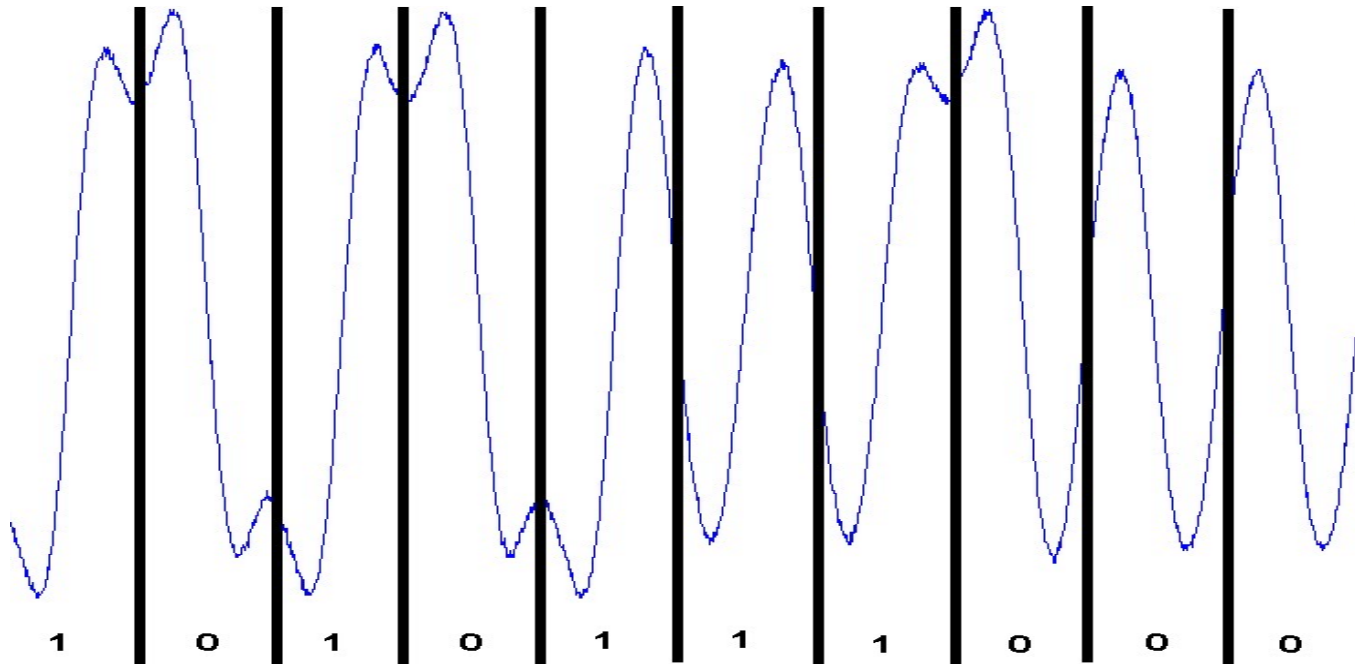
10BASE-T Overview

- 10 Mbps (Million Bits per second)
- Category 3 cable (voice-grade) or better
- 2 pairs DATA (1,2 & 3,6) leaves center pair (4,5) for PHONE
- 100 meter runs (limited by cable attenuation)
 - All that's necessary to support building structured cabling
- Star/Hub Topology
- Inherently Full Duplex (Upper layer MAC/Repeater may be half duplex though)

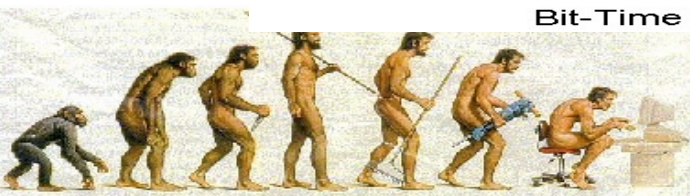


10BASE-T: Data Encoding

- 10base-T uses Manchester Encoding
- + to - transition = "0" - to + = "1"
- Always DC balanced & Always has a transition each bit-time for clock recovery.



Bit-Time Boundaries occur every 100ns



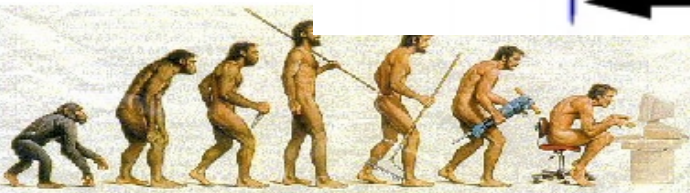
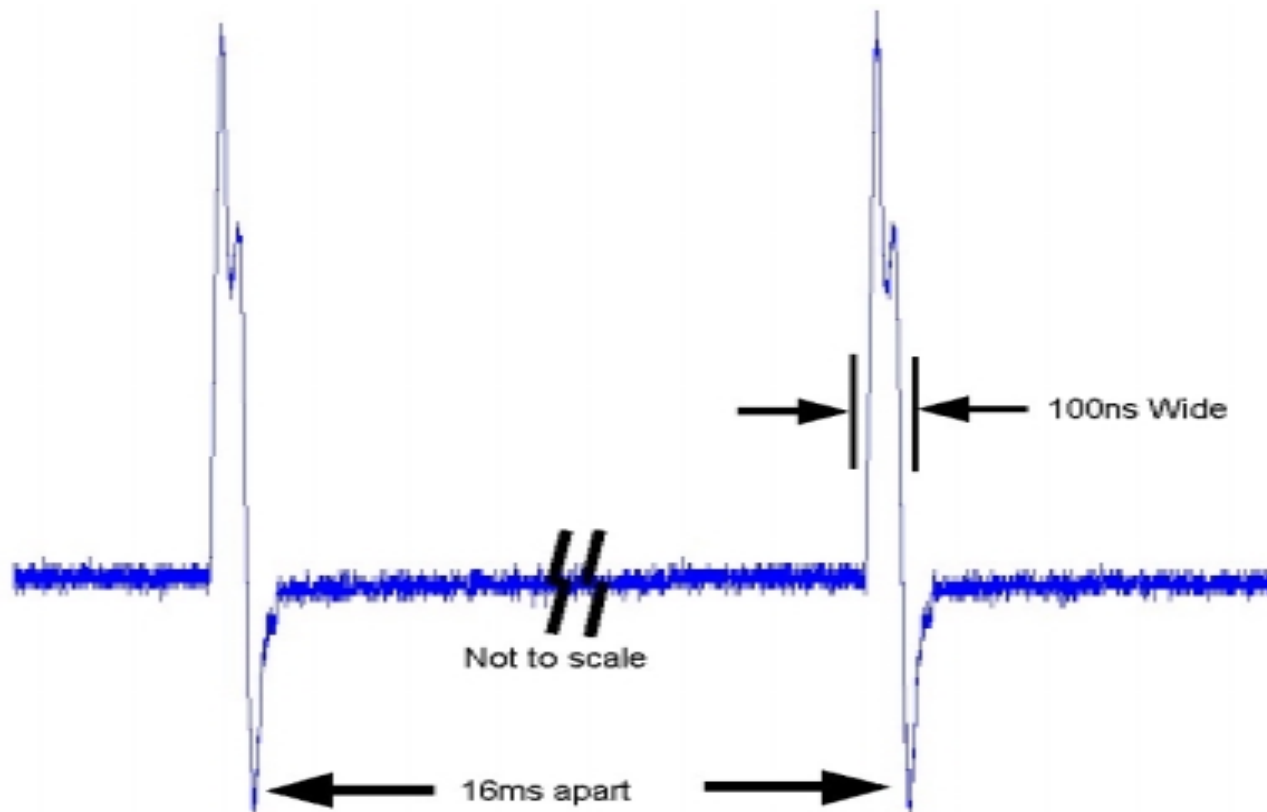
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

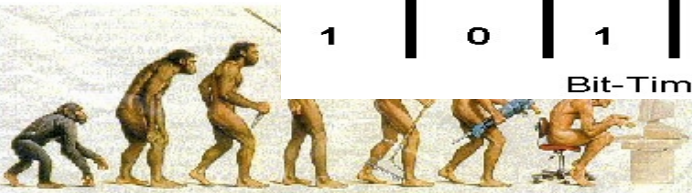
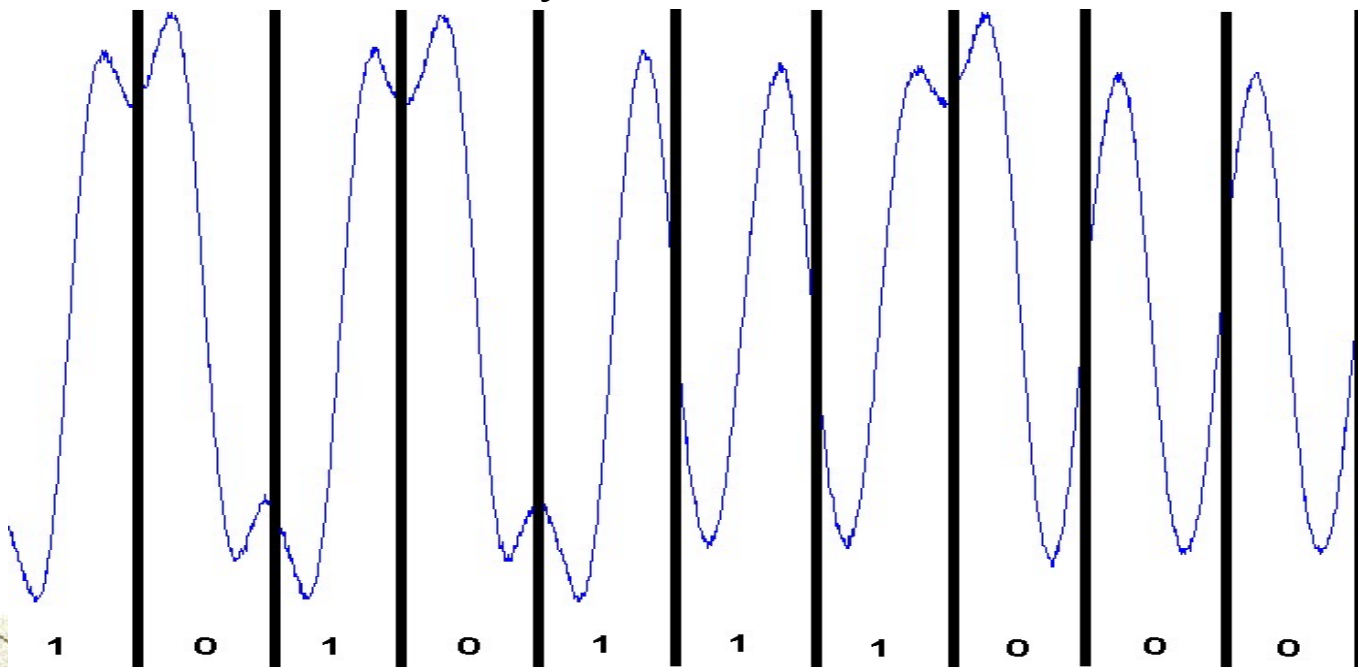
10BASE-T: Between the frames

- Idle between frames is filled with Link Test Pulses (LTPs)
- Periodically signal presence of link partner



10BASE-T: Preamble and SFD

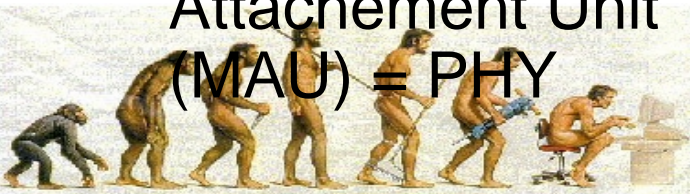
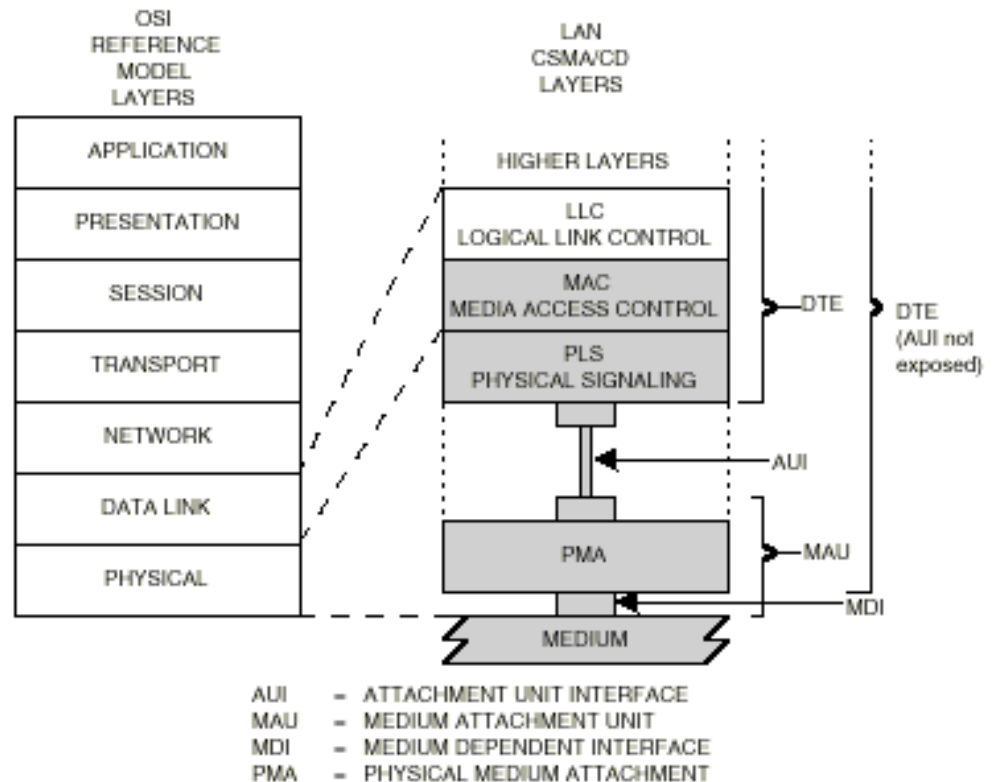
- Preamble allows device to recover clock of link partner
- Since “x” bits will be lost until clock is recovered, start frame delimiter (SFD) marks byte boundary
- Recall: Preamble: 7bytes 10101010 SFD= 10101011



ution

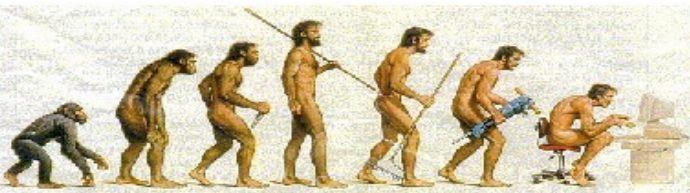
10BASE-T: AUI

- Model at right for all 10Mbps Ethernet PHYs including:
 - 10BASE-2, 10BASE-5, 10BASE-T, 10BASE-FP / -FB / -FL
- Attachment Unit Interface (AUI) allows different PHYs to be attached to the MAC. Also used in repeaters.
- Medium Attachment Unit (MAU) = PHY



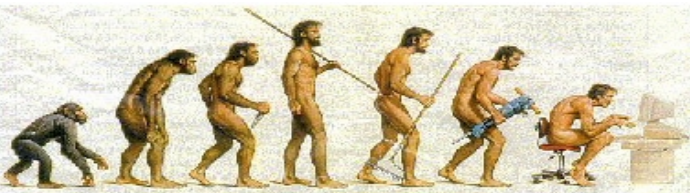
100BASE-T4 Overview

- 100Mbps
- Cat-3 or Better, 100 meter max
- Uses all 4 pair
 - Transmits on 3 pair, listens for collision on 4th.
 - Creates half-duplex only limitation in PHY
- Extinct – Why?
 - Introduced at same time as 100Base-TX(Cat-5 or better) when 10Base-T(Cat-3 or better) was prevalent.
 - 100Base-TX based on pre-existing CDDI, thus low cost TX chips emerged rapidly
 - Provided customers with option –
 - » Buy T4 equipment supporting installed base of Cat-3
 - » OR... Install new Cat-5 cable and buy TX equipment
 - » Most choose to install new cable to “future proof” network



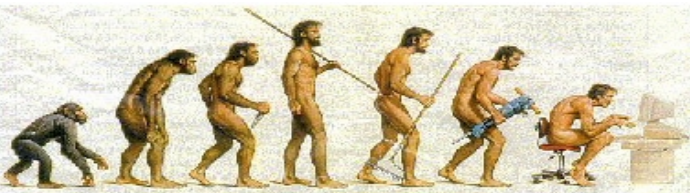
100BASE-TX Overview

- The critter commonly called Fast Ethernet
- 100 Mbps
- Category 5 cable (data-grade) or better
- Same 2 pairs for DATA as 10BASE-T (1,2 & 3,6)
- 100 meter runs (limited by cable attenuation)
 - All that's necessary to support building structured cabling
- Inherently Full Duplex (Upper layer MAC/Repeater may be half duplex though)



100BASE-TX: Data Encoding

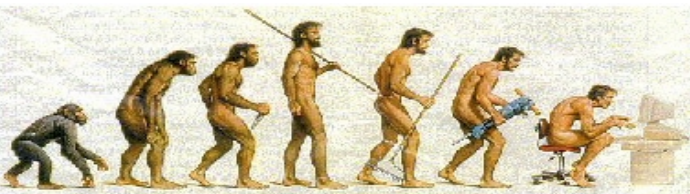
- 4B/5B Block Code
- 4B/5B means 100Mbps data requires 125Mbps on the media (a 25% speedup – resulting in 20% overhead(non-databits transmitted))
- Why is this useful? Creates control codes.
- “I” (Idle) = 11111 Which is sent continuously to provide a good clock to the link partner’s receiver
- “J” and “K” = 11000 & 10001 – Start of Frame, unique bit pattern, cannot be made from any combination of other VALID symbols
- $2^5=32$ symbols, only need 16 (2^4) symbols for data, 1 for idle, 2 for start of frame, 2 for end of frame, rest are invalid



100Base-TX: 4B/5B Tables

D A T A	PCS code-group [4:0] 4 3 2 1 0	Name	MII (TXD/RXD) <3:0> 3 2 1 0	
	1 1 1 1 0	0	0 0 0 0	Data 0
	0 1 0 0 1	1	0 0 0 1	Data 1
	1 0 1 0 0	2	0 0 1 0	Data 2
	1 0 1 0 1	3	0 0 1 1	Data 3
	0 1 0 1 0	4	0 1 0 0	Data 4
	0 1 0 1 1	5	0 1 0 1	Data 5
	0 1 1 1 0	6	0 1 1 0	Data 6
	0 1 1 1 1	7	0 1 1 1	Data 7
	1 0 0 1 0	8	1 0 0 0	Data 8
	1 0 0 1 1	9	1 0 0 1	Data 9
	1 0 1 1 0	A	1 0 1 0	Data A
	1 0 1 1 1	B	1 0 1 1	Data B
	1 1 0 1 0	C	1 1 0 0	Data C
	1 1 0 1 1	D	1 1 0 1	Data D
	1 1 1 0 0	E	1 1 1 0	Data E
	1 1 1 0 1	F	1 1 1 1	Data F

C O N T R O L	PCS code-group [4:0] 4 3 2 1 0	Name	MII (TXD/RXD) <3:0> 3 2 1 0	Interpretation
	1 1 1 1 1	I	undefined	IDLE; used as inter-stream fill code
	1 1 0 0 0	J	0 1 0 1	Start-of-Stream Delimiter, Part 1 of 2; always used in pairs with K
	1 0 0 0 1	K	0 1 0 1	Start-of-Stream Delimiter, Part 2 of 2; always used in pairs with J
	0 1 1 0 1	T	undefined	End-of-Stream Delimiter, Part 1 of 2; always used in pairs with R
	0 0 1 1 1	R	undefined	End-of-Stream Delimiter, Part 2 of 2; always used in pairs with T
	0 0 1 0 0	H	Undefined	Transmit Error; used to force signaling errors
	0 0 0 0 0	V	Undefined	Invalid code
	0 0 0 0 1	V	Undefined	Invalid code
	0 0 0 1 0	V	Undefined	Invalid code
	0 0 0 1 1	V	Undefined	Invalid code
	0 0 1 0 1	V	Undefined	Invalid code
	0 0 1 1 0	V	Undefined	Invalid code
	0 1 0 0 0	V	Undefined	Invalid code
I N V A L I D	0 1 1 0 0	V	Undefined	Invalid code
	1 0 0 0 0	V	Undefined	Invalid code
	1 1 0 0 1	V	Undefined	Invalid code



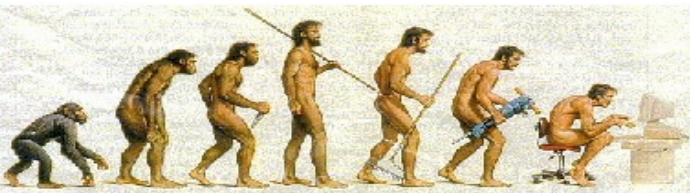
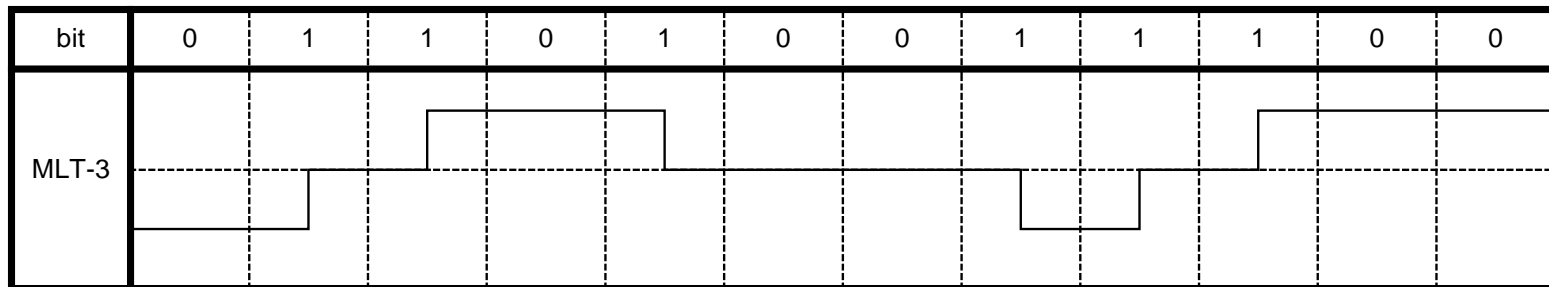
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

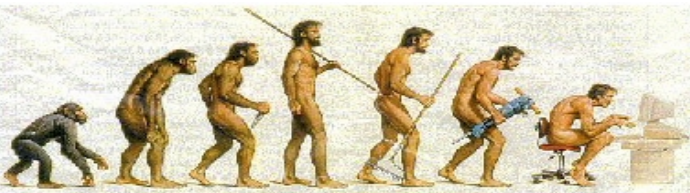
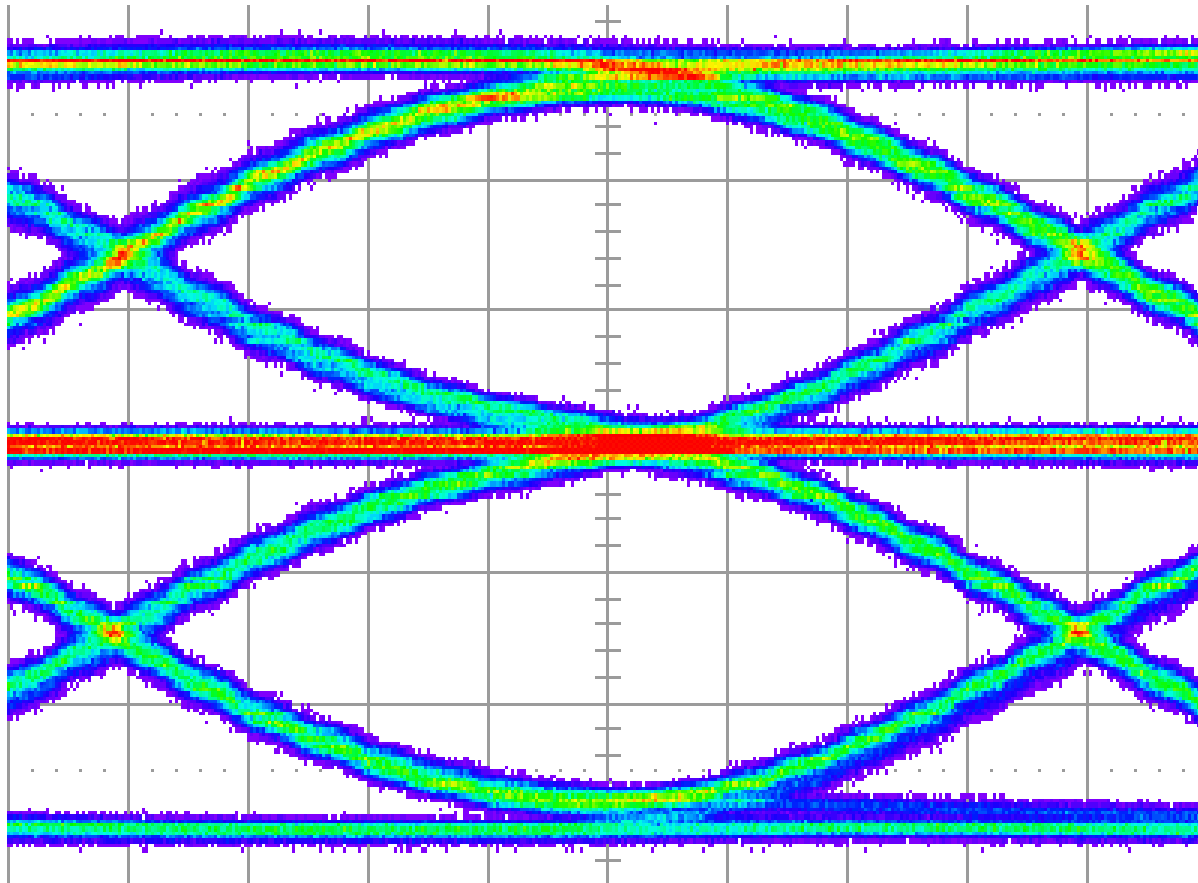
100BASE-TX: MLT-3 Coding

- 5B Idle code is 11111, at 125Mbps (NRZI) that's a 125MHz tone, recall cat-5 performance specification ends at 100MHz!
- Multi-Level Transition with 3 levels a.k.a. MLT-3
Reduces frequency of 5B encoded data.
- If data is “1”, then transition from current level to next level. ie 1111= $+1, 0, -1, 0$ reducing freq by 1/4
- If bit data is “0”, then don't transition



100Base-TX: MLT-3 Eye Diagram

- 1-Bit Time shown, with all possible transitions



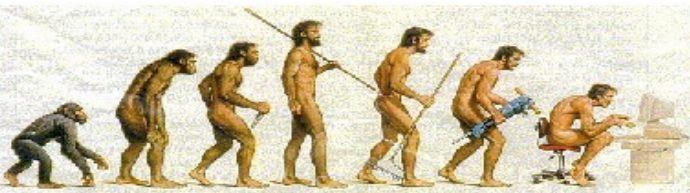
The Ethernet Evolution

10 to 10,000 Mbps

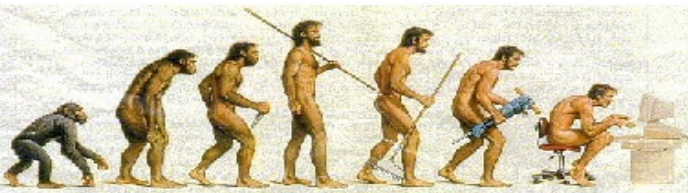
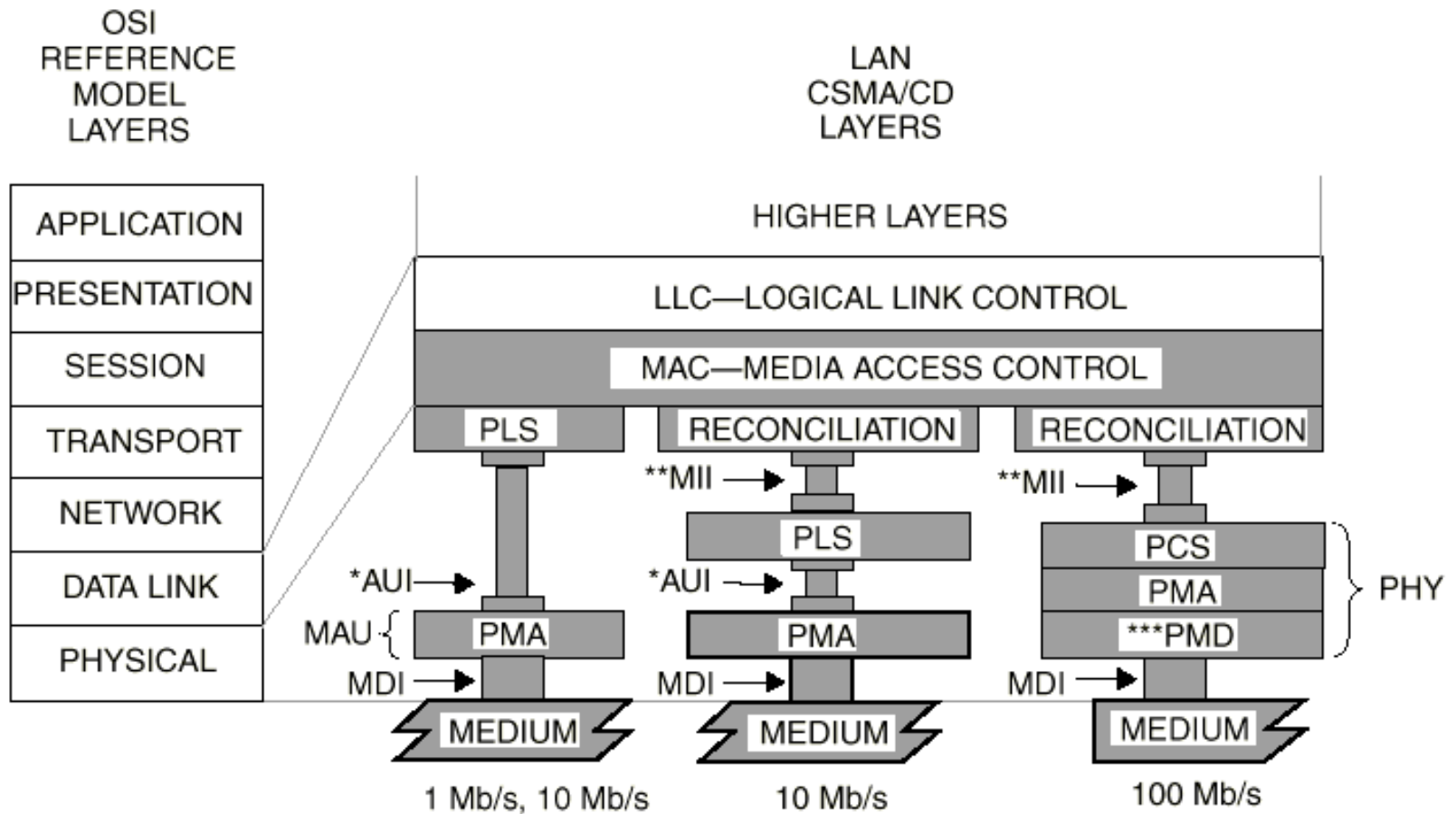
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

100BASE-TX: Scrambling

- MLT-3 successfully reduces frequency of 5B encoded data, but recall the 5B idle stream is sent continuously between frames.
- After MLT-3 encoding, Idle went from a 125MHz tone to a 31.25MHz tone ($125/4$). Since most of the transmitted energy is at this frequency, it radiates strongly, which the FCC does not like, hence...
- Scrambling of the 5B encoded stream is required.
- Prior to MLT-3 encoding, the 5B stream is pseudo-randomly scrambled, this breaks up the repeating 1s and spreads the transmitted energy across many frequencies, thus no single frequency has much energy.

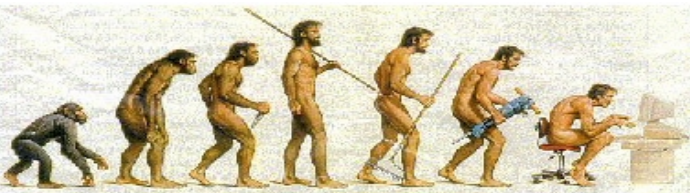


100BASE-TX: Media Independent Interface



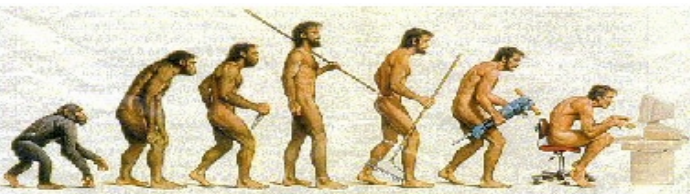
100BASE-T2

- 100Mbps
- Cat-3 or Better
- Uses only 2 pair
 - Uses 5 level signaling
 - Transmits and receives on both pairs simultaneously
 - Allows use of other pairs.
 - Full-duplex capable
- Extinct – Why?
 - Limited market potential due to 100Base-TX



1000BASE-T Overview

- 1000Mbps
- Cat-5 or Better
- Uses all 4 pair
 - Uses 5 level signaling
 - Transmits and receives on both pairs simultaneously
 - Full-duplex capable
- Emerging –
 - Growing number of PHY chip vendors
 - 10/100/1000 single chip solutions and multiport chips



10
0

The Ethernet Evolution

10 to 10,000 Mbps

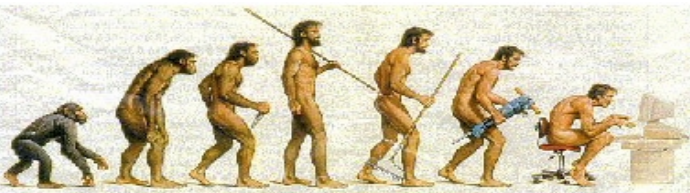
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

The 1000BASE-T Solution

- Start with 125 MHz signaling rate
125Mbps
 - Exactly like 100BASE-TX
- Transmit/Receive on all 4 pairs of cable
 - Similar to 100BASE-T4

x 4
500Mbps
- Use multilevel signaling – PAM5
 - Similar to 100BASE-TX (3-level MLT-3 signaling)
 - Exactly like 100BASE-T2 (5-level PAM5 signaling)
 - Allows 2 bits per symbol

500Mbps
x 2
1000Mbps
- Allow simultaneous bi-directional signaling on all 4 pair
 - Exactly like 100Base-T2
 - Allows full-duplex communication at 1000Mbps



10
1

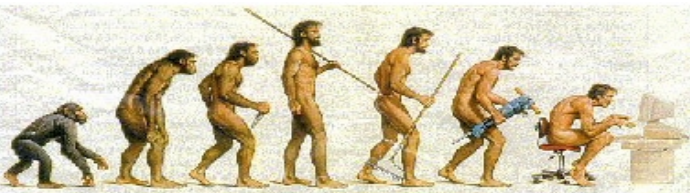
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

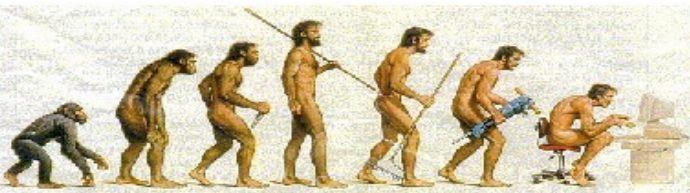
1000Base-T: 4D-PAM5

- 4 Dimensional Pulse Amplitude Modulation with 5 levels
- 5 levels on 4 pairs yields 625 possible symbols (5^4)
- To encode 8bits every 8ns (125MHz), only 256 symbols are needed (2^8).
- Remaining symbols can be used for control characters (idle, start of packet, end of packet, etc).
- Data frames are transmitted using a convolutional (Trellis) code
 - Adds structure to the transmitted symbols needed for Viterbi Decoding

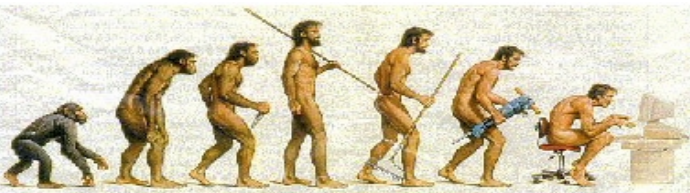
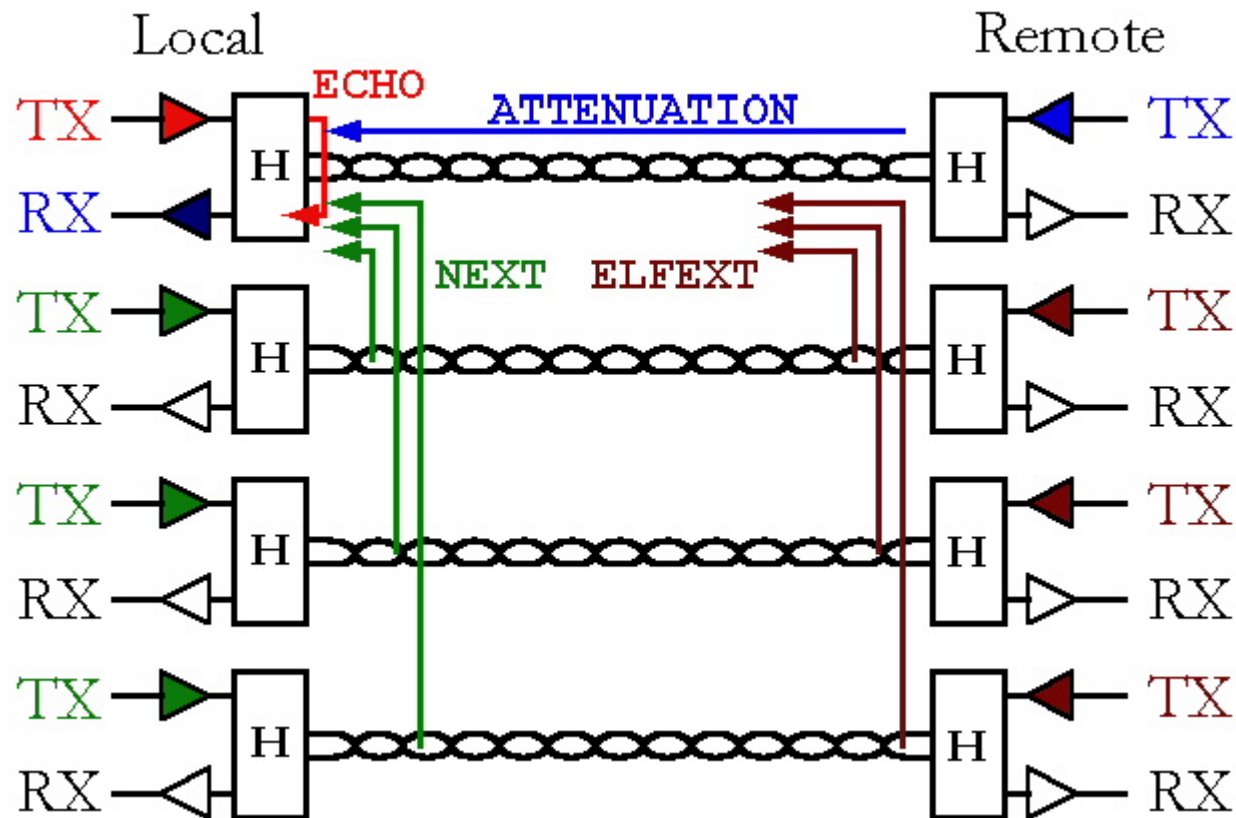


1000Base-T: Error Correction

- Viterbi Decoder
 - Takes the last few received symbols
 - Calculates the most likely sequence of symbols
 - A symbol error results in an impossible or unlikely sequence
 - The most likely sequence probably contains the correct symbols
 - “Most likely sequences” are known in advance due to the trellis code structure



1000Base-T: Noise Environment



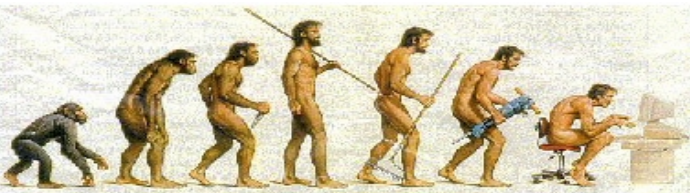
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

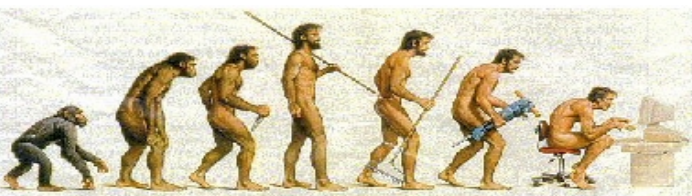
1000Base-T: DSP

- Digital signal processing (DSP) techniques can reduce effects of echo and NEXT
 - These are caused by local transmissions
 - Tap the signal on all 4 transmitters
 - Cancel out these signals from the received waveform
- ELFEXT is not easily cancelled with DSP
 - Originates from remote transmitters
 - Relies on cable and connectors to meet new Cat5 specification, so that ELFEXT is minimal
- Alien Crosstalk also unpredictable
 - Unknown source
 - Thus, cannot be cancelled
- 25-pair bundles not allowed!
 - Crosstalk between bundled pairs too high

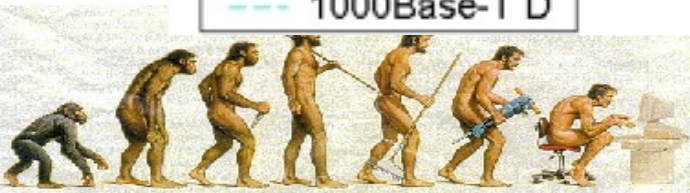
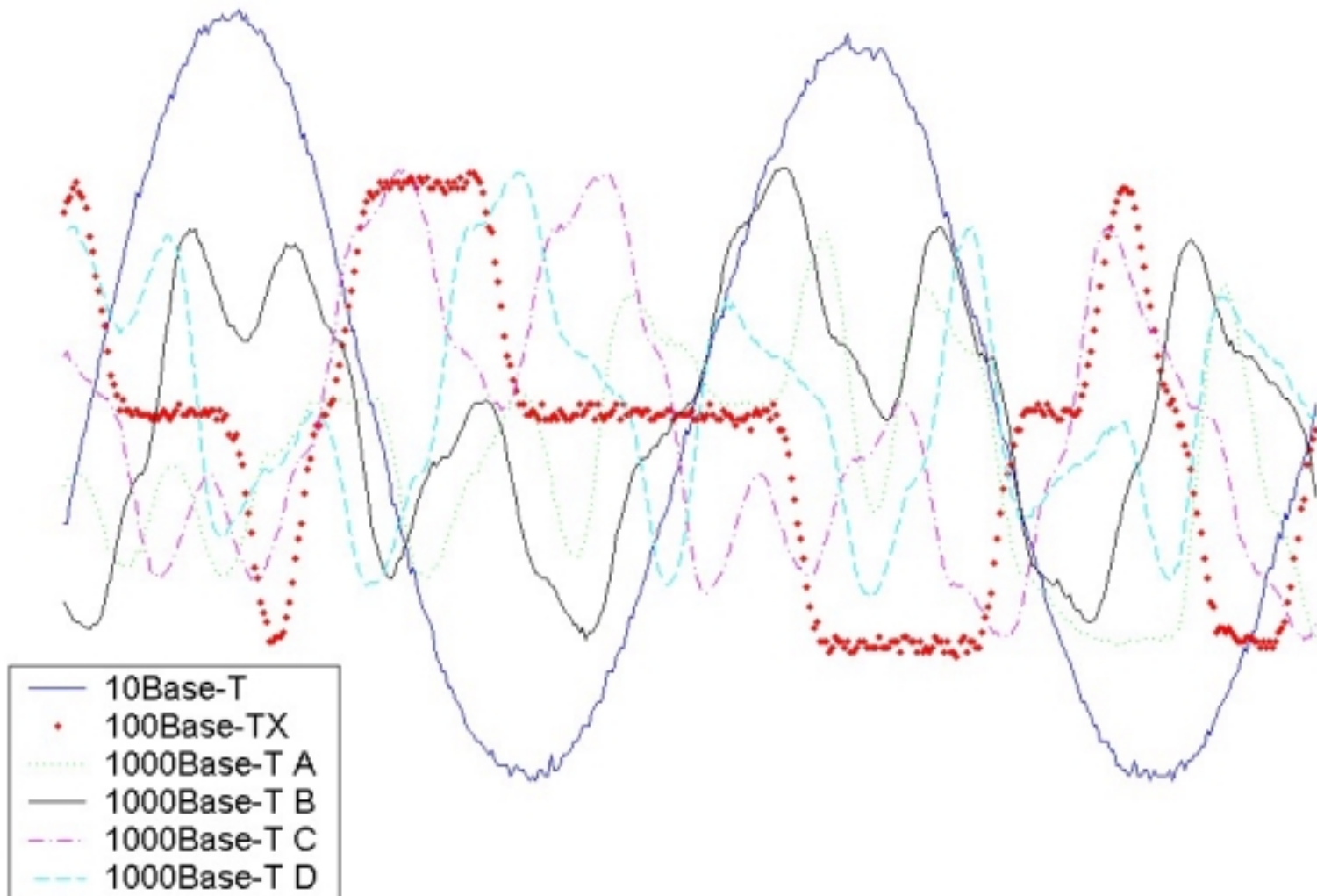


1000Base-T: Master/Slave Timing

- Synchronization to implement adaptive filters for canceling echo and NEXT
 - Need a common master clock to reference
 - Symbols transmitted & received at the same rate
- Link pair must have a Master and a Slave
 - Master uses internal clock to transmit
 - Slave recovers the master clock from the data received
 - Slave uses recovered clock to transmit its data
- Need a means of deciding who is who –
 - uses Auto-Negotiation (more on that later)

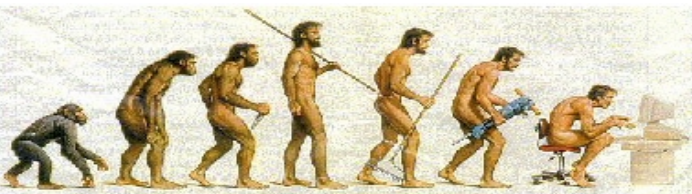


10, 100, 1000 Signal Comparison



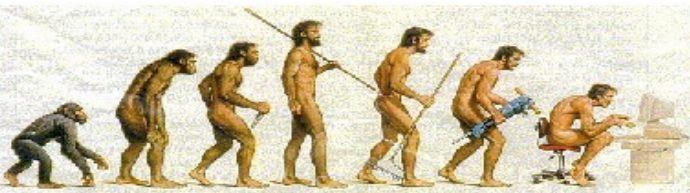
UTP Copper Phys Wrap-up

- Combo PHYs
 - 10/100 PHYs – 10Base-T & 100Base-TX
 - 100/1000 PHYs – 100Base-TX & 1000Base-T
 - 10/100/1000 PHYs – emerging
- Quad / Hex / Octal PHYs
 - Reduces chip and other component count required to manufacture multiport devices, thus lower cost.
 - Most multi-phy devices are also Combo PHYs



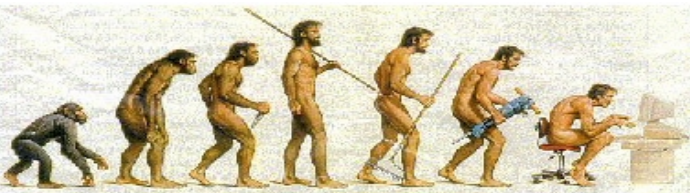
Fiber Evolution Summary

- 1987 – Fiber Optic Inter-Repeater Link (FOIRL):
 - Link 10Mbps repeaters via multimode fiber (2km max)
- 1993 – 10Base-F:
 - 10Mbps to end stations via multimode fiber (2km max)
- 1995 – 100Base-FX:
 - Capitalizing on FDDI work
 - 100Mbps via multimode fiber (2km max) to repeaters or end stations
- 1998 – 1000Base-SX / LX:
 - Capitalizing on FibreChannel work, but spedup
 - 1000Mbps



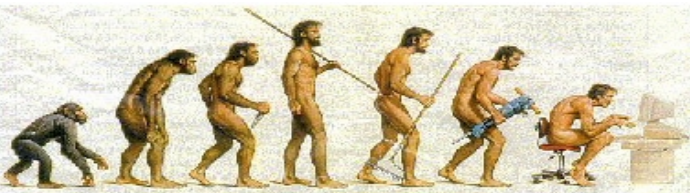
10Mbps Fiber Overview

- All use multimode fiber, driven by LEDs
- FOIRL: Fairly common
 - FOIRL = Fiber Optic Inter-Repeater Link
 - Typical use: FOIRL Medium Attachment Unit (MAU) attached to Attachment Unit Interface (AUI) of repeaters w/o built in fiber support
- 10BASE-F
 - Catch all for 10Base-FP, 10Base-FB, 10Base-FL
- Distances:
 - 10Base-FP: 500m max
 - All others: 2000m max – limited by cable attenuation
- Same speed, but different techniques – identical type of MAU required on both sides of fiber



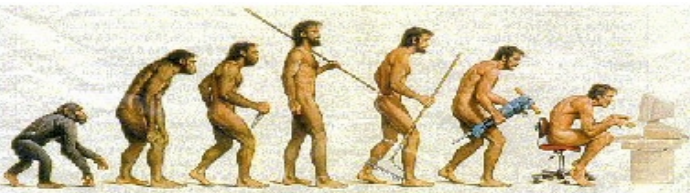
100BASE-FX Overview

- 100 Mbps
- Multimode Fiber
- Inherently Full Duplex (Upper layer MAC/Repeater may be half duplex though)
- Full Duplex Distance: 2000 meter runs
 - Limited by cable attenuation
- Half Duplex Distance: 412 meters
 - Point-to-point link, no repeaters
 - Limited by collision domain



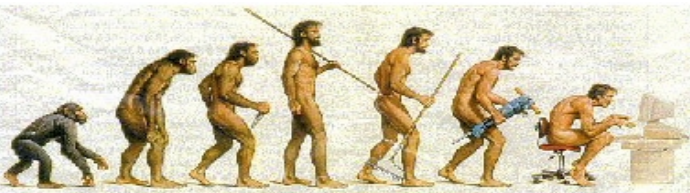
100BASE-FX: Similar to -TX

- Like all 100Mbps Fast Ethernet, supports Media Independent Interface (MII)
- Like 100BASE-TX, uses 4B/5B Coding to encode data and idle
- Similarity ends here though – No need for:
 - Scrambling
 - MLT3 Encoding
- Why?
 - MLT3 is unnecessary due to high bandwidth of fiber
 - Scrambling is unnecessary as fiber does not radiate, thus no possibility of interference.



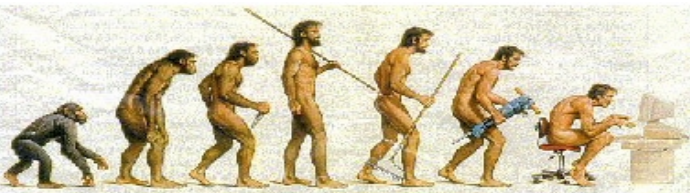
1000BASE-SX/LX Overview

- 1000 Mbps
- Multimode Fiber for SX or LX
- Singlemode Fiber for LX only
- Inherently Full Duplex (One fiber for Tx, one for Rx)
 - Upper layer MAC may be half duplex, but this is highly unlikely
- Distance:
 - Tricky issue



1000Base-SX/LX: Distance and Launch Cable

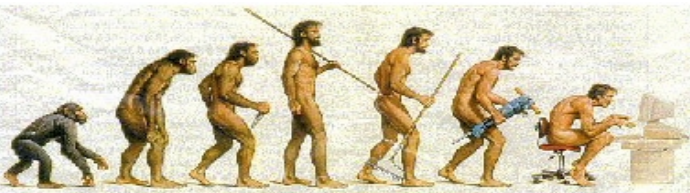
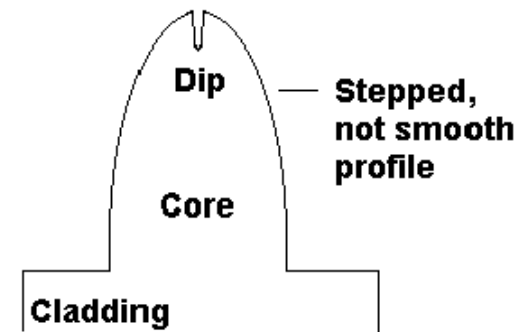
- Remember from the physical media section the following two slides...



Recent Modal Dispersion Issues

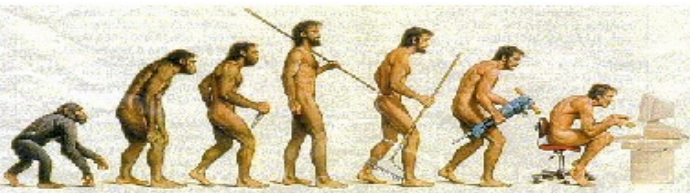
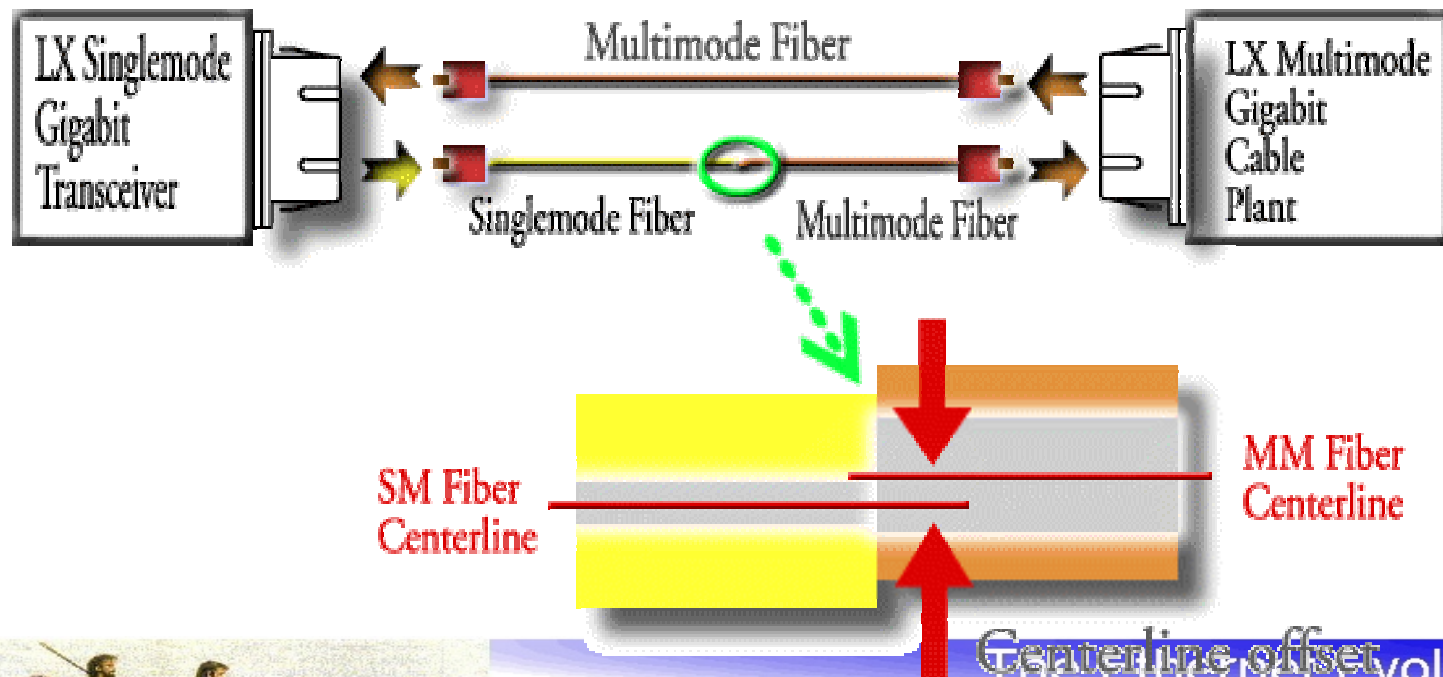
- As a side-note: During the development of Gigabit Ethernet, it was discovered that a sizeable percentage of installed (old) multimode graded index fiber had a deviation in the center of the core, as depicted at left. This was determined to cause increased modal dispersion (DMD, differential modal delay as the gig folk called it)
- The result of this discovery of a flaw in the installed media was the subsequent reduction in supported link lengths for GbE on installed multimode fiber.
- Which leads to the following new term...

Graded-Index Multimode Fiber Index Profile



Launch Conditioning

- Many options exist to minimize the DMD effect, one of which is to use a single-mode “pig-tail” to condition/reduce the modes launched into the multi-mode segment, by launching them off-center (and thus avoiding the index “notch” in the center of the installed fibers)



1000Base-SX Distances

Table 38-2—Operating range for 1000BASE-SX over each optical fiber type

Fiber type	Modal bandwidth @ 850 nm (min. overfilled launch) (MHz · km)	Minimum range (meters)
62.5 μm MMF	160	2 to 220
62.5 μm MMF	200	2 to 275
50 μm MMF	400	2 to 500
50 μm MMF	500	2 to 550
10 μm SMF	N/A	Not supported

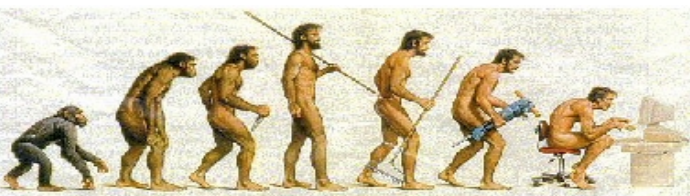
- Modal Bandwidth refers to the quality of your installed MM fiber.
- Recall, SingleMode Fiber (SMF) is not supported for SX
- Greater distances are likely, but not recommended by the standard



1000Base-LX Distances

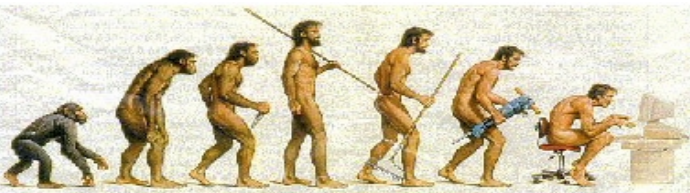
Table 38-6—Operating range for 1000BASE-LX over each optical fiber type

Fiber type	Modal bandwidth @ 1300 nm (min. overfilled launch) (MHz · km)	Minimum range (meters)
62.5 μm MMF	500	2 to 550
50 μm MMF	400	2 to 550
50 μm MMF	500	2 to 550
10 μm SMF	N/A	2 to 5000



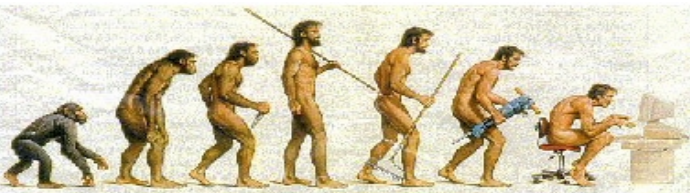
1000Base-SX/LX: Data Encoding

- 8B/10B Block Code
- Code developed by IBM in the 80s, used by FibreChannel.
- Like 4B/5B, has 20% overhead, used not only for control characters (idle, start of packet, end of packet) but also for data code redundancy.
- Running Disparity:
 - Every data code has a + and a – running disparity version
 - The form of each data code determines whether the NEXT data code should be + or –
 - If data is received that does not follow the running disparity rules, then an error has occurred.



Autonegotiation

Clauses 28 and 37 of 802.3



12
0

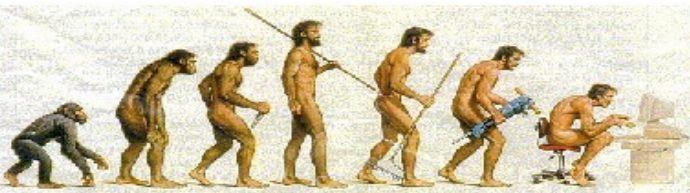
The Ethernet Evolution

10 to 10,000 Mbps

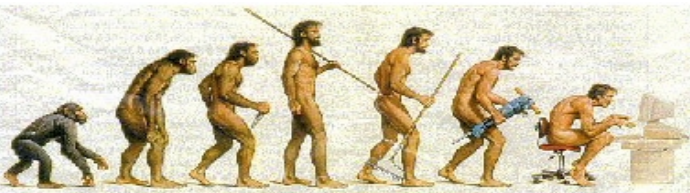
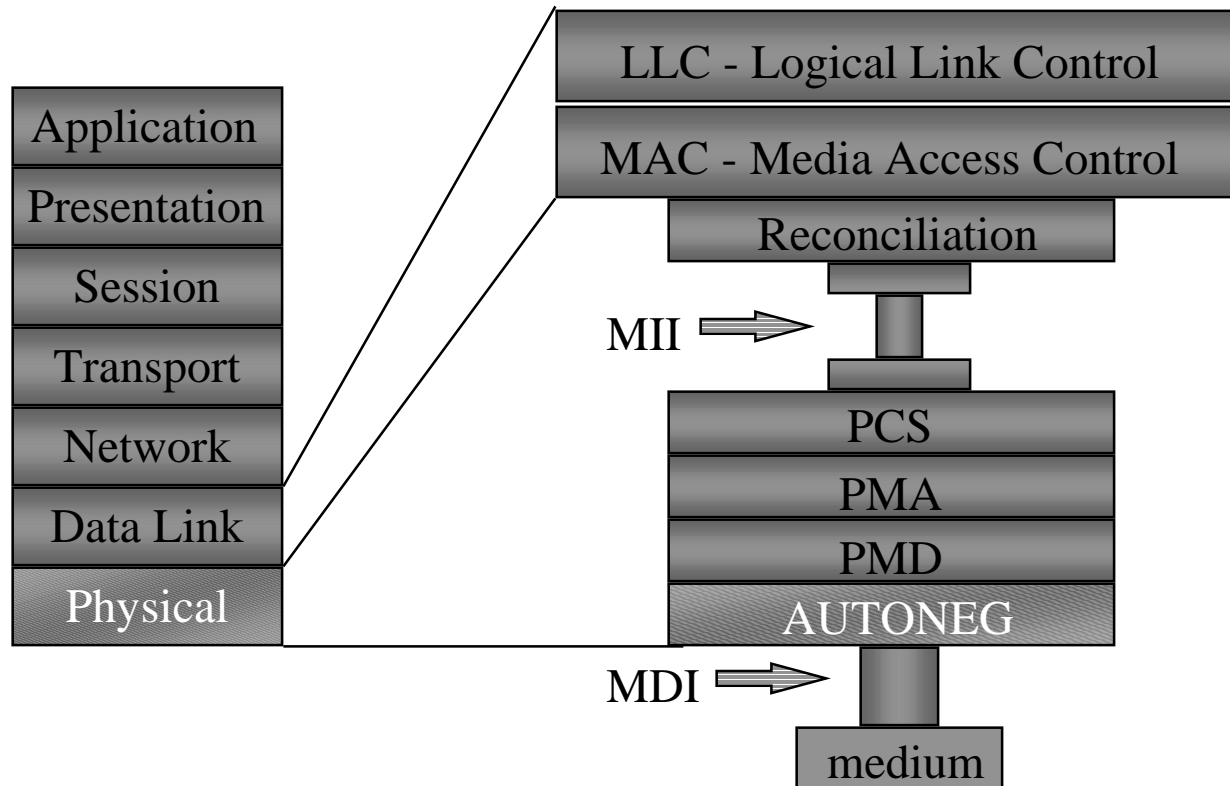
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Autonegotiation vs. Autosensing

- Autonegotiation
 - standardized speed handshake
 - auto-configures to best possible link (e.g., 100 full duplex)
 - still links with older or non-autoneg devices
 - sometimes causes autosensing (NOT autonegotiating) devices to link at 10 and not 100
 - user error/misunderstanding cause problems
- Autosensing/Speed Detection
 - several different proprietary methods
 - only auto-configures to 10 or 100, not duplex settings
 - creates many interoperability headaches

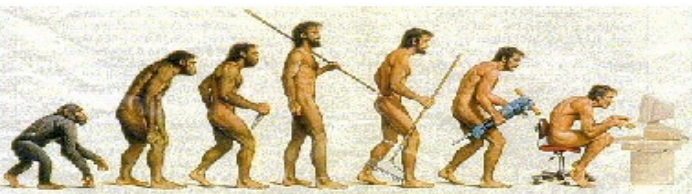


ISO-OSI model

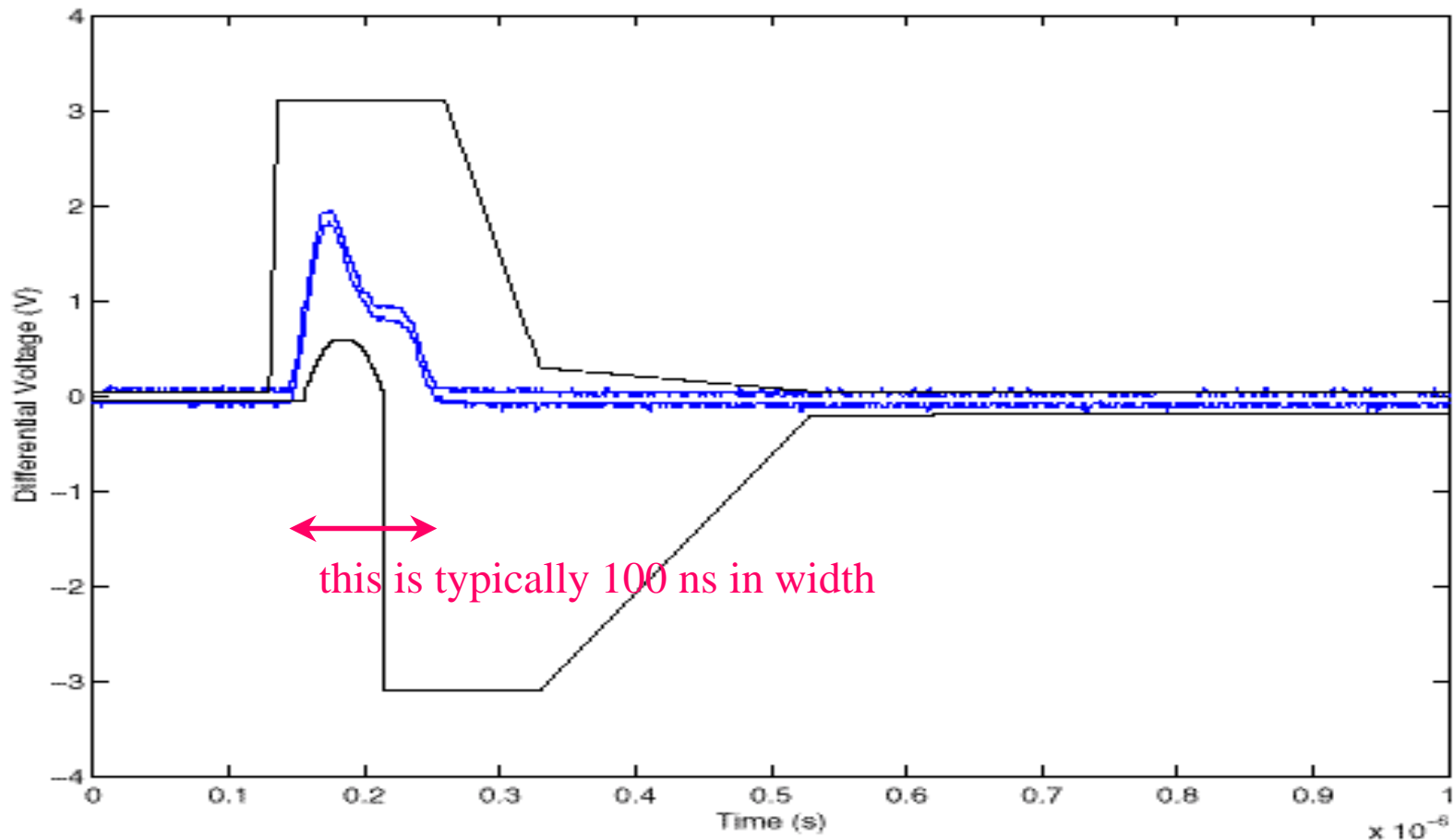


Autonegotiation - How?

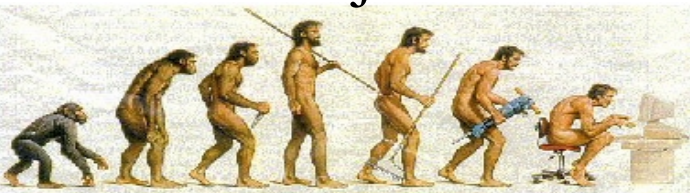
- Constantly sends out 10base-T Link Test Pulses before linking
- The pulses are grouped together in defined “words” that convey meanings, such as “I can do 100 half and full duplex”
- Older 10base-T devices just think they’re LTPs
- Autoneg devices understand them as words and exchange handshake info to link at best possible link
- If the autoneg device sees regular LTPs coming in (not autoneg words), it just links at 10 half duplex
- If the autoneg device sees Fast Ethernet IDLE stream coming in, it just links at 100 half duplex.
- Unfortunately, this only works on copper for 10/100 - it’s supported on fiber and copper for gig (i.e., there is no autonegotiation for 100base-FX)



NLP (Normal Link Pulse)



- Looks just like a 10base-T Link Test Pulse (LTP)



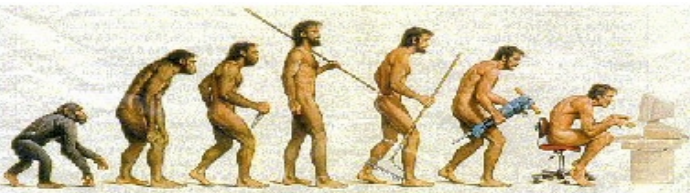
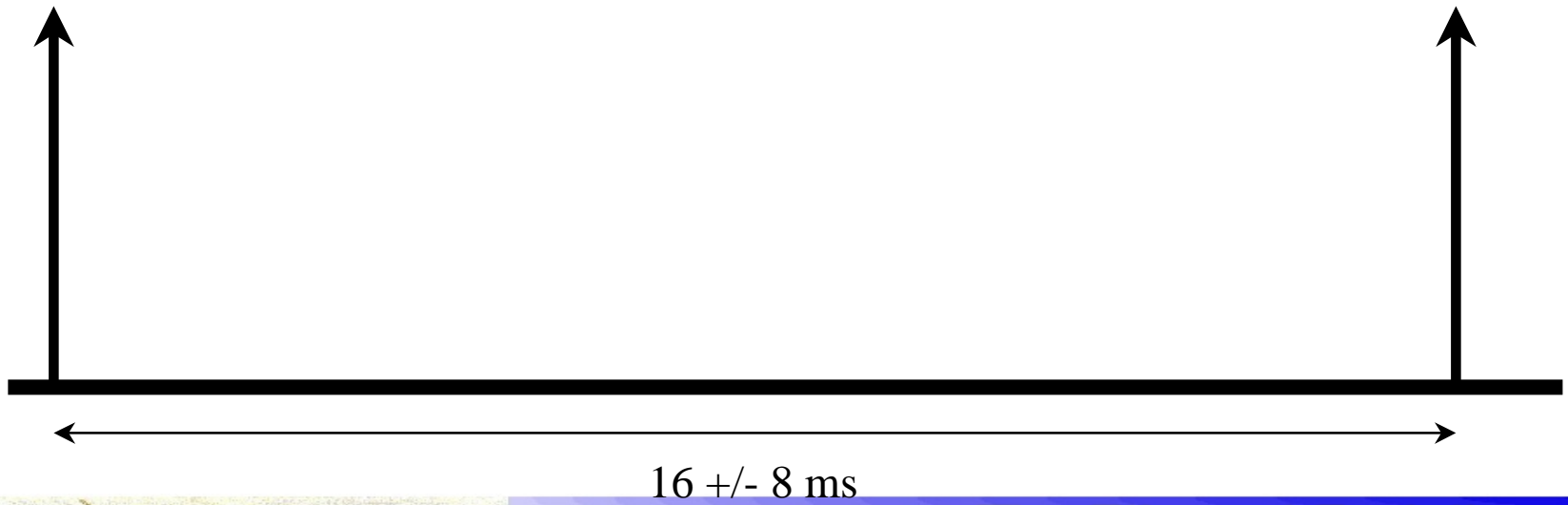
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

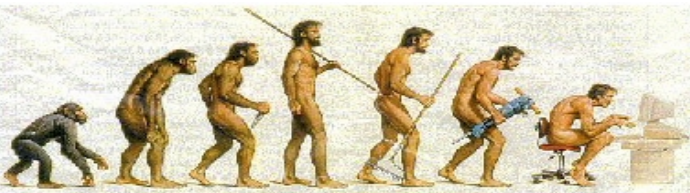
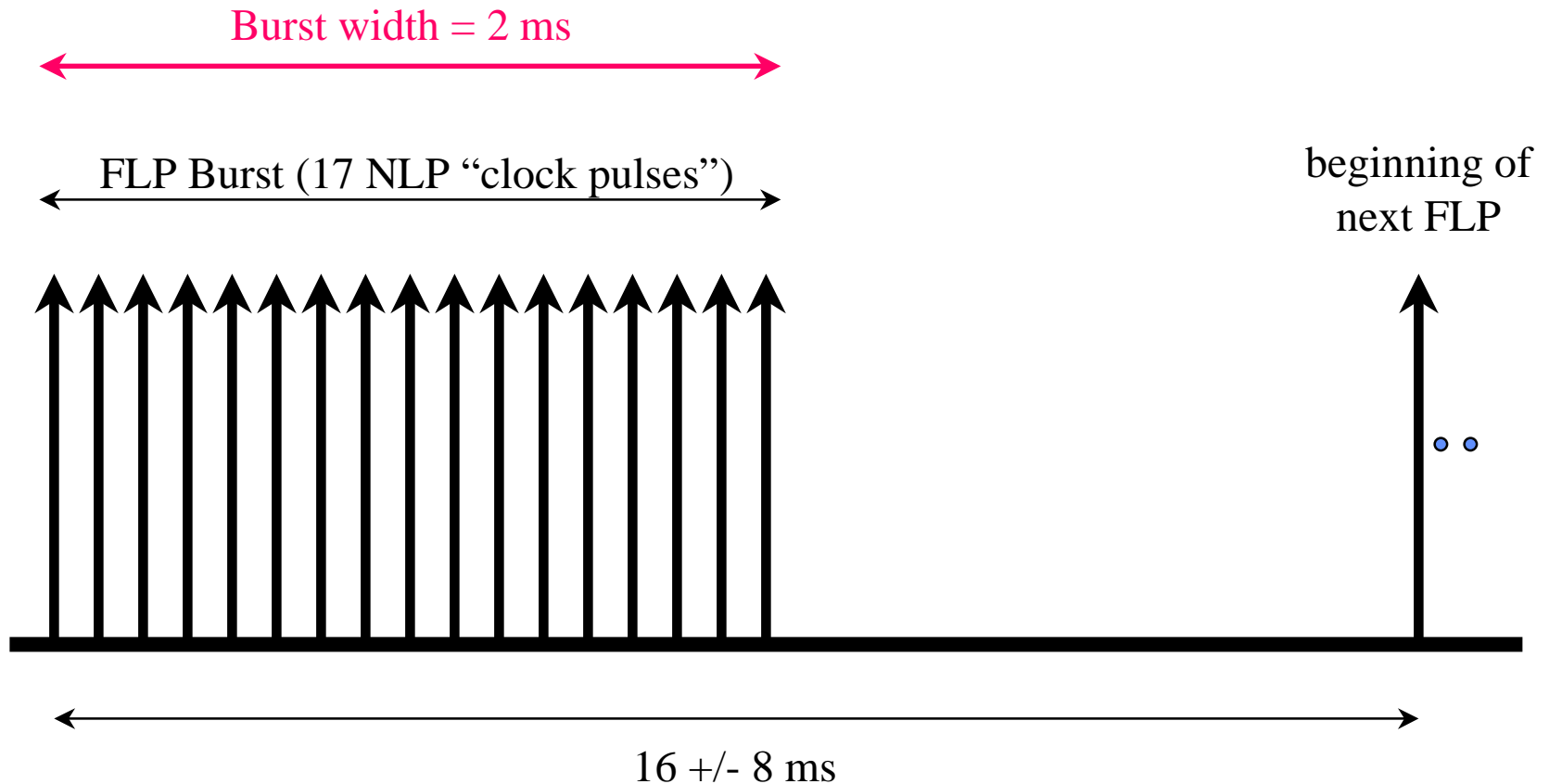
When do we care about NLPs?

- When the NLPs are spaced apart they look just like normal 10base-T Link Test Pulses (LTPs)
- So an old 10base-T legacy device will see them as normal link
- But a legacy device won't care if we send the NLPs more often than normal, so...



When they are FLP Bursts!

- Stick a bunch of NLPs together and use it to signal info (like Morse code)



12
6

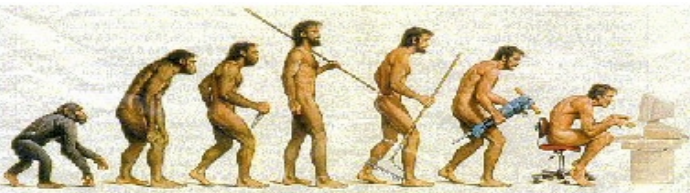
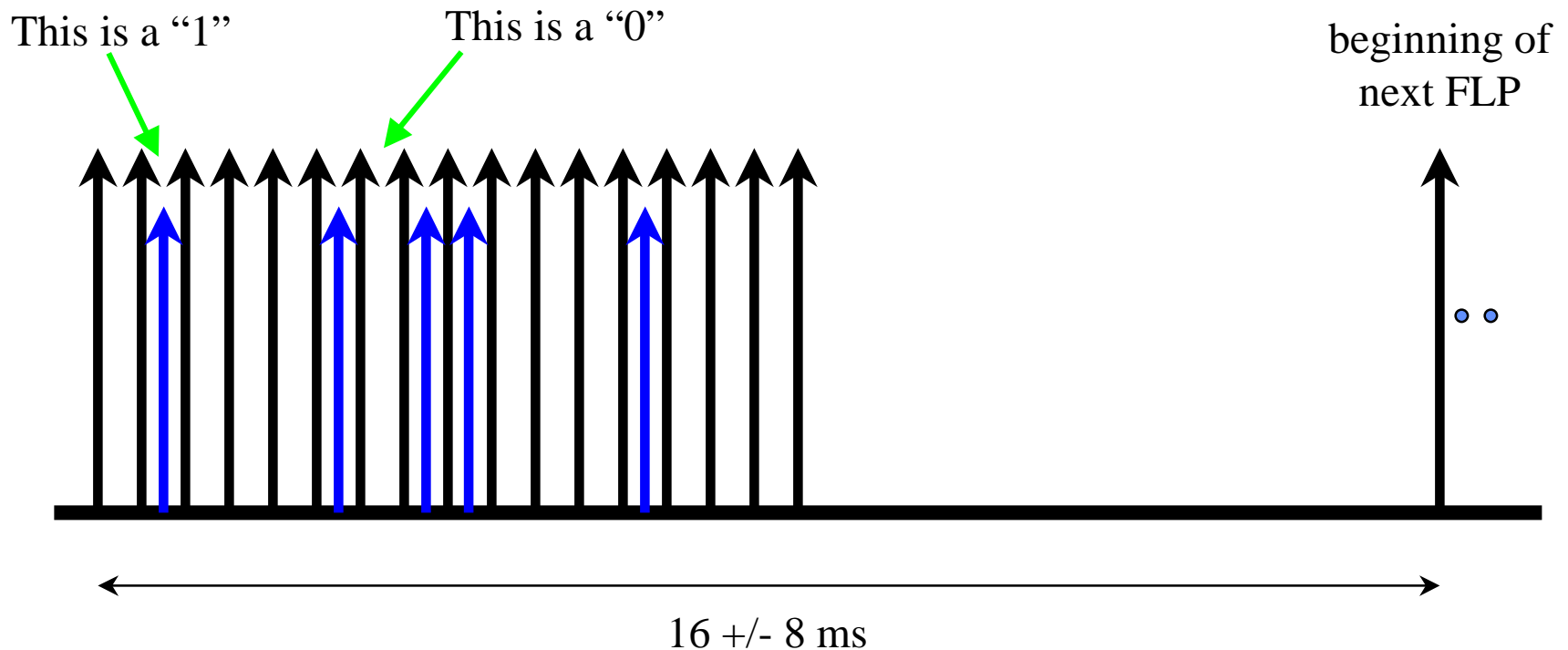
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Data Pulses

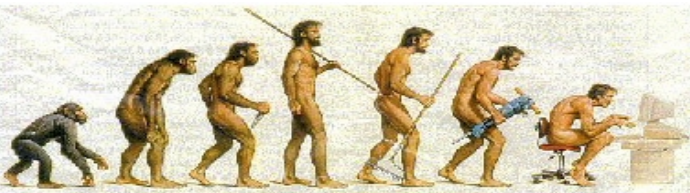
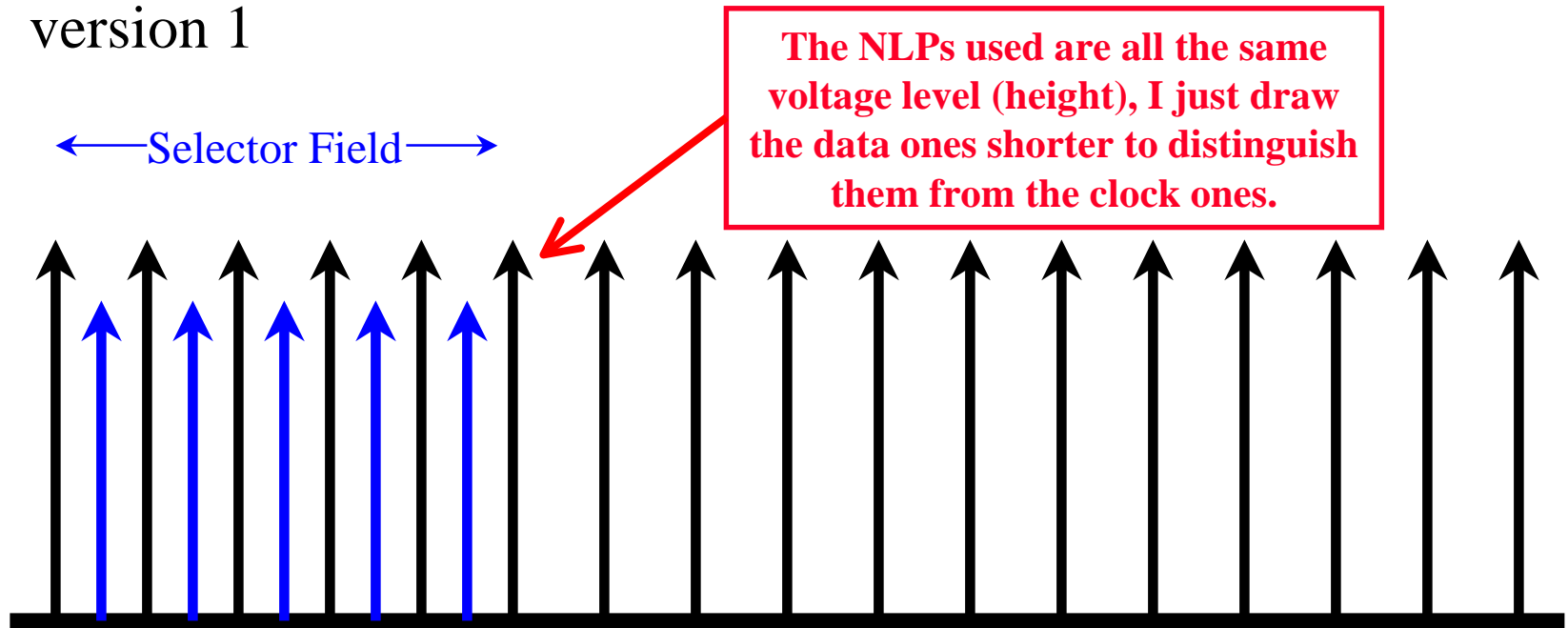
- In between the “clock pulses” we can put a data pulse to signify a value of 1, or not put one to signify value of 0



right....What's an FLP for again?

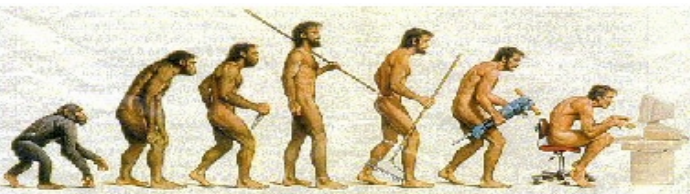
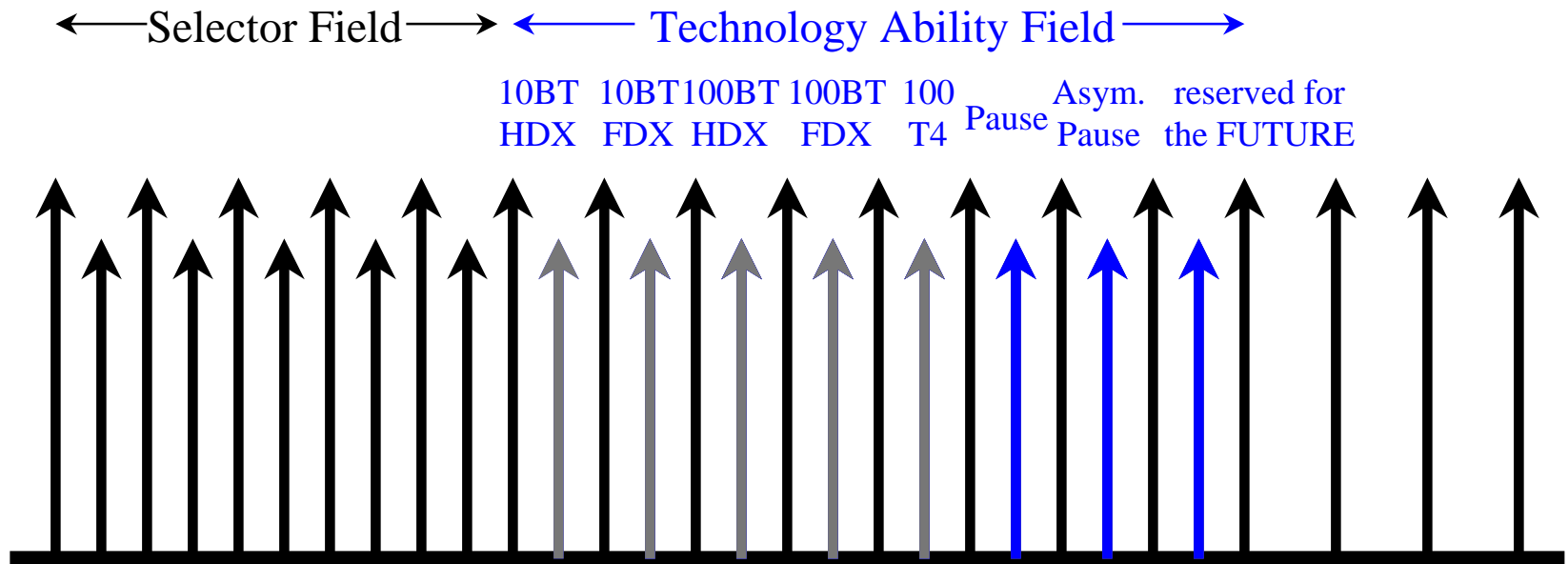
The data fields are pre-defined for specific meanings

The selector field is used for version info, such as “802.3 version 1



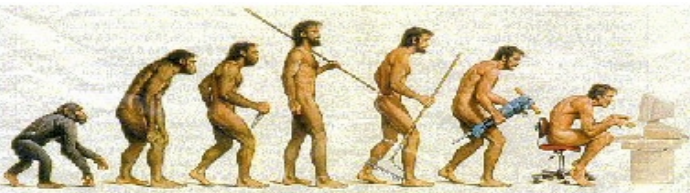
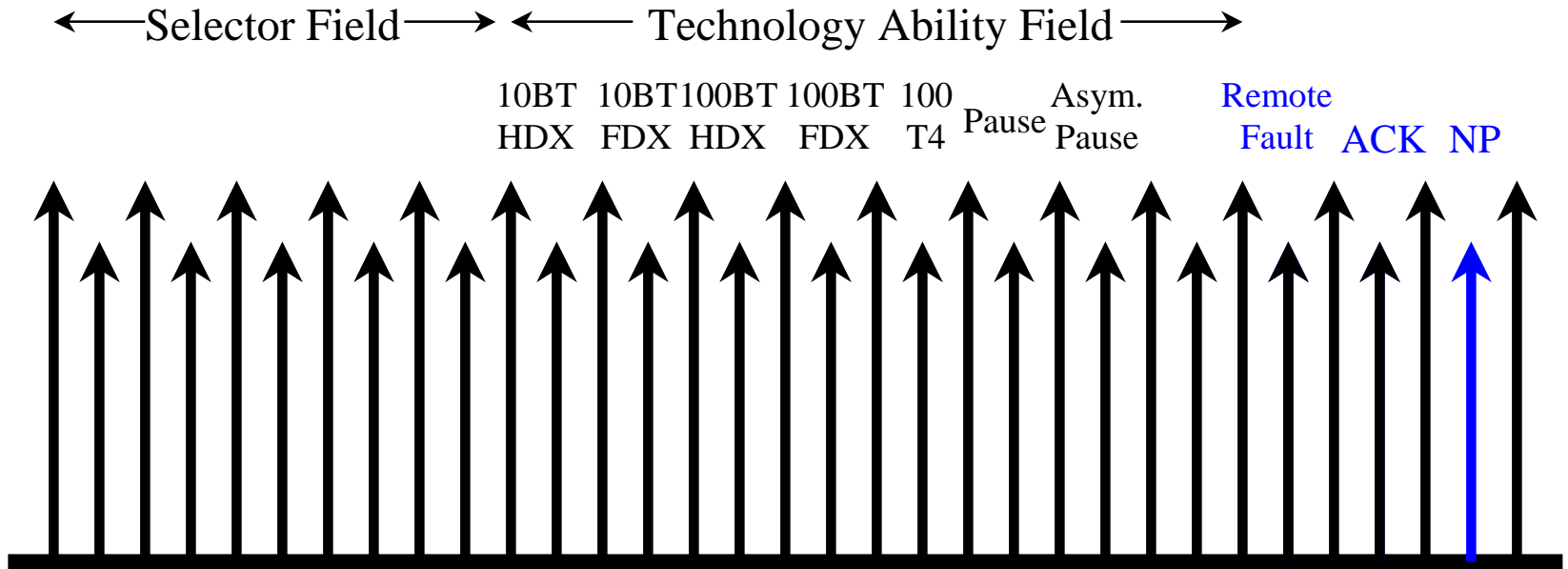
right....What's an FLP for?

The Technology Ability Field is used to signal which modes the device can handle, for example 100base-T Full Duplex

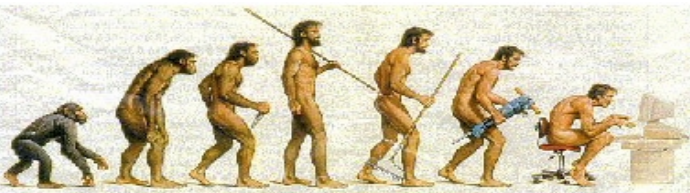
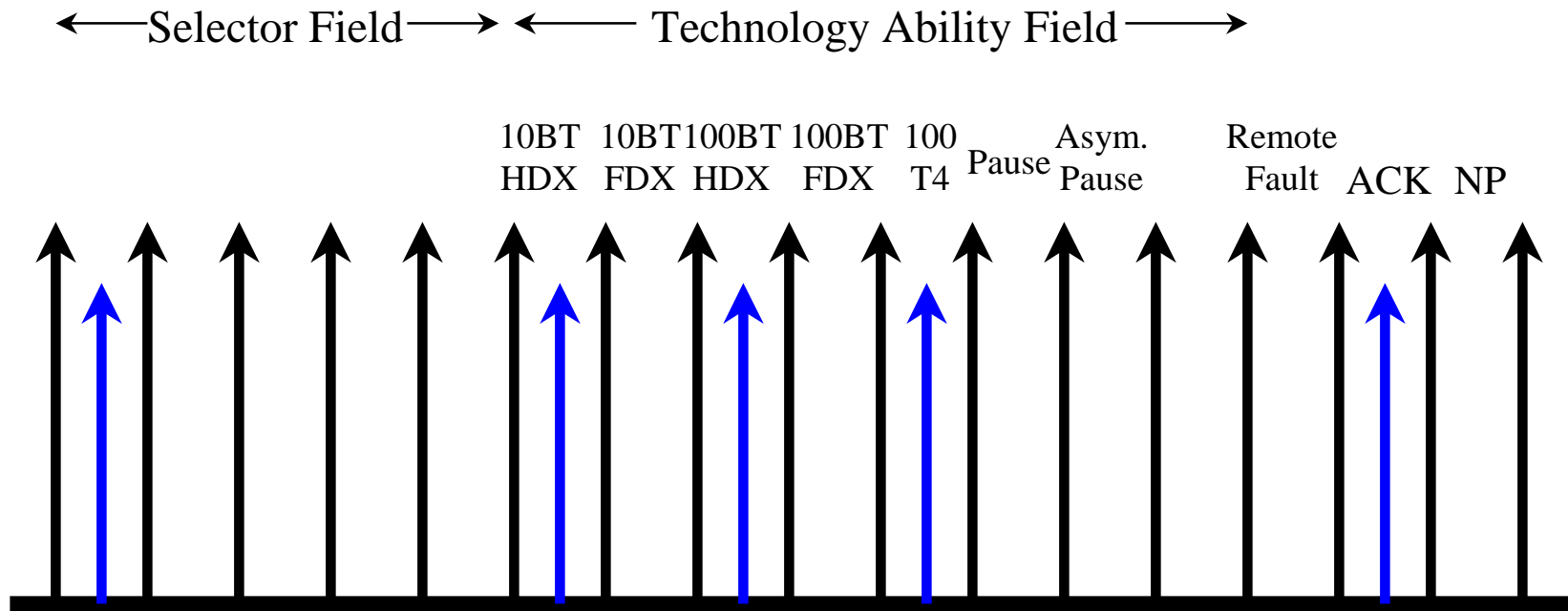


right....What's an FLP for again?

The last fields are used to tell the link partner of a remote fault, acknowledge the receipt of its partner's FLPs, or that there are more “pages” of FLP fields to send



A Sample FLP



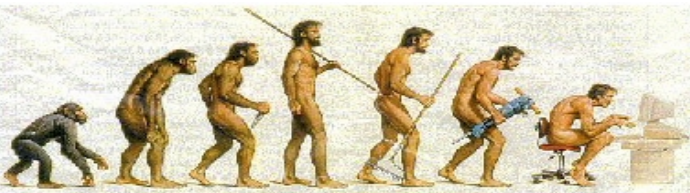
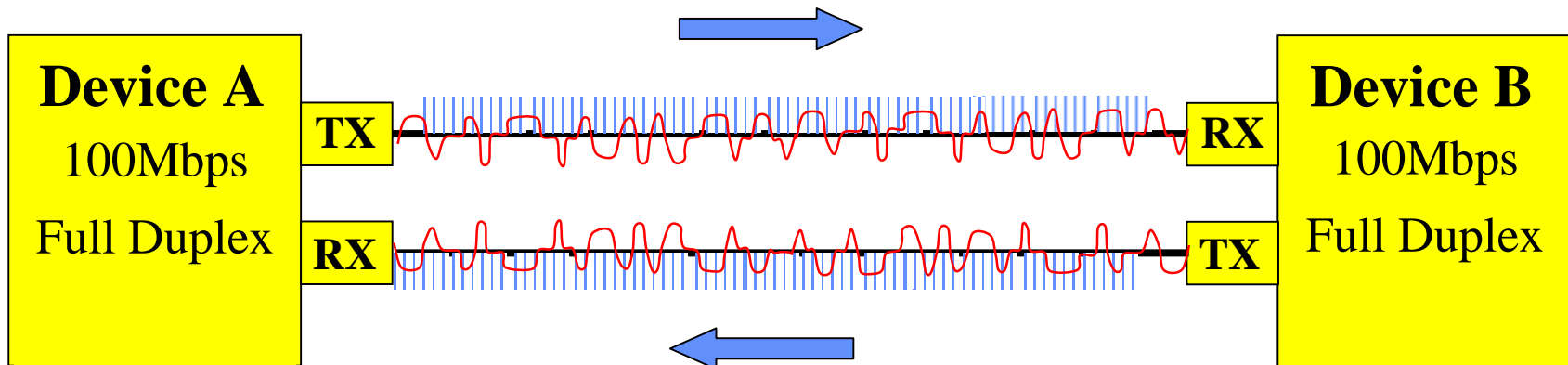
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

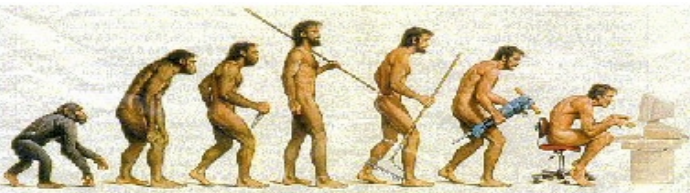
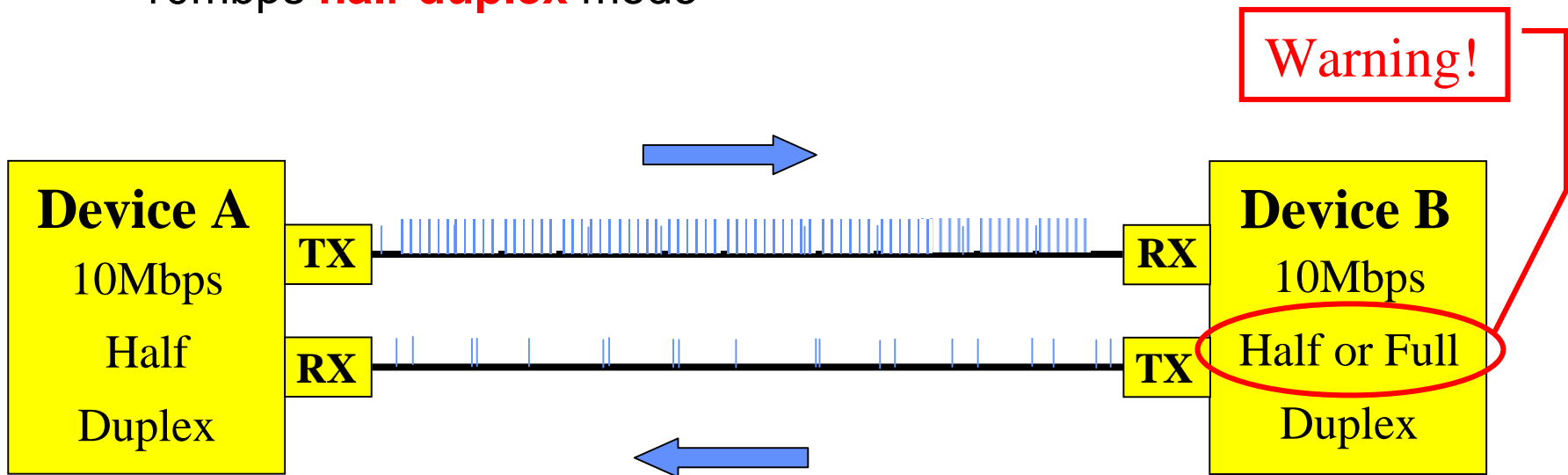
How It Works: 2 Aneq Devices

- Device A sends FLPs out, while Device B does the same
- Each device receives the other's FLPs and sets their ACK bit to true (a value of "1")
- They then choose the best possible mode that both support and start transmitting IDLE, and it's linked!



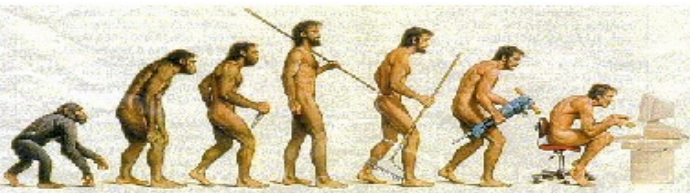
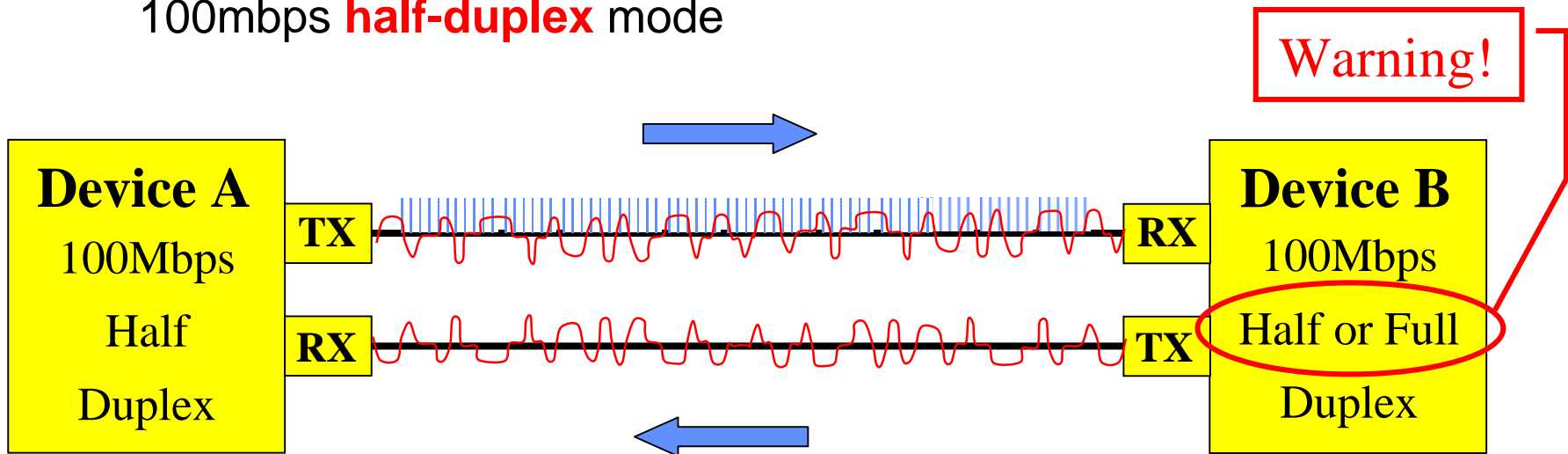
How It Works: an Anege Device with a 10Mbps Legacy Device

- Device A sends FLPs out, while Device B sends out LTPs
- Device A “parallel detects” the LTPs and links with Device B in 10mbps **half-duplex** mode



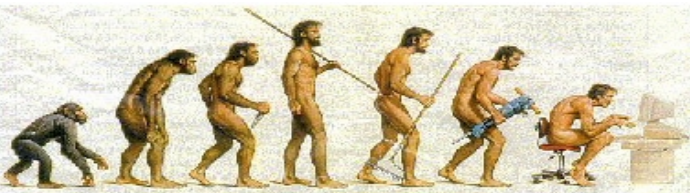
How It Works: an Anege Device with a 100Mbps Legacy Device

- Device A sends FLPs out, while Device B sends out Fast Ethernet IDLE
- Device A “parallel detects” the IDLE and links with Device B in 100mbps **half-duplex** mode



Autoneg Problems

- Three types of problems are sometimes encountered:
 - 1) A Duplex mismatch occurs whereby one side is Half-Duplex, one side is Full-Duplex
 - 2) The wrong speed link is used (10 instead of 100)
 - 3) The devices never link
- Let's look at how these happen...

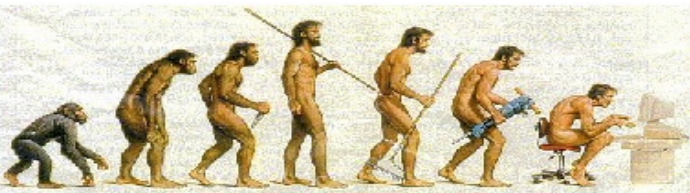
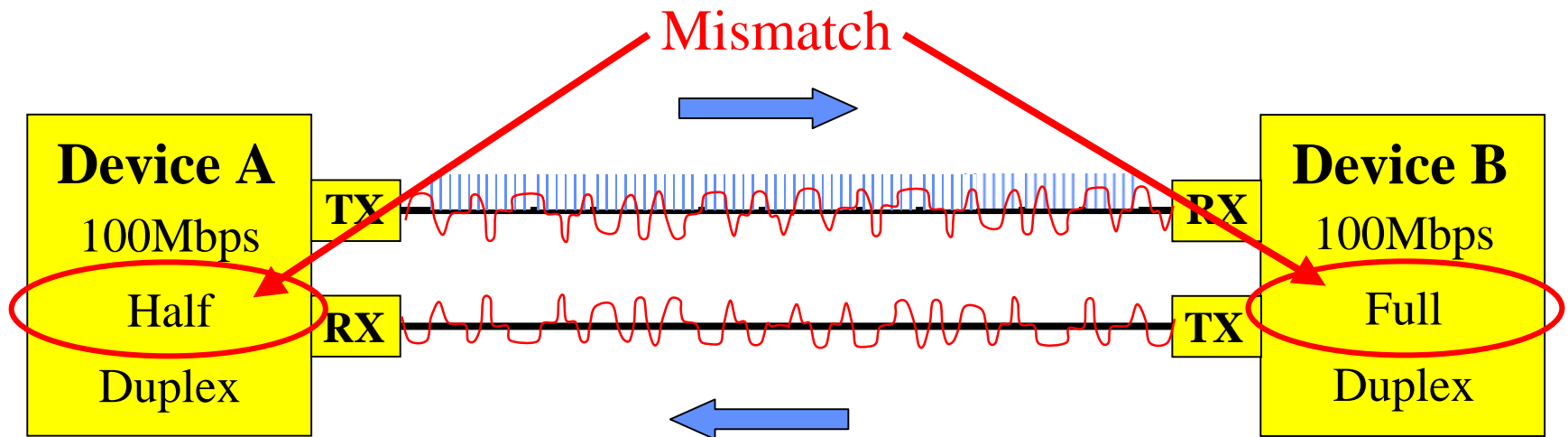


Aneg Error #1: User Misconfiguration

User configures Device B to be 100Mbps Full Duplex, not knowing this disables Autoneg (sending FLPs)

Device A sends FLPs out, while Device B sends out IDLE

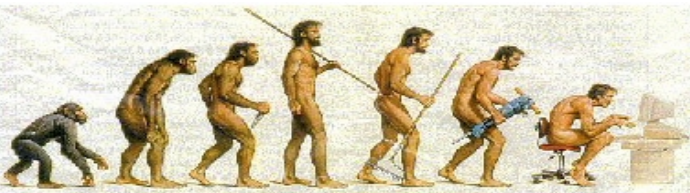
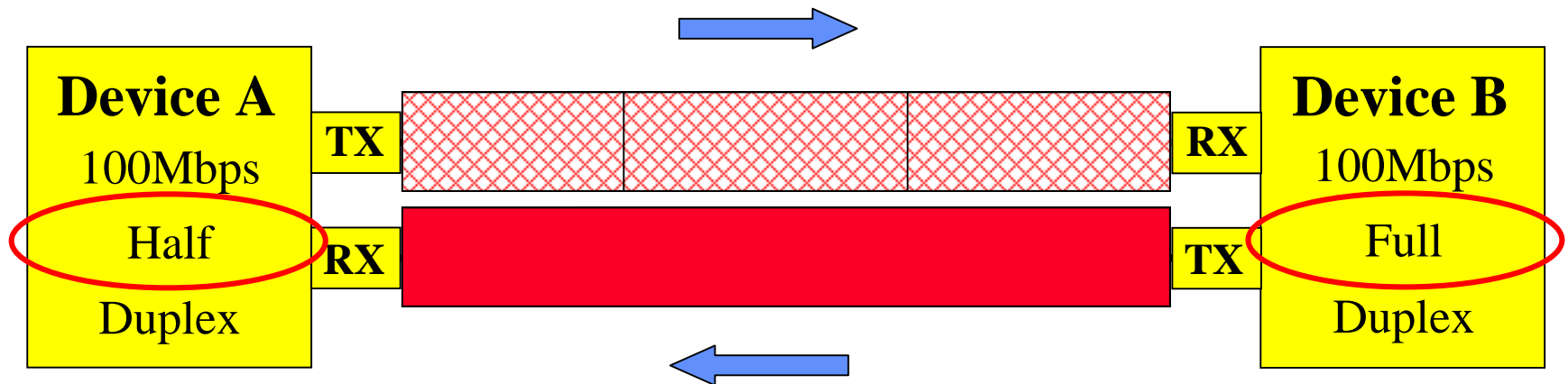
Device A sees IDLE and assumes Device B is 100Mbps Half-Duplex, thus it links in Half-Duplex mode



Duplex Mismatch Problem

If Device A and B send out a frame at the same time, then:

- 1) Device A will believe a collision occurred and corrupt its outgoing frame while discarding Device B's frame, and then attempt to resend its own frame
- 2) Device B will not resend its frame, and see Device A's frame as corrupted

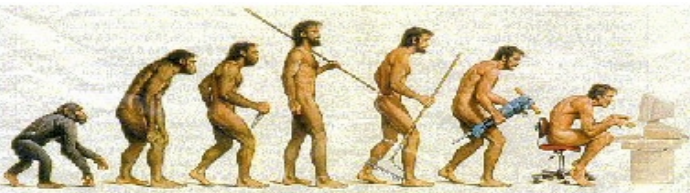
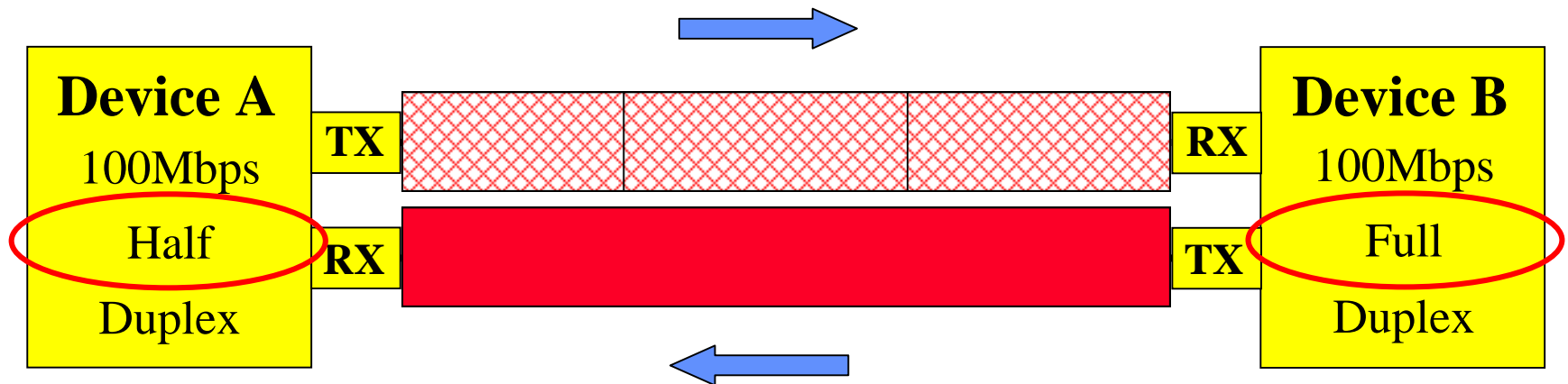


Duplex Mismatch Symptoms

Device A will record many “Late Collisions” in its counters

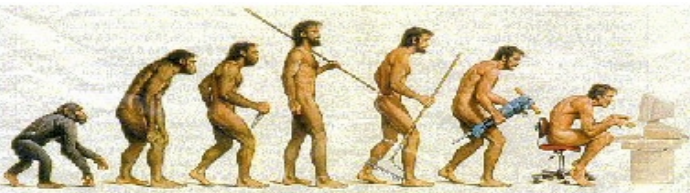
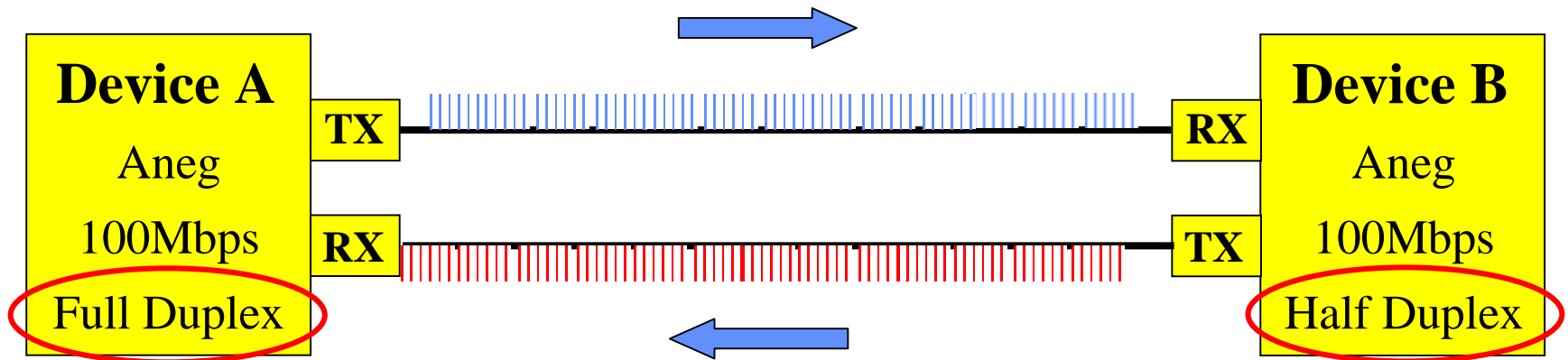
Device B will record many “CRC Errors” in its counters

The connection will appear slow, as the dropped frames aren’t resent until a higher layer times out (usually 1 second later)



Aneg Error #2: User Misconfiguration

- User configures Device A for 100Mbps Full-Duplex and Device B for Half-Duplex (or Device B is only Half-Duplex) while leaving Aneg (sending FLPs) turned on for both
- Each device receives the other's FLPs, but the mismatch never resolves to a link because there is no common option



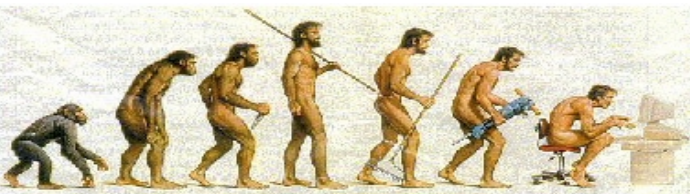
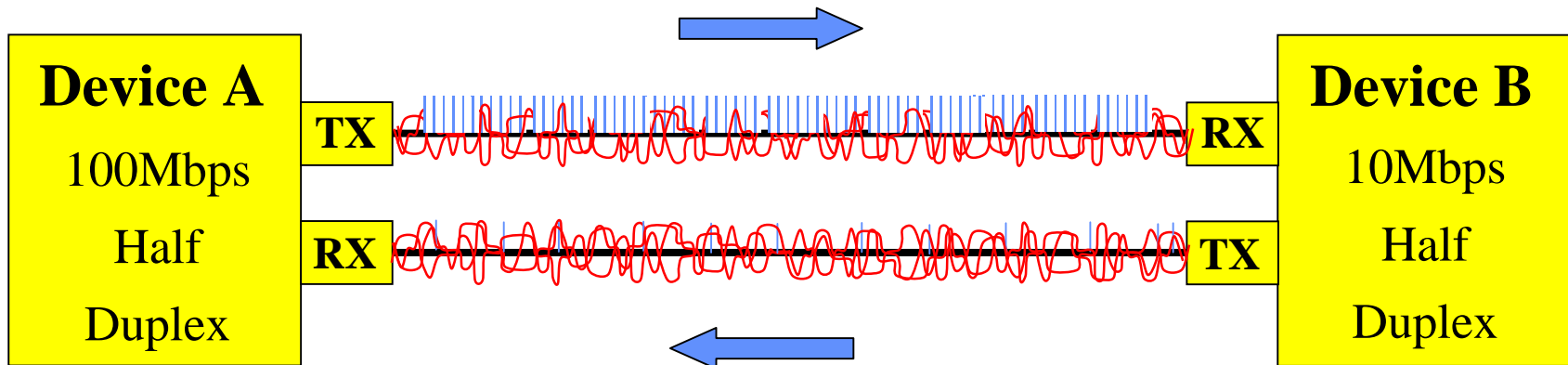
Aneg Error #3: Autosensing Devices

Autosensing devices do not use FLPs

Instead, they send IDLE while watching the RX line for LTPs or IDLE and decide based on that

So Device A will send FLPs, which Device B will interpret as LTPs and switch to transmitting LTPs itself

Meanwhile, Device A may have seen the IDLE from B and “parallel detected” over to sending 100Mbps IDLE, which Device B will now see as 10Mbps junk data (yes, this is legal) - this should fix itself in a few seconds, or never



14
0

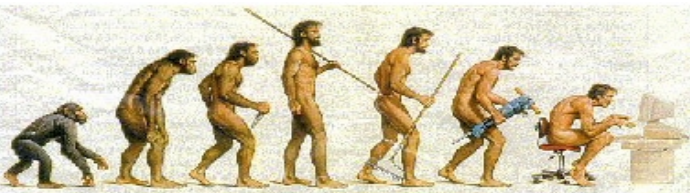
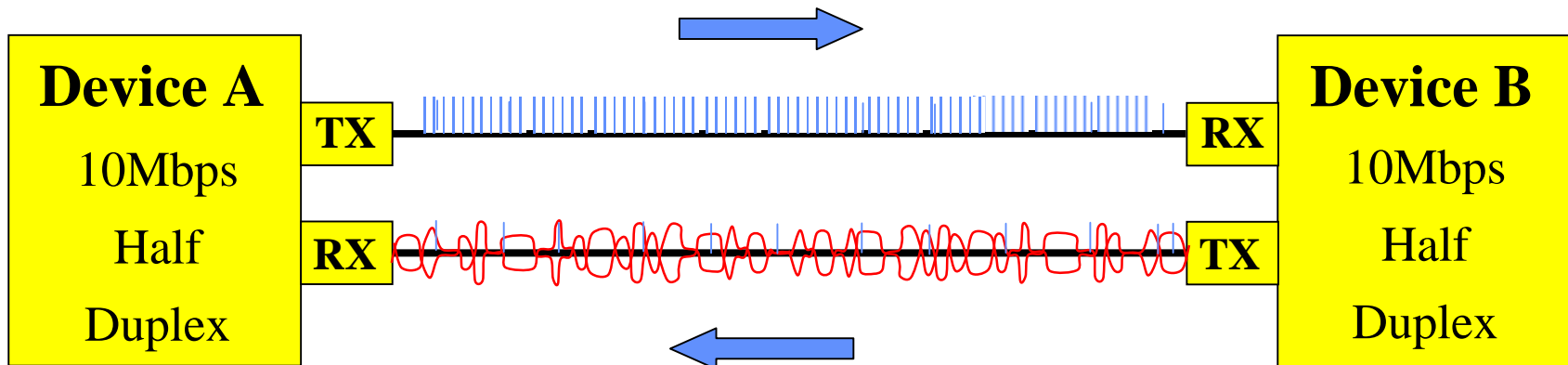
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

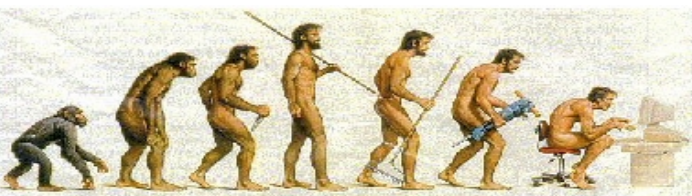
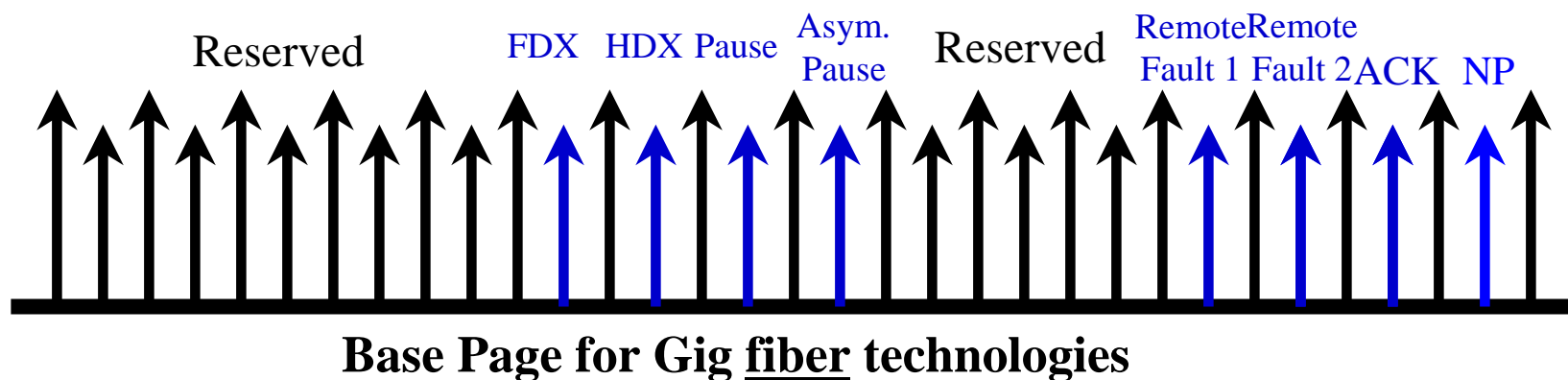
Aneg Error #3: Autosensing Devices

OR, Device A will see the LTPs from Device B and parallel detect to 10mbps, thereby successfully forming a 10mbps link, which is not optimal



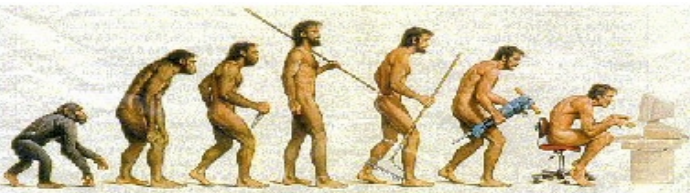
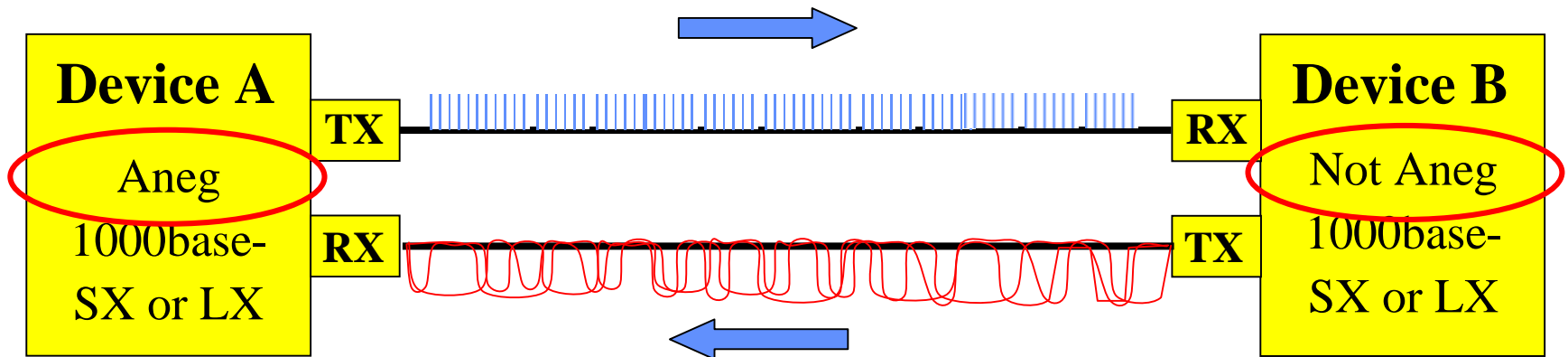
Gigabit Ethernet Autonegotiation

- Gigabit Ethernet uses the same Aneg mechanism as 10/100 copper Ethernet (remember there is no Aneg for 10 or 100 fiber)
- For 1000base-SX and 1000base-LX, it's used to determine duplex and flow-control parameters using the pre-defined fields shown below
- For 1000base-T, the Next Page field is used to indicate there is additional info that must be exchanged using more “pages”



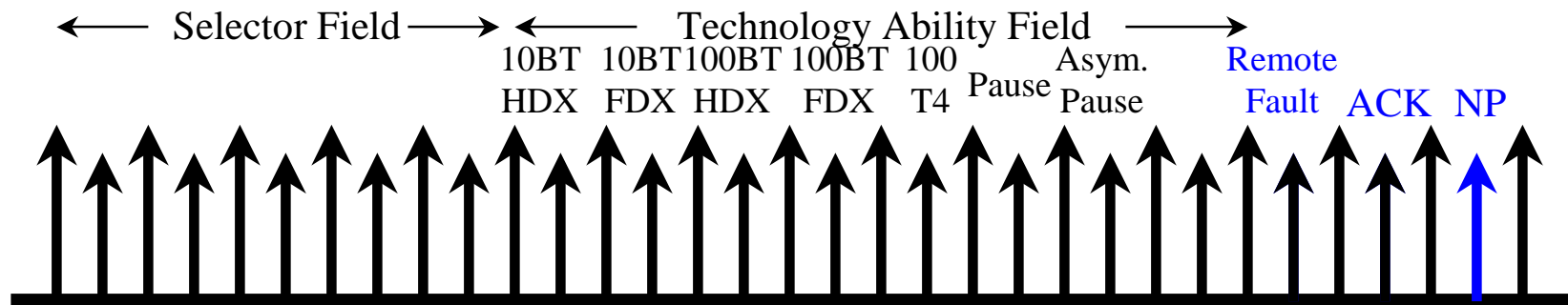
Gig Aneg Error #1: User Misconfiguration

- There are almost no Half-Duplex Gig devices, so we never have the mismatch problem of 10/100 Aneg
- The biggest problem is when users configure one device to “Autoneg Enabled” and the other device to “Autoneg Disabled”
- Such misconfigured devices will not link, or only one side will think it’s linked

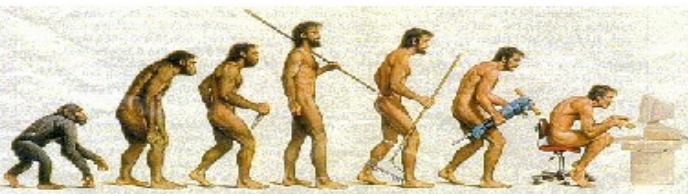


Gigabit Copper Autonegotiation

- For 1000base-T, the Next Page field is used to indicate there is additional info that must be exchanged using more “pages”
- These pages contain the same type of info regarding duplex and speeds, but they also contain a couple new things:
 - a field to indicate whether the device is a single-port or multi-port device
 - a field for whether the master/slave determination is manually configured
 - a field for whether this device is the master or slave
 - a seed value



Base Page for Gig copper technologies



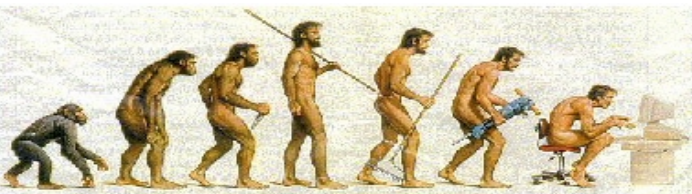
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

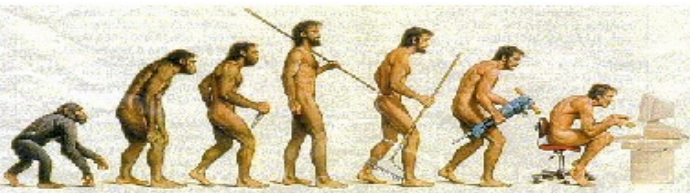
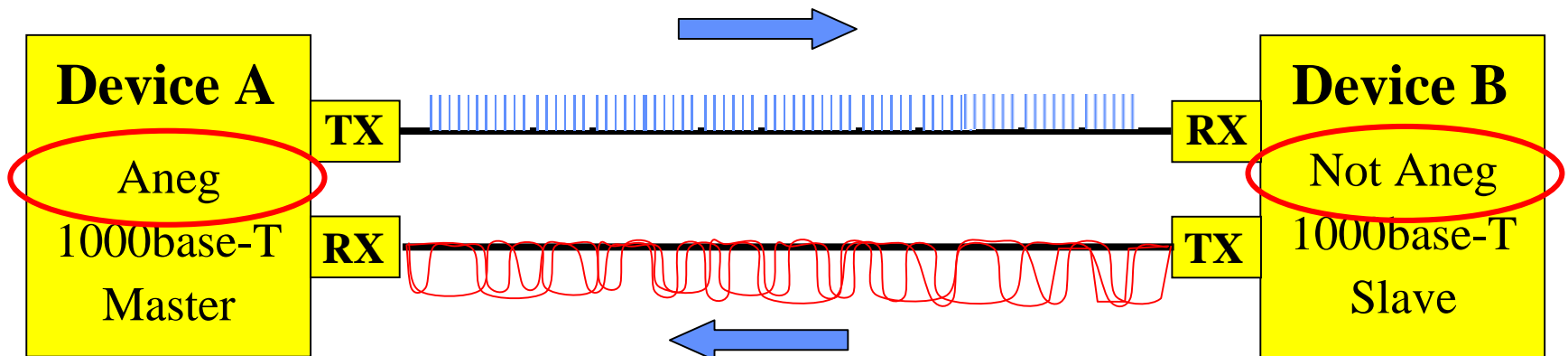
Gigabit Copper Autoneg (cont.)

- The purpose for the new fields is to resolve which side is the master and which the slave for a training sequence to adjust to the cable characteristics (like a DSL modem does)
- If the user does nothing, then it will automatically select a master based on:
 - if one side is multi-port, it is the master
 - if both sides are or both sides are not multi-port, then the seed is used to randomly choose
- Or the user can manually configure one to be the master and the other a slave



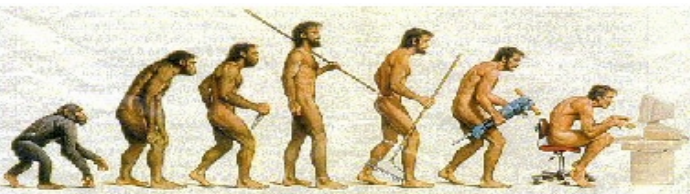
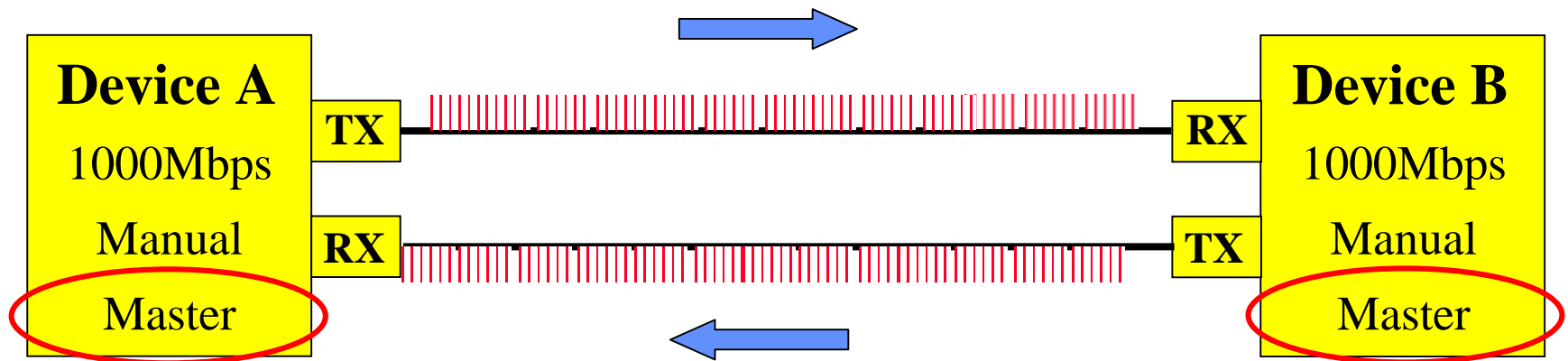
Gig Copper Aneg Error #1: User Misconfiguration

- There are almost no Half-Duplex Gig devices, so we never have the mismatch problem of 10/100 Aneg
- The biggest problem is when users configure one device to “Autoneg Enabled” and the other device to “Autoneg Disabled”
- Such misconfigured devices will not link, or only one side will think it’s linked

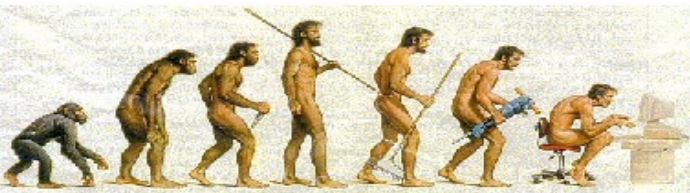


Gig Copper Aneg Error #2: User Misconfiguration

- A user configures both devices to be master or both to be slave
- Such misconfigured devices will never link - even if they're also both autonegotiating, or both manual
- The problem is manually configuring master/slave overrides automatically selecting which side is which



Power over Ethernet - 802.3af



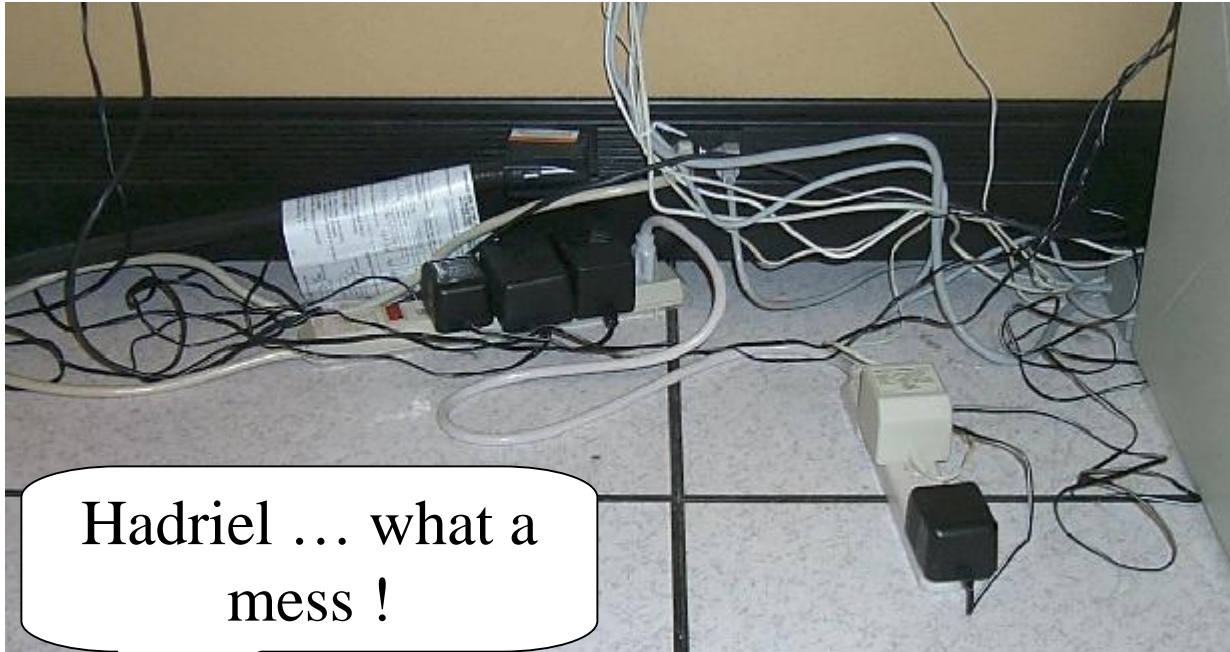
14
8

The Ethernet Evolution

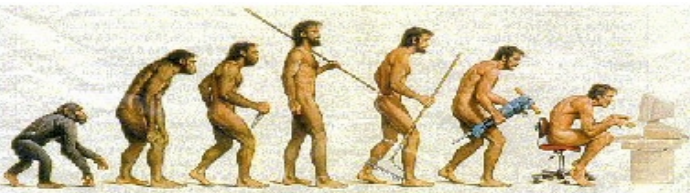
10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Etherpower = No More Wall Warts !!



Hadriel ... what a mess !



14
9

The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Power over Ethernet: the concept



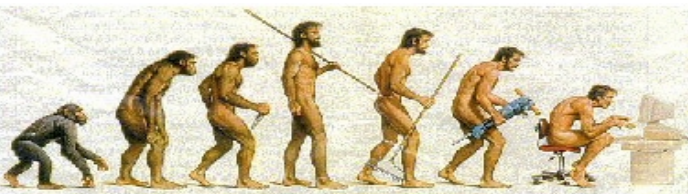
*48VDC Modular
Rectifier and
Battery Tray*



*Ethernet Switches
or
'mid-span insertion'*



Data Appliances

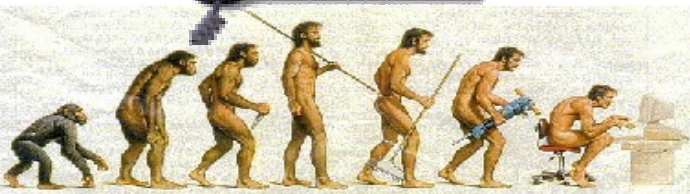
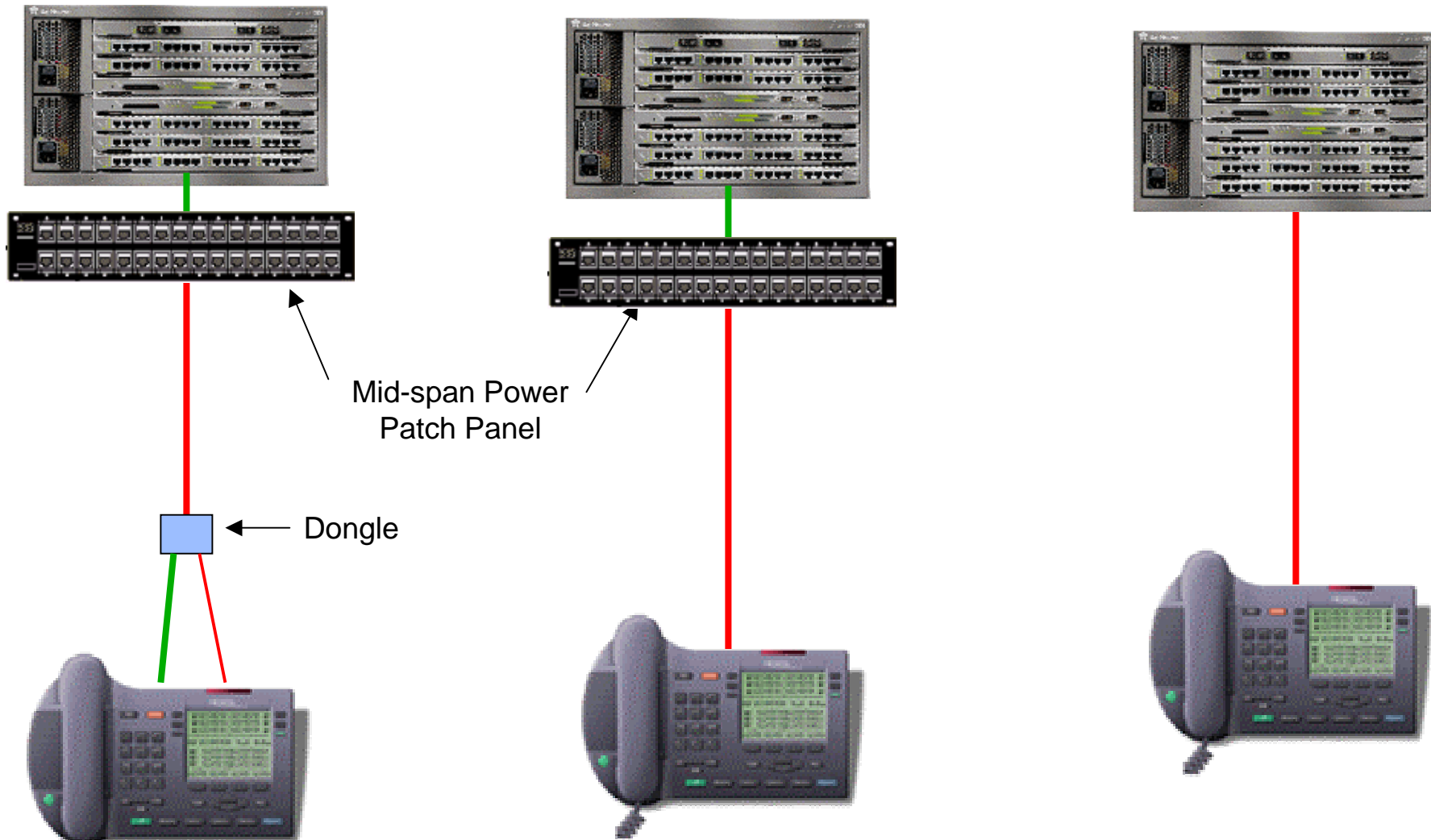


The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Power over Ethernet: the details



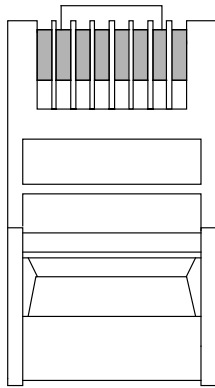
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

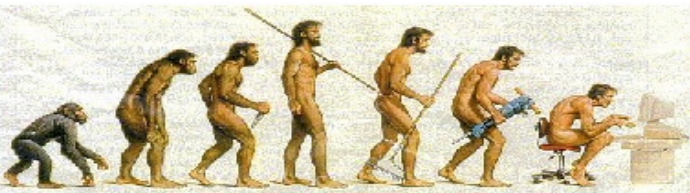
UTP Copper

- looking down at the top of the RJ-45 (locking tab facing down)



1	2	3	4	5	6	7	8
Solid White, Orange Stripe	Solid Orange, White Stripe	Solid White, Green Stripe	Solid Blue, White Stripe	Solid White, Blue Stripe	Solid Green, White Stripe	Solid White, Brown Stripe	Solid Brown, White Stripe

- 10base-T and 100base-TX/T2 only use wires 1, 2, 3, and 6 - wires 4, 5, 7, and 8 are unused
- 100base-T4, and 1000base-T use all 8 wires



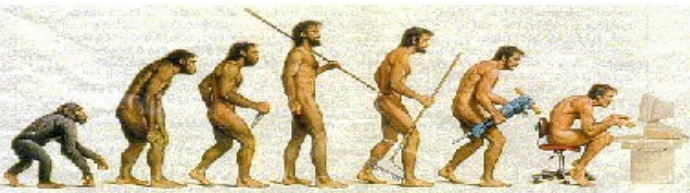
Power Over What?

Unused pairs (4/5, 7/8)

- Safer
- Cheaper for phone side
- Allows for mid-span insertion
- Allows for cheap “dongles” and non-data equipment

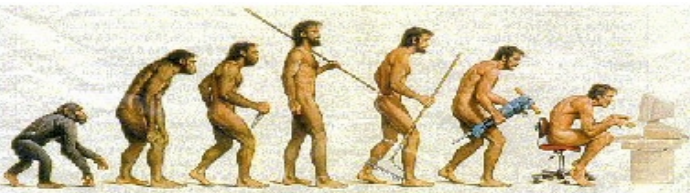
Signal pairs (1/2, 3/6)

- Doesn't need all 4 pairs
- *May* work with 1000base-T
- Cheaper for switch side



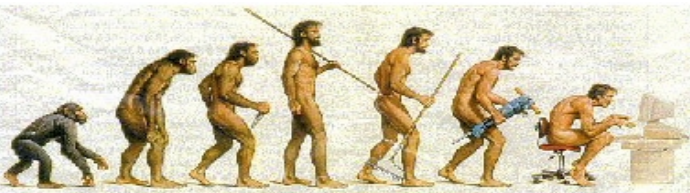
Power Method Options

- Use only unused pairs
- Use only data pairs
- Use both: the source can feed power on one or the other, but the sink (device to be powered) will draw it from either
 - this makes the receiver (the phone side) more expensive
 - this allows for both a cheap powered patch panel and cheap powered switch method
 - this does not allow both data and unused pairs to send power simultaneously (not safe)



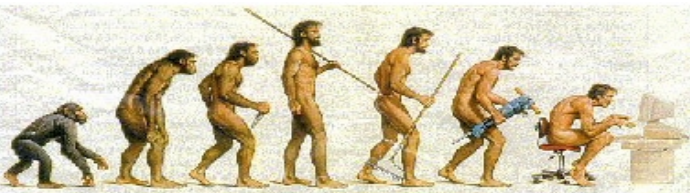
Power Discovery

- Method for power source to determine if connected device needs power
- Needs to be fool-proof (doesn't fry old stuff)
- Needs to be cheap
- Needs to work before the other side is powered up (obviously)
- Needs to be safe regardless of cable-type connected: could be regular, crossover, flipped, wrong, broken

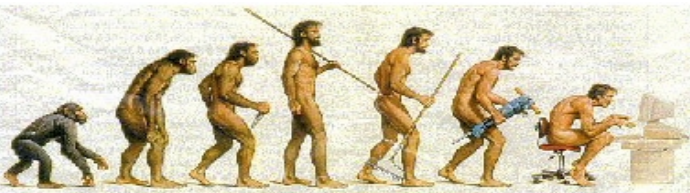


TBD

- Voltage - probably 48v nominal DC (definitely DC)
- Amperage - prob. 300mA
- Watts - prob. about 10-12 watts (this isn't much)
- Which pairs, or both
- Which discovery method(s)
- Basically Everything is To Be Decided



10 Gigabit Ethernet - 802.3ae



15
7

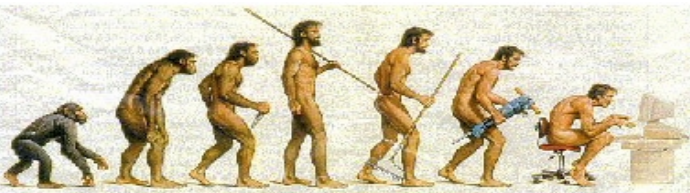
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

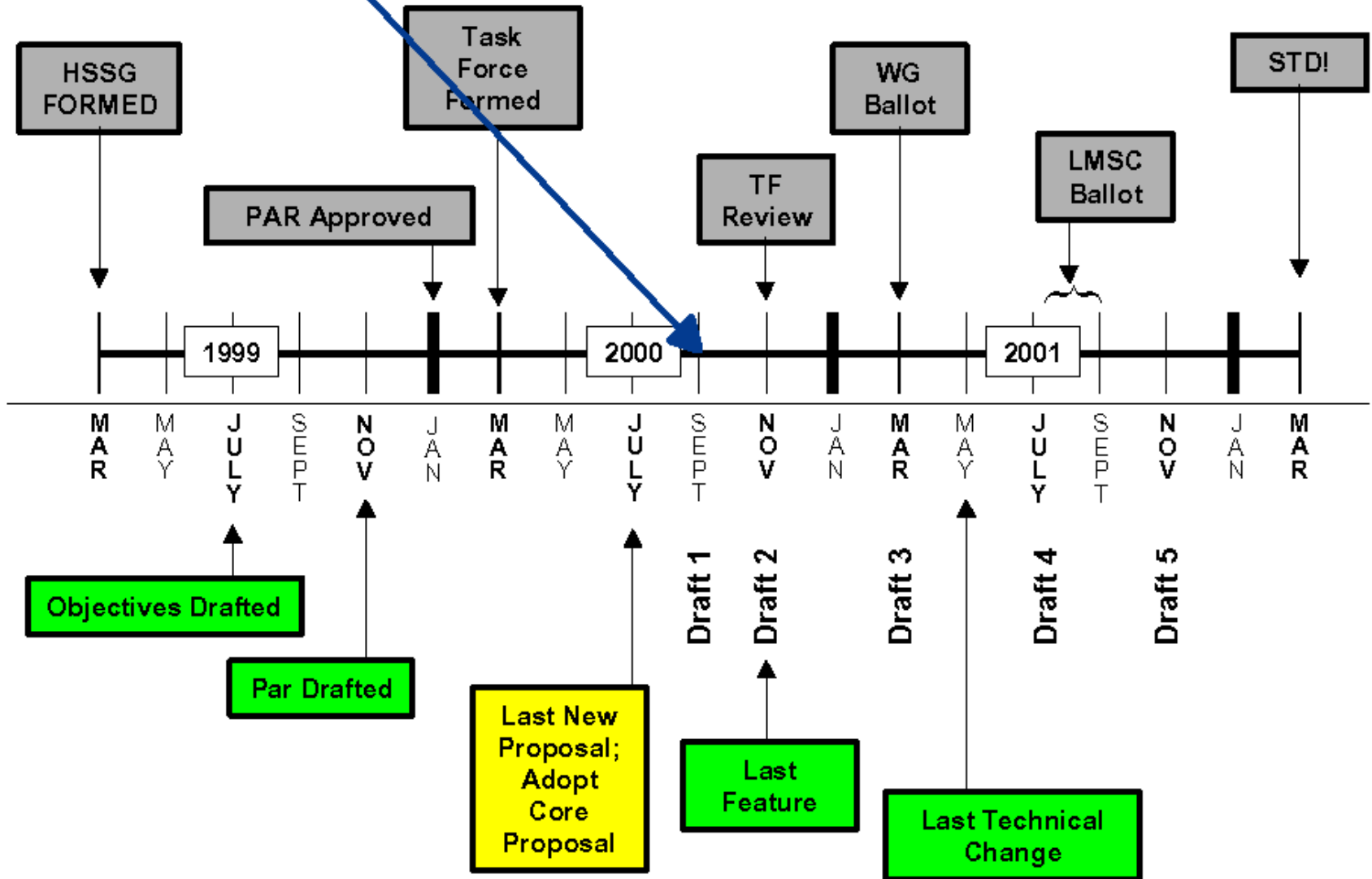
IEEE 802.3ae Objectives

- **Support full-duplex operation only.**
- **Provide Physical Layer specifications which support link distances of:**
 - – At least 300 m over installed MMF
 - – At least 65 m over MMF
 - – At least 2 km over SMF
 - – At least 10 km over SMF
 - – At least 40 km over SMF
- **Define two families of PHYs**
 - – A LAN PHY, operating at a data rate of 10.000 Gb/s
 - – A WAN PHY, operating at a data rate compatible with the payload rate of OC-192c/SDH VC-4-64c



Long Term Schedule

You are here



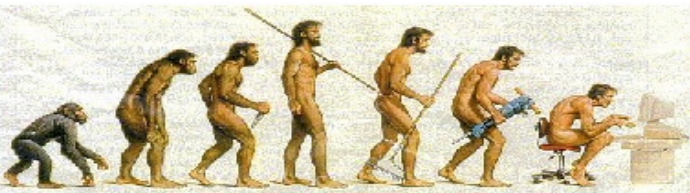
The PMDs

- Set of 4 PMDs (Physical Media Dependent) to optimize balance between distance objectives, cost

Application

Optimal Solution

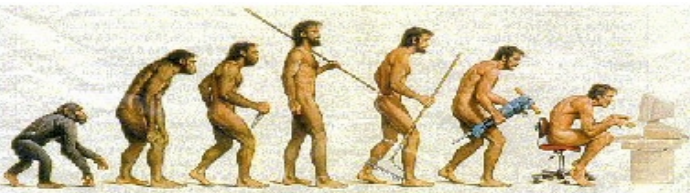
Longest Distance (40+ km)	1550 Serial
Med. reach, lower cost, transponder compat.	1310 Serial
Max reuse of installed MM / SM (Building LAN)	1310 WWDM
Low cost on MM (Equipment Room)	850 Serial



PMD Distance

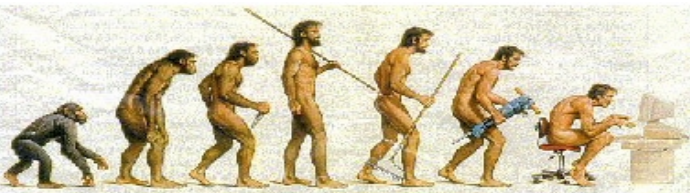
Fiber Type per 11801 (Bandwidth @ 850nm//1310nm)	850 Serial	1310 WWDWM	1310 Serial	1550 Serial
Legacy 62.5 MMF (160-200//500)	1-25 m	*1-300 m	NA	NA
Legacy 50 MMF (400-500//500)	1-75 m	*1-300 m	NA	NA
SMF	NA	1-10 km	1-10km	1-40km+

* Offset Launch Patch Cord required for distances $> 100\text{m}$



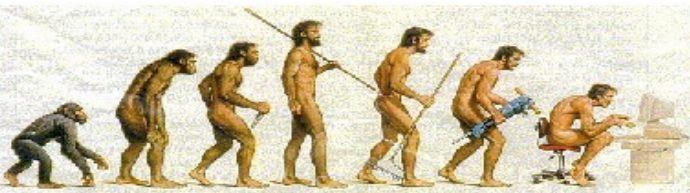
PMD Names

- Wavelength: S=850nm L=1310nm E=1550nm
- PMD Type:
 - X=WDM LAN(Wave Division Multiplexing – 4 wavelengths on 1 fiber)
 - R=Serial LAN using 64B/66B coding (LAN Application)
 - W=Serial WAN – SONET OC-192c compatible speed/framing
- 10GBASE-LX
- 10GBASE-SR / -LR / -ER
- 10GBASE-LW / -EW



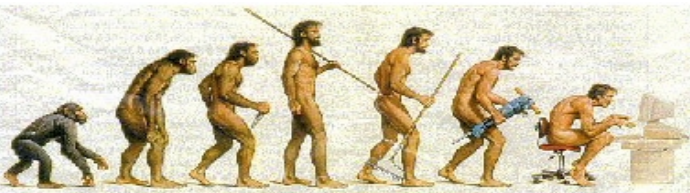
LAN PHY

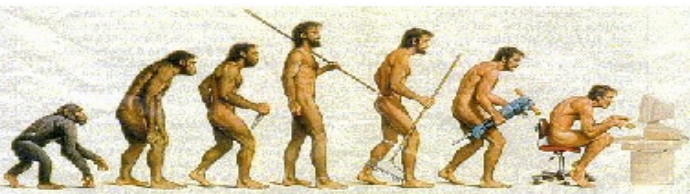
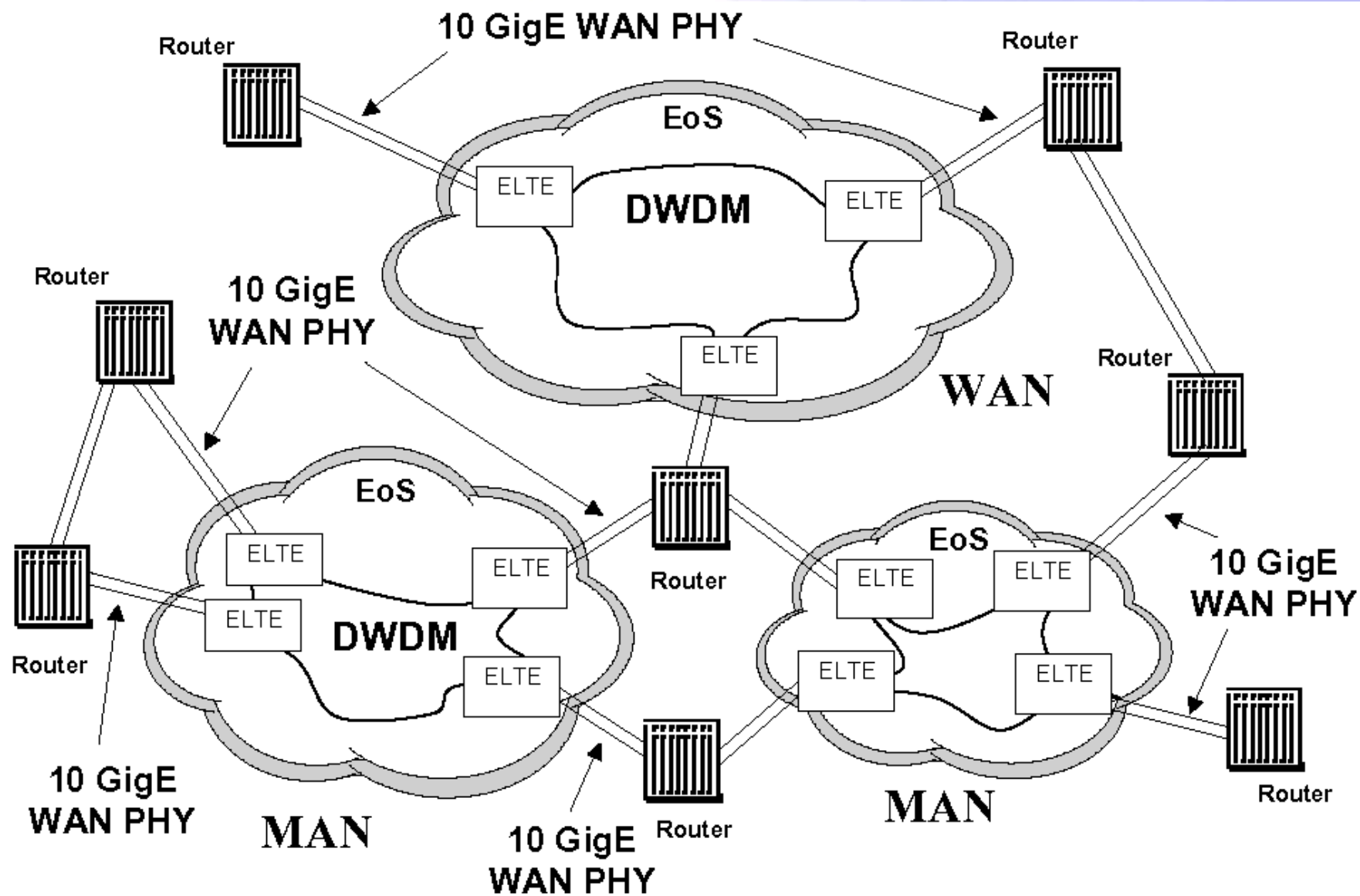
- Simple Ol' Ethernet
- Transmits/Receives MAC frames at 10.000Gbps
- Allows easy/simple speed scaling/aggregating of 10 1-Gbps links
- Supports Link Aggregation
- Can drive 2m to 40km depending on PMD
- Useful for:
 - Campus backbones (connect Gig E switches together)
 - Dark Fiber runs
 - Computer Rooms
 - SANs, etc...



WAN PHY

- Transmit/Receive MAC Frames at 9.29419 Gig
Why?
- SONET/SDH OC-192c Data rate compatible
- Take coded (via 64b/66b) MAC frames and place in the payload section of SONET frame and transmit onto SONET at OC-192c speeds (9.95328)
- For what purpose?
 - Make use of the existing SONET Photonic Network
 - Predominate optical infrastructure in North America, Europe, and China
 - Avoid use of costly features of SONET – Optics, 1ppm Stratum Clock, and many management features



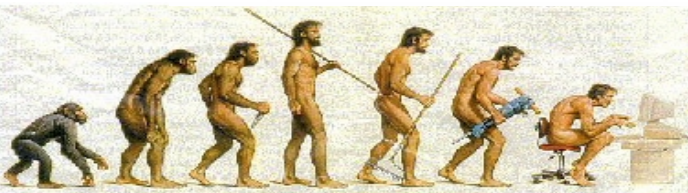
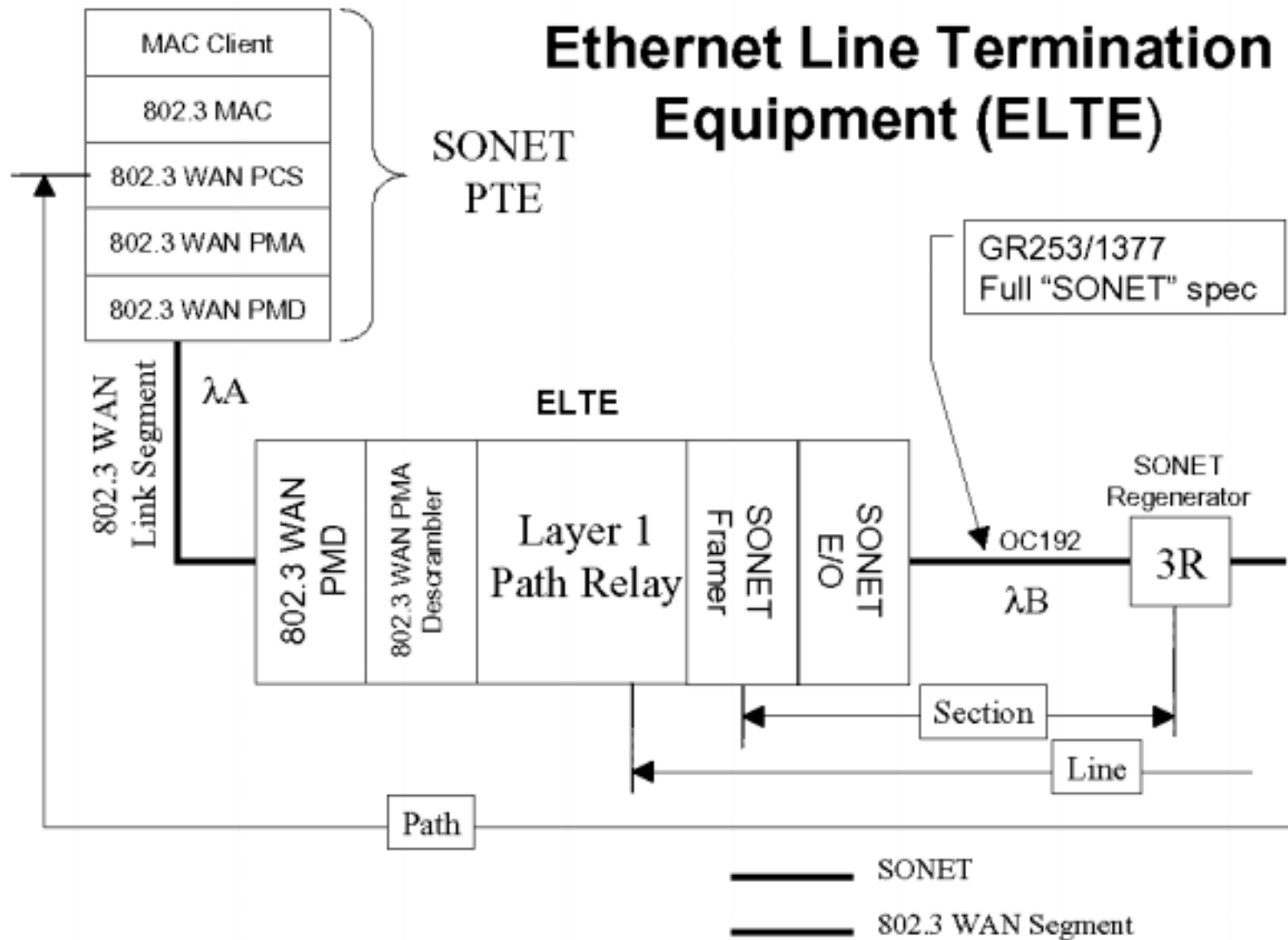


The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Ethernet Line Termination Equipment (ELTE)

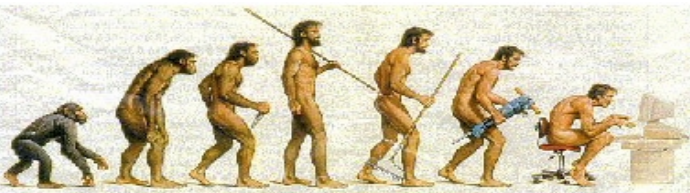
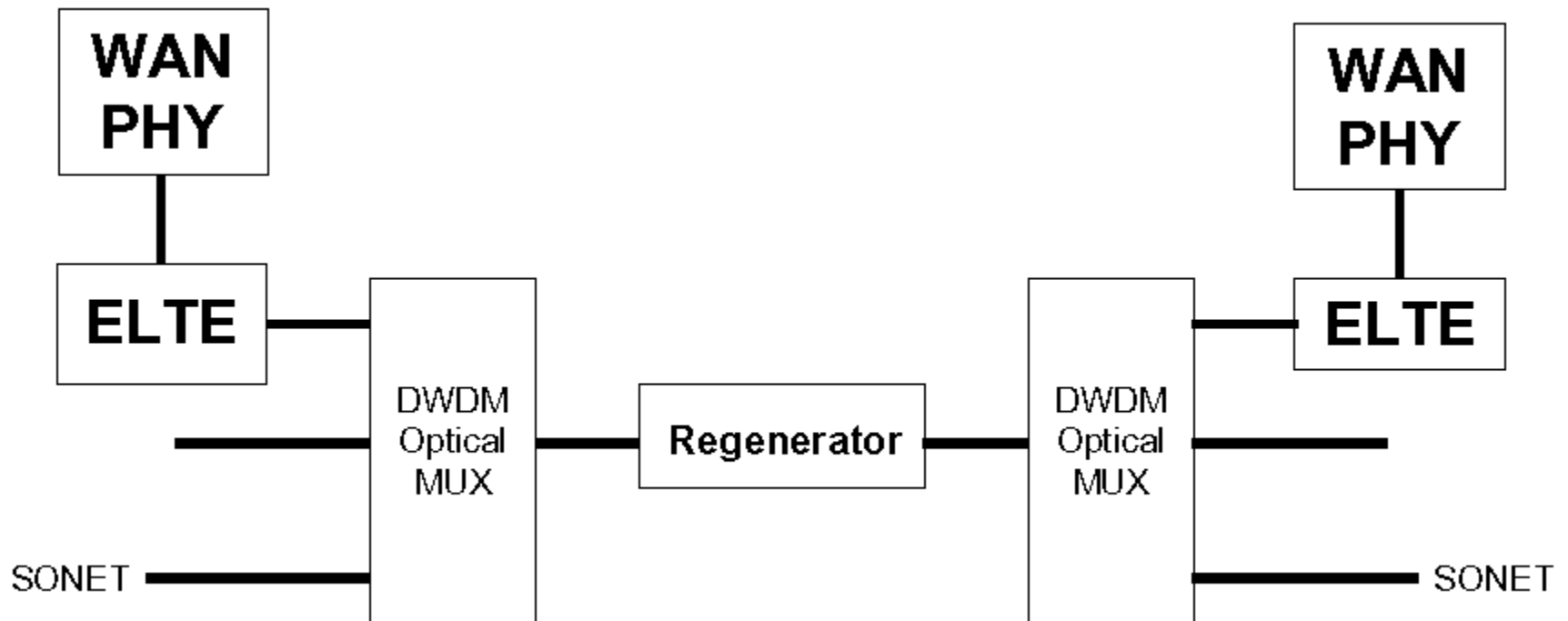


The Ethernet Evolution

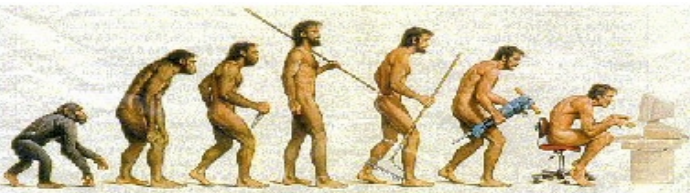
10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Use in today's DWDM



Switched Networks



16
8

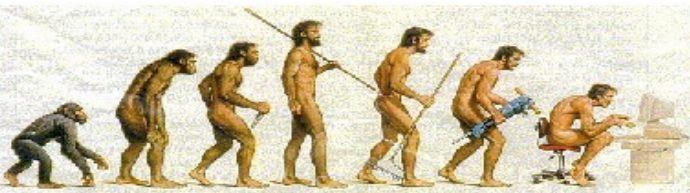
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

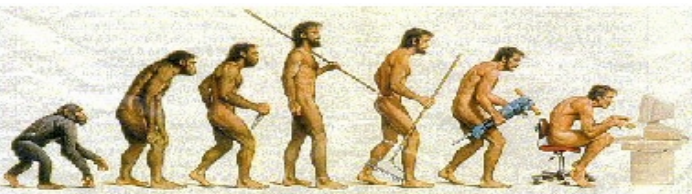
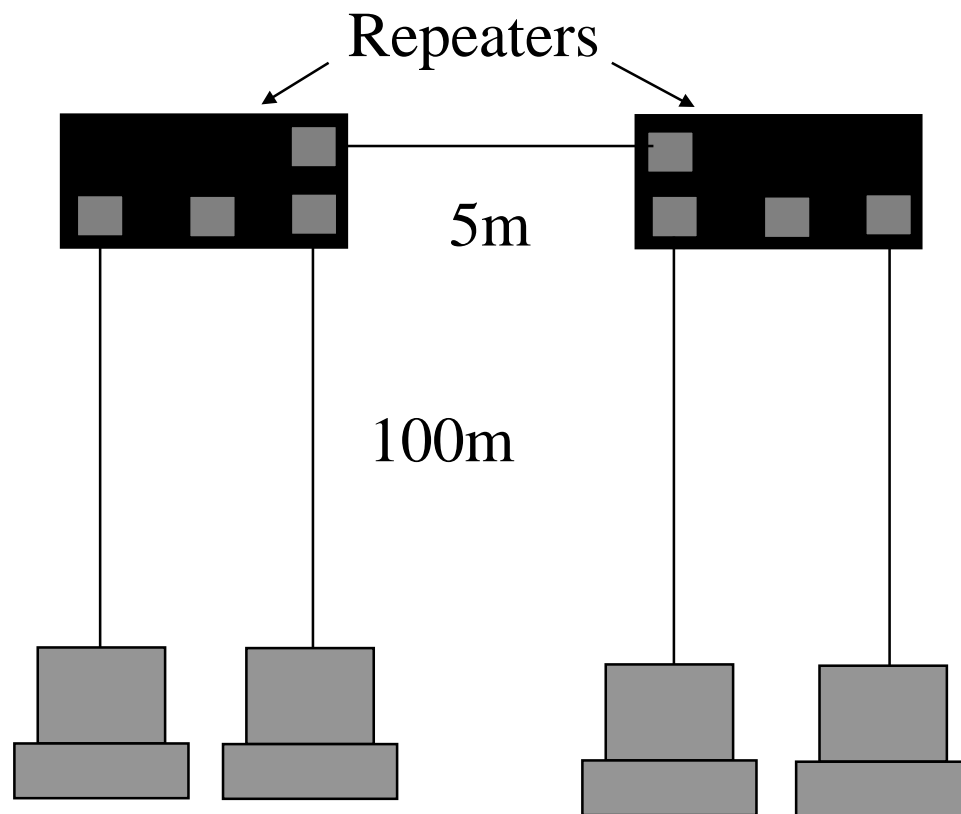
Topics

- Quick Background
- Spanning Tree
- VLANs
- 802.1p/QoS
- L3 Switching
- Link Aggregation
- Multiple Spanning Trees
- Rapid Reconfiguration



Shared Medium (Repeated Network)

- All machines “share” the network
- Only one machine can talk at any one time
- Distance limitations
 - At most 205m for Fast Ethernet
- Total throughput limited
- Single collision domain



End Stations
The Ethernet Evolution

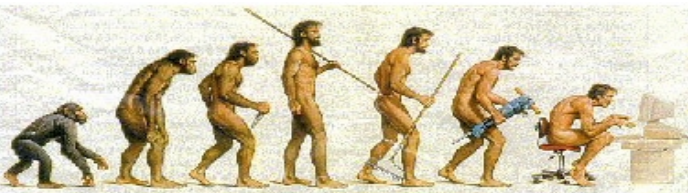
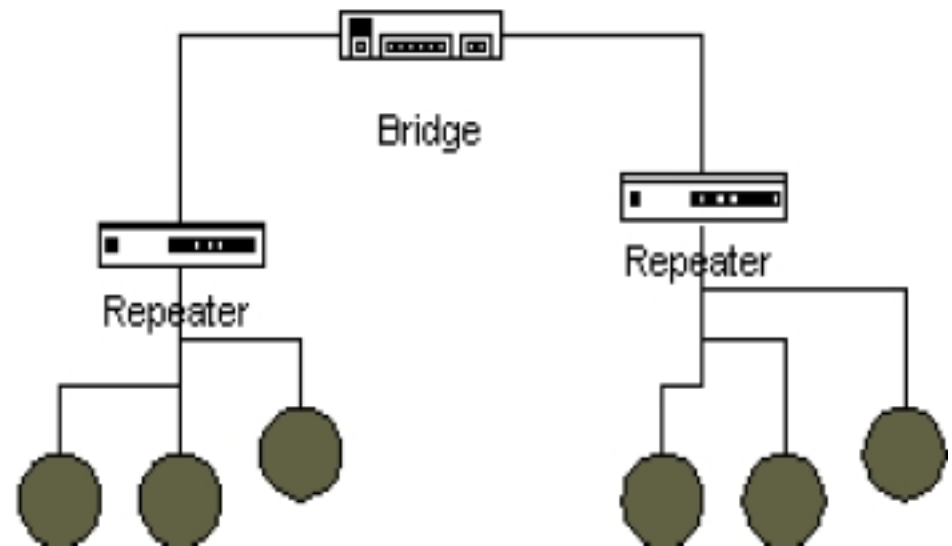
10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Bridging Review

- Connects Separate shared Networks
- Frame Translation/ Encapsulation (Token Ring to Ethernet)
- Reduces Unicast Traffic
- Switches: Allow for multiple conversations

One Broadcast Domain
Two Shared Mediums

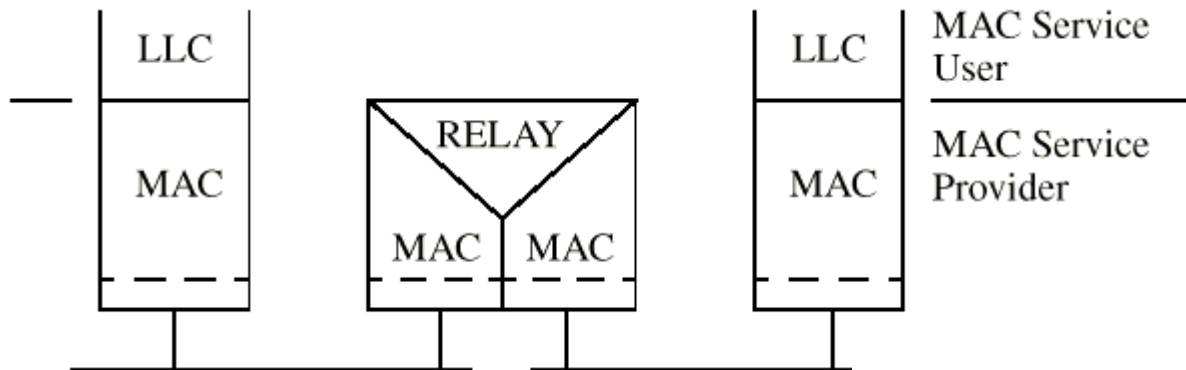


The Ethernet Evolution

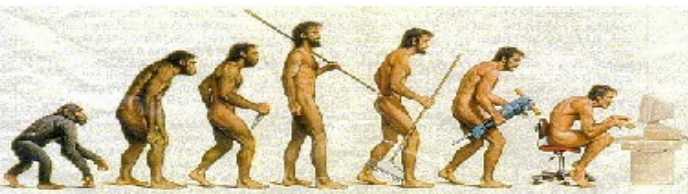
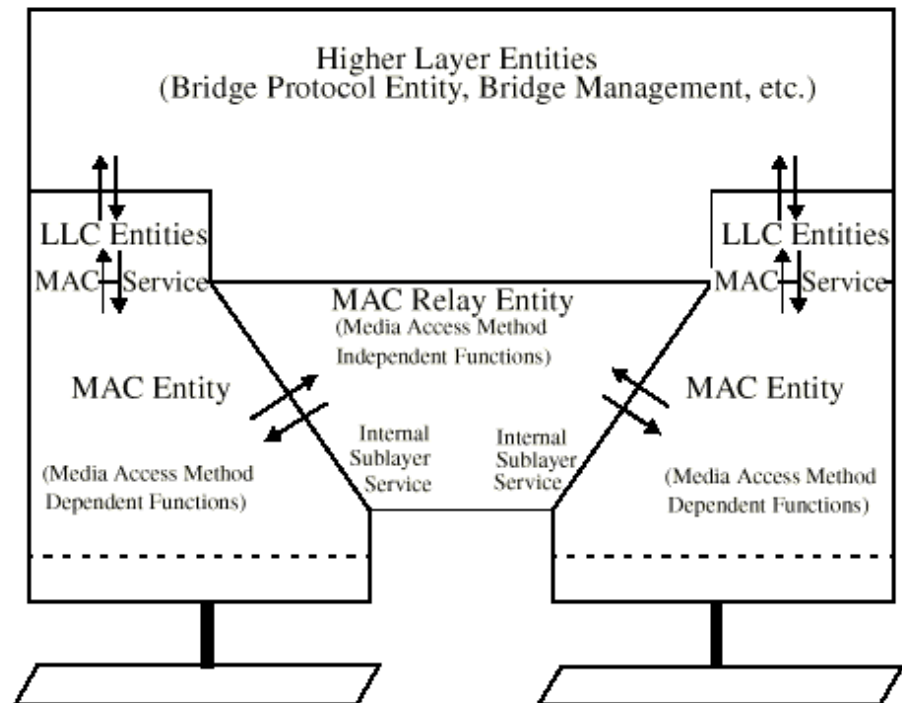
10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Bridging Background



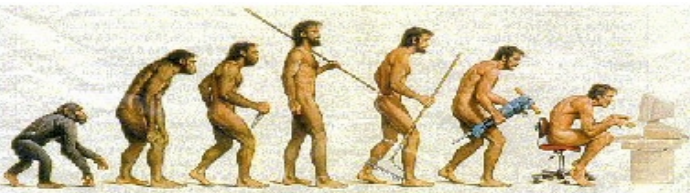
- Bridges work at layer 2 of the OSI Model
- Their primary function is to relay frames



Bridge Tables

- One table lists MAC addresses, which port they're on, and if they're active or disabled

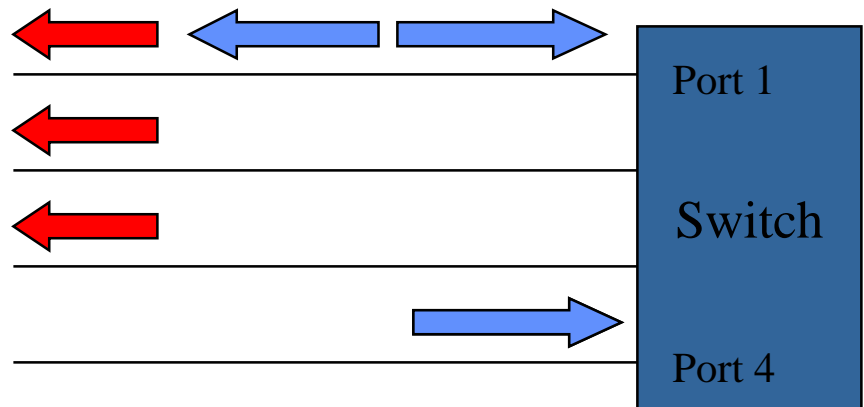
Entry	MAC Addr	Port	active
1	0800900A2580	1	yes
2	002034987AB1	1	yes
3	0500A1987C00	2	yes
4	00503222A001	2	yes
5			
6			
7			
8			
9			
10			
11			
12			



Learning of addresses

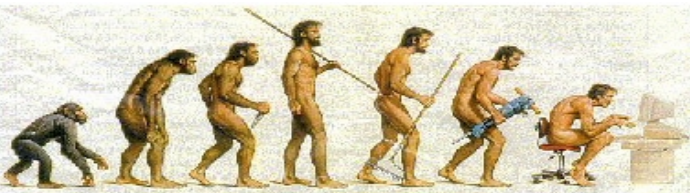
The Filtering Database learns a station's location from the source address on an incoming frame.

This source address is "learned" by the filtering database. All future frames destined for this MAC address 002222333344 are will be forwarded ONLY out of this Port.



Frame with destination address 002222333344 is received on Port 4.

Since this is not learned, it is FLOODED out all of the other ports.



The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

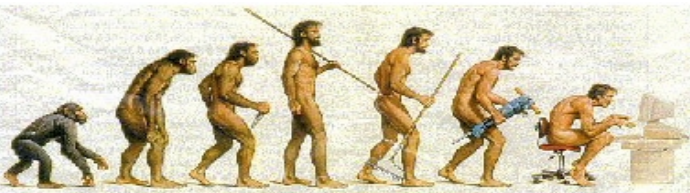
The Learning Bridge

That was a bit fast and complex. Let's review.

Every bridge has a table called a Filtering Database. Entries in this table are updated upon receipt of frames, the source addresses and the ports they arrive on are learned.

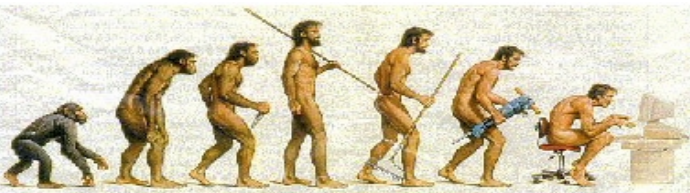
Once a MAC address is associated with a port, frames containing that destination address are only forwarded out of that port.

In the real switches these tables vary in size, most have the capability of holding several thousand MAC addresses. I've seen one that has the capacity for more than 150,000 addresses.



Spanning Tree

Part of 802.1D



802.1D - Spanning Tree

- Configures arbitrary physical topology into one loop-free spanning tree
- provides fault-tolerance by using redundant links as backups
- configures deterministically/reproducibly
- low overhead/bandwidth
- auto-configures

International Standard ISO/IEC 10038 : 1993
ANSI/IEEE Std 802.1D, 1993 edition

(This edition contains ANSI/IEEE Std 802.1D-1990,
ANSI/IEEE Std 802.1i-1992, and IEEE Std 802.5m-1993)

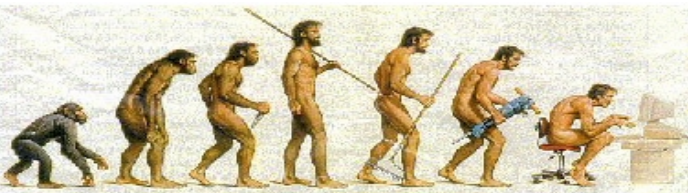
Information technology—
Telecommunications and information exchange
between systems—Local area networks—
Media access control (MAC) bridges

Sponsor

Technical Committee on Computer Communications
of the
IEEE Computer Society

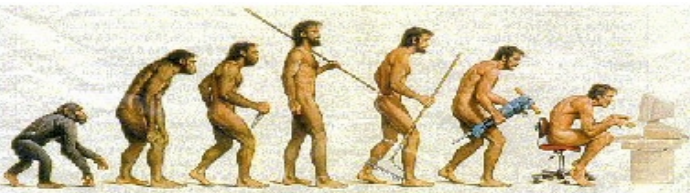
Abstract: An architecture for the interconnection of IEEE 802 Local Area Networks (LANs) below the level of the MAC Service, which is transparent to logical link control (LLC) and higher layer protocols, is defined. Transparent Bridging between Fibre Distributed Data Interface (FDDI) LANs and between FDDI LANs and IEEE 802 LANs is included. The operation and management of the connecting Bridges is specified. A Spanning Tree Algorithm and Protocol ensures a loop-free topology and provides redundancy. The Bridging method is not particular to any MAC Type; criteria for additional MAC-specific Bridging methods are defined. Source-Routing Transparent (SRT) Bridges are defined in an annex, and the protocols for the operation of source routing in an SRT Bridge are specified.

Keywords: data processing, information interchange, local area networks, metropolitan area networks, fibre distributed data interface (FDDI), mode of data transmission, network interconnection, models, source routing, Source-Routing Transparent (SRT) Bridge



Auto-Configuration

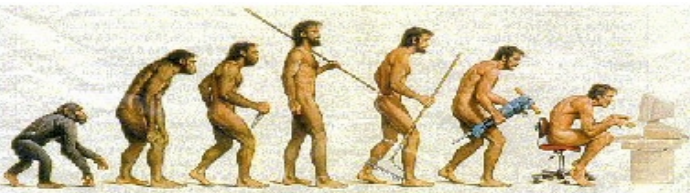
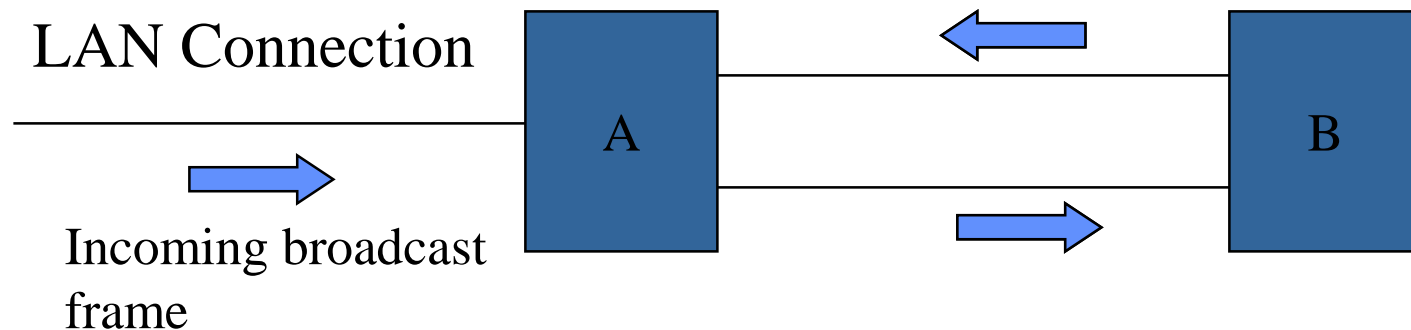
- Default values allow out-of-box configuration without user interaction
- If the user wants, though, they can control the relative priority of bridges and ports to be used
- The user can also control the “path cost” of each port, so a least-cost tree can be built according to the user’s specifications



Spanning Tree

Why a tree?

If you have 2 switches that are connected in parallel, it could create a loop.

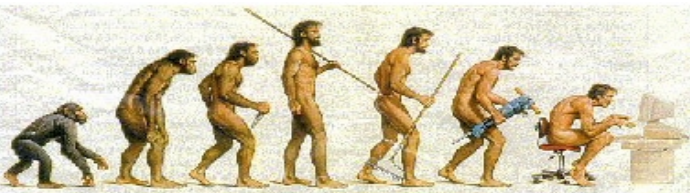


More Reasons

Spanning Tree Disables one of these connections.

It also keeps track of each of these connections. If the active connection becomes disconnected, it will activate a disabled one to take over and make a new tree.

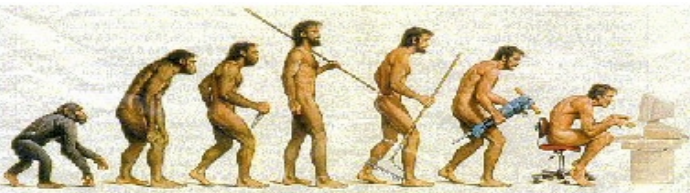
How does it do this?



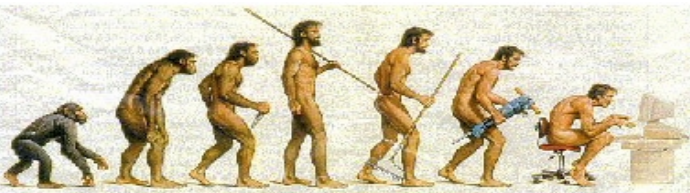
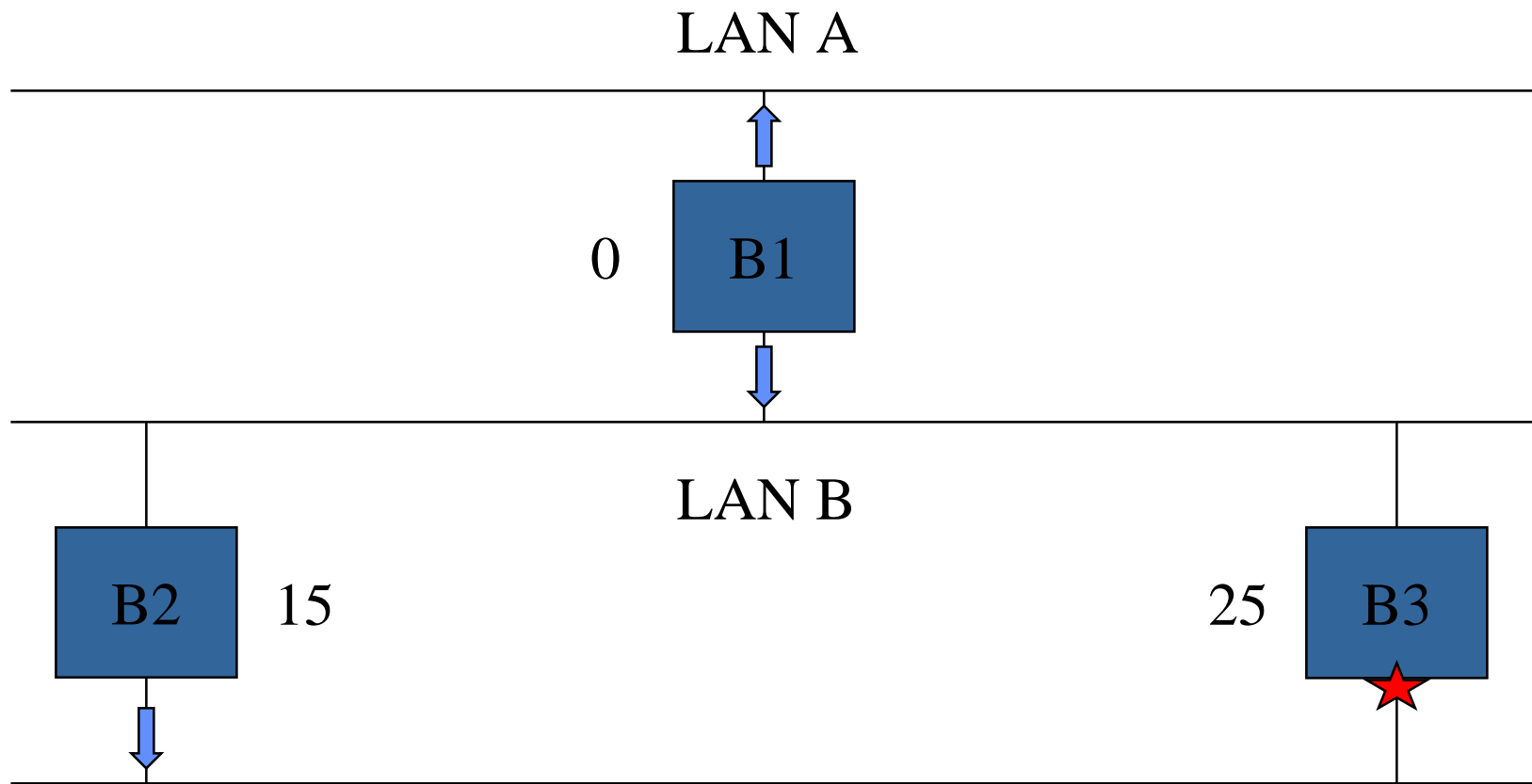
Initial Bridge Parameters:

Bridge	Priority	Path Cost
B1	1	20
B2	2	15
B3	2	25

- All Ports on each bridge have the same Path Cost in this example.
- The Max Age, Hello Time, and Forward Delay parameters are left at their default values of 20.0, 2.0, and 15.0 respectively.



Initial Bridged LAN Topology



LAN C

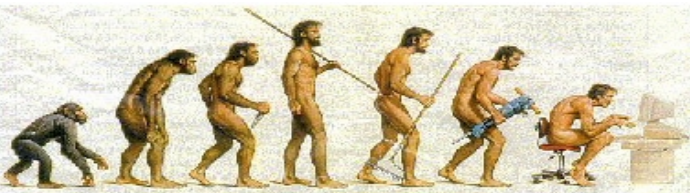
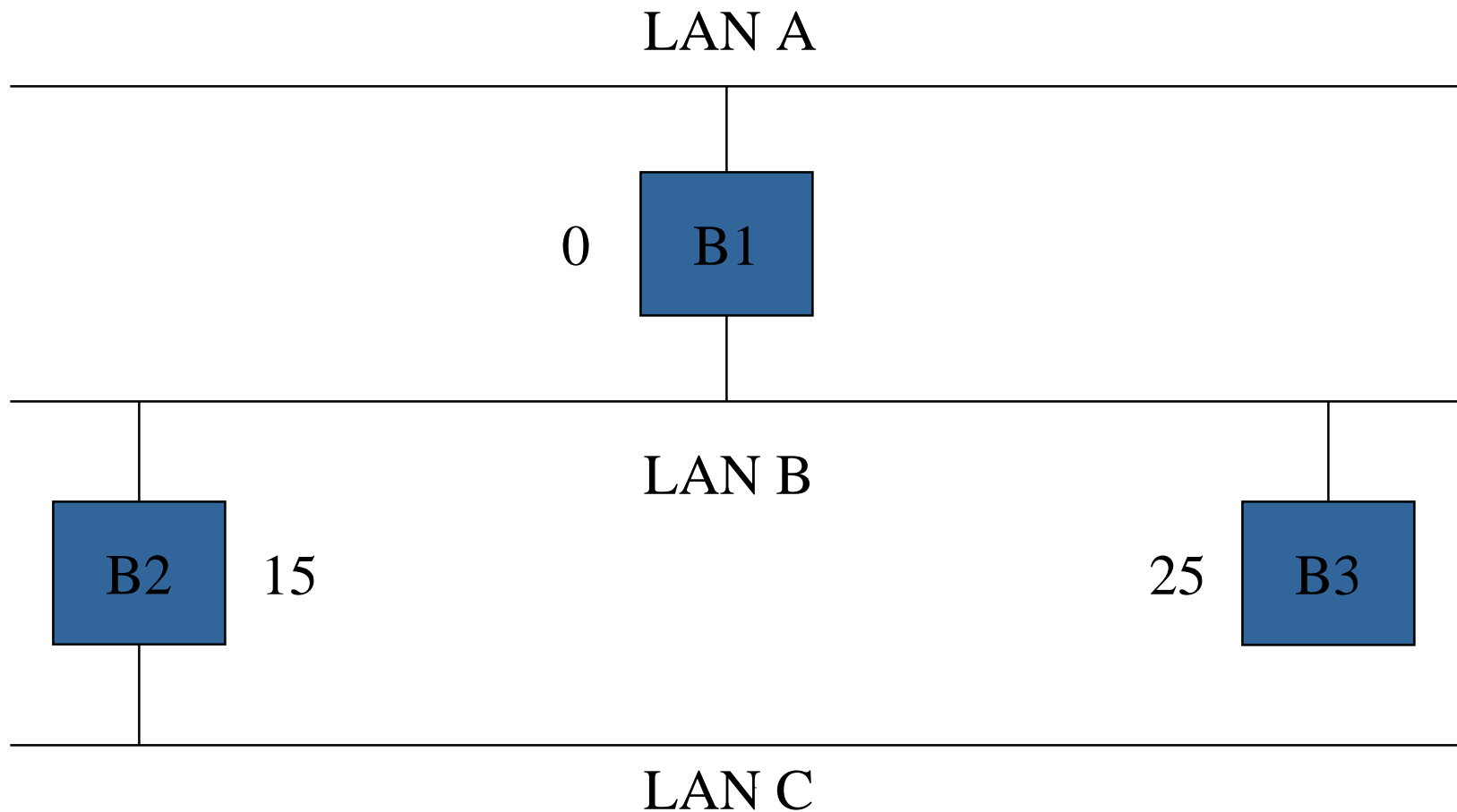
18
2

The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Active Bridged LAN Topology after Bootup



The Ethernet Evolution

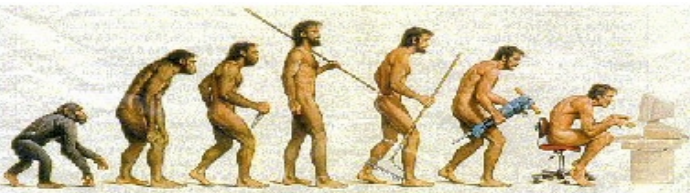
10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

18
3

VLANs

802.1Q



802.1Q - Standard for VLANs

- Defines a method of establishing VLANs
- Establishes the Tagged Frame
- Provides a way to maintain priority information across LANs

P802.1Q/D7
October 3, 1997

Draft Standard P802.1Q/D7 IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks

Sponsor
**LAN MAN Standards Committee
of the
IEEE Computer Society**

Prepared by the Interworking Task Group of IEEE 802.1

Abstract:

This Standard defines:

- An architecture for Virtual Bridged LANs;
- The services provided in Virtual Bridged LANs;
- The protocols and algorithms involved in the provision of those services.

Keywords:

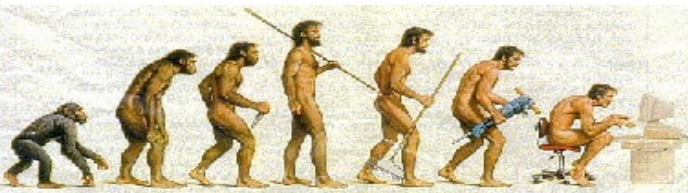
local area networks, media access control bridges, virtual LANs, MAC Bridge management

Copyright © 1997 by the Institute of Electrical and Electronics Engineers, Inc.
345 East 47th Street
New York, NY 10017, USA
All rights reserved.

This is an unapproved draft of a proposed IEEE Standard, subject to change. Permission is hereby granted for IEEE Standards Committee participants to reproduce this document for the purposes of IEEE standardization activities. If this document is submitted to ISO or IEC, notification shall be given to the IEEE Copyright Administrator. Permission is also granted for member bodies and technical committees of ISO and IEC to reproduce this document for the purposes of developing a national position. Other entities seeking permission to reproduce portions of this document for these or other uses, must contact the IEEE Standards Department for the appropriate license. Use of information contained in this unapproved draft is at your own risk.

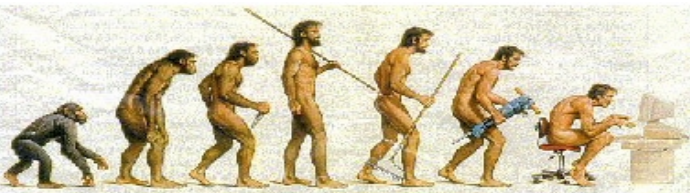
IEEE Standards Department
Copyright and Permissions
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331, USA

Copyright © 1997 IEEE. All rights reserved.
This is an unapproved IEEE/ISO/IEC Standards Draft, subject to change.



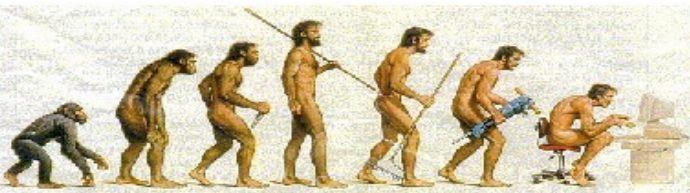
Reasons For Standardizing VLANs

- Old implementations could only be defined in one switch
- To connect a VLAN to another network, each VLAN needed a router port
- The only multi-switch VLANs were proprietary:
 - Cisco: ISL
 - Bay: Lattispan
 - 3Com: VLT
 - Cabletron: SecureFast



Standards Based VLANs

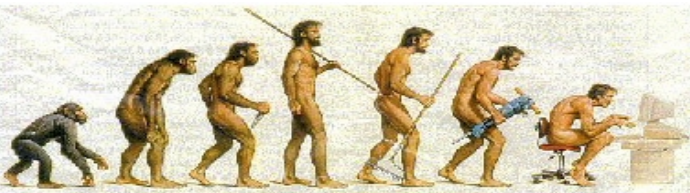
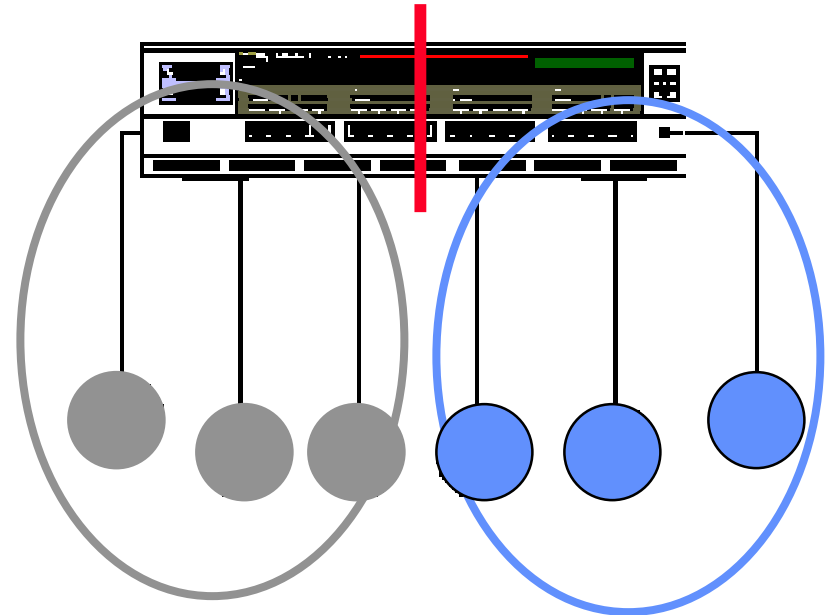
- Includes definition for a new GARP application called GVRP (GARP VLAN Registration Protocol)
 - **Propagate VLAN registration across the net**
- Associate incoming frames with a VLAN ID
- De-associate outgoing frames if necessary
- Transmit associated frames between VLAN 802.1Q compliant switches



What are VLANs - Virtual Local Area Networks ?

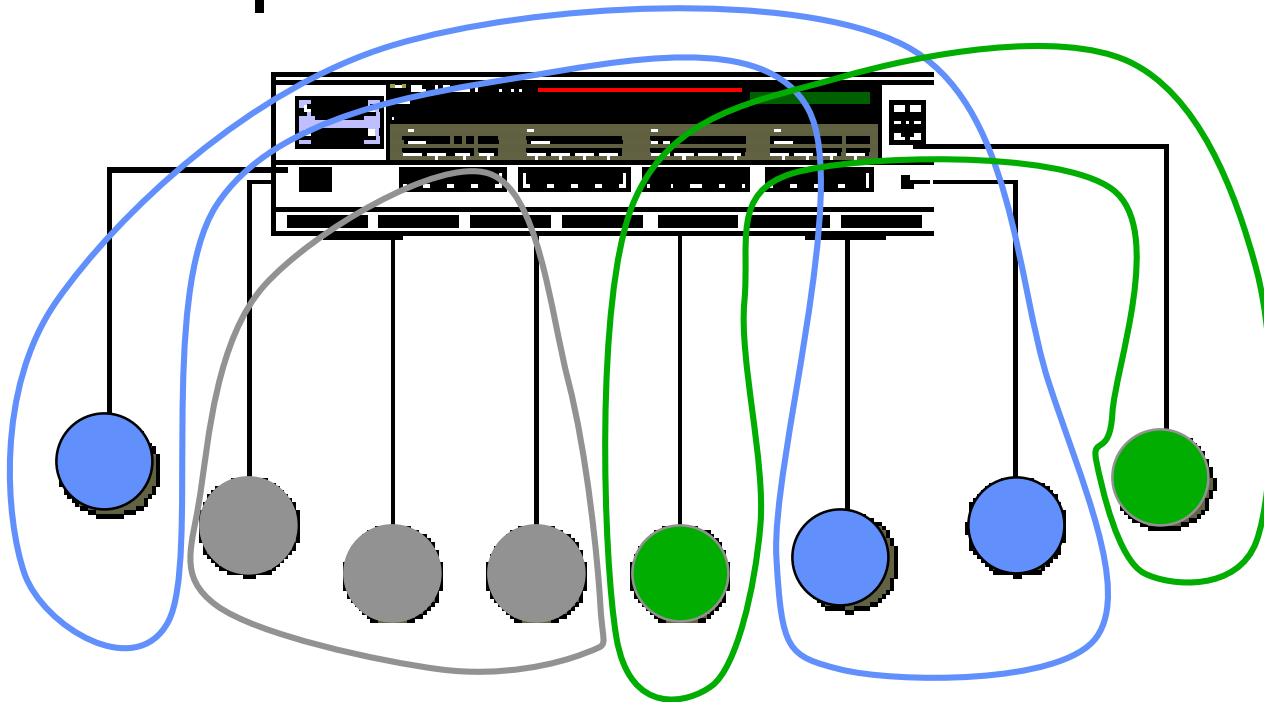
- Divides switch into two or more “virtual” switches with separate broadcast domains
- Achieved by manual configuration through the switches’ management interface
- Only that switch will be segmented

One Broadcast Domain

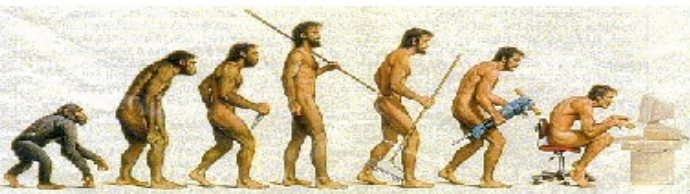


Multiple VLANs in One Switch

Multiple Broadcast Domains



- Multiple VLANs can be defined on the same switch



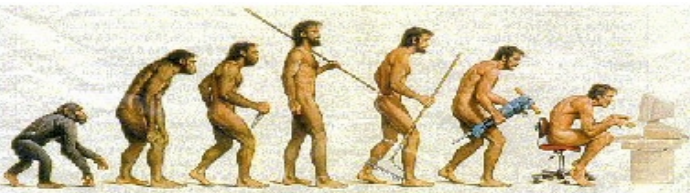
The Ethernet Evolution

10 to 10,000 Mbps

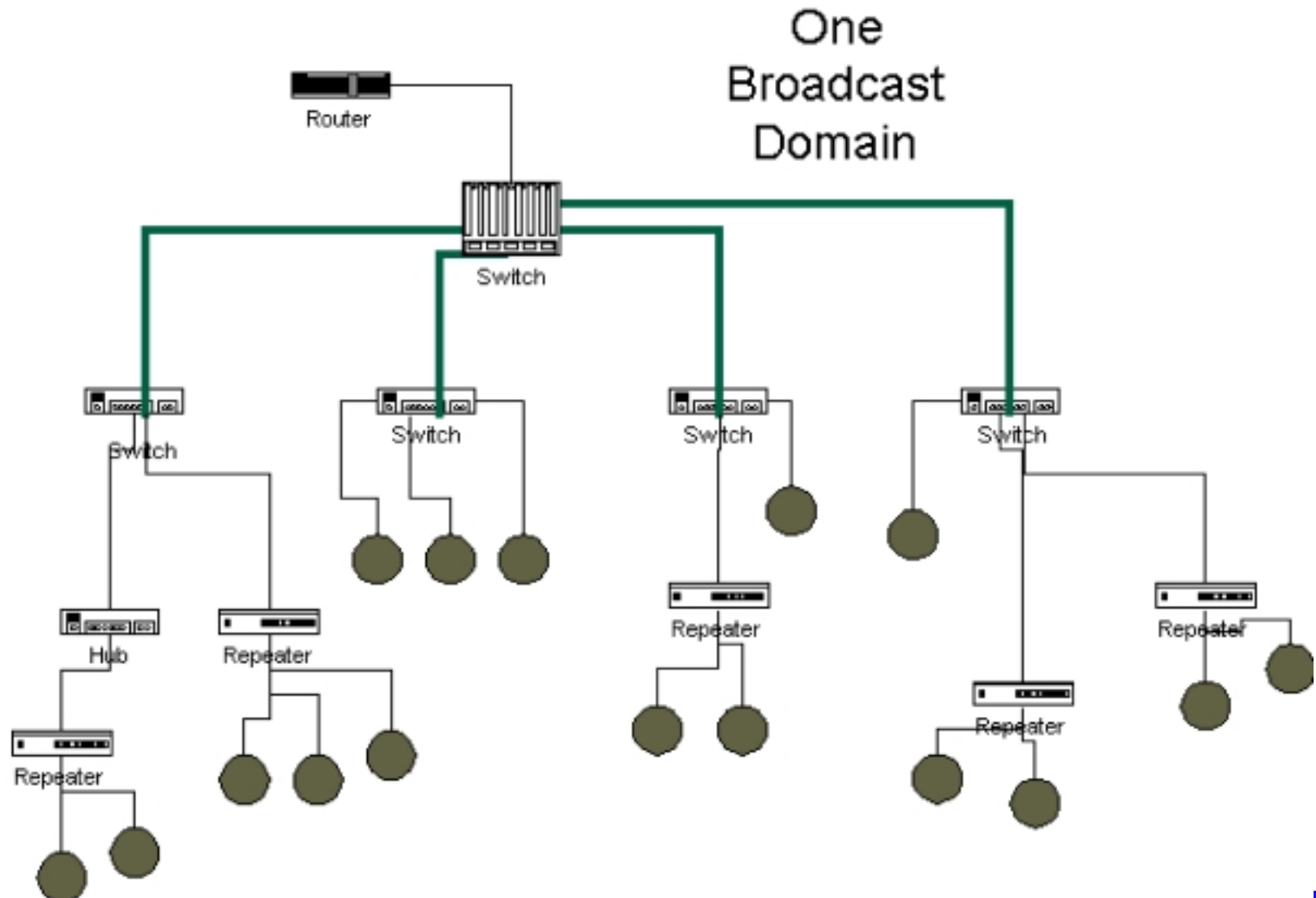
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Why VLANs?

- Lots of broadcast traffic wastes bandwidth
 - VLANs create separate broadcast domains
 - » Microsoft Networking
 - » Novell Networking
 - » NetBEUI
 - » IP RIP
 - » Multicast (sometimes acts like broadcast)
- VLANs can span multiple switches and therefore create separate broadcast domains that span multiple switches



Why Are VLANs Needed?



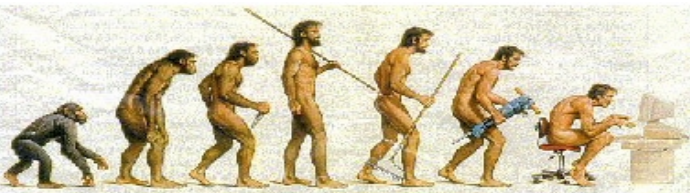
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

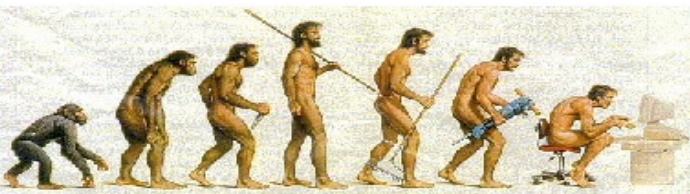
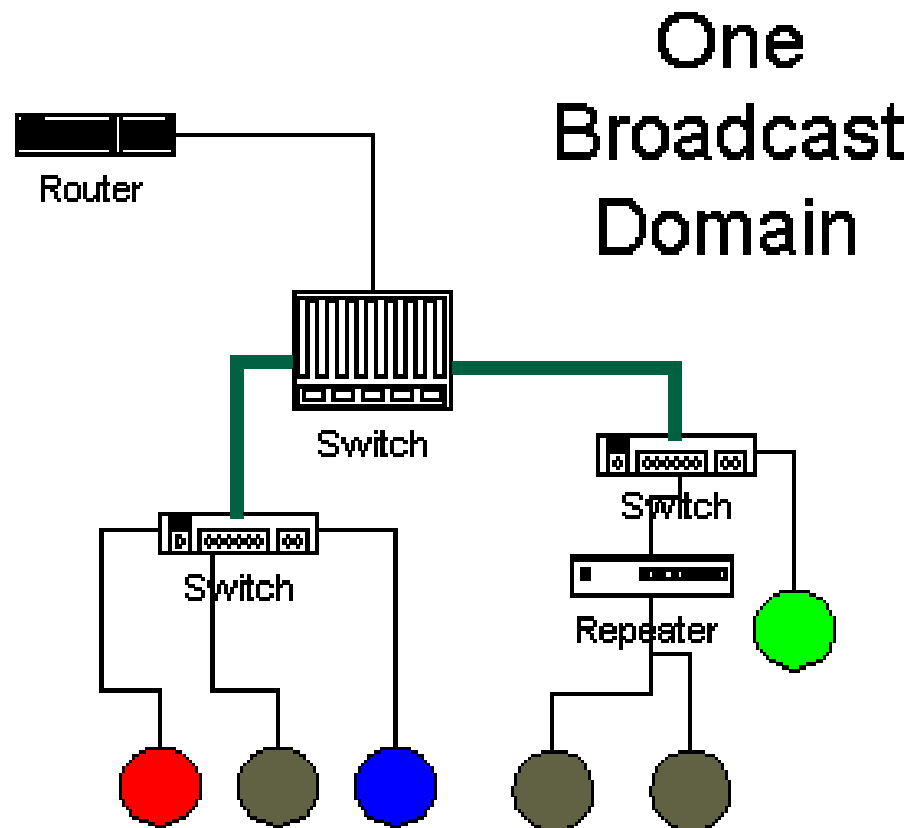
Why VLANs?

- Lots of broadcast traffic wastes bandwidth
 - Microsoft Networking
 - Novell Networking
 - NetBEUI
 - IP RIP
 - Multicast (acts like broadcast)
- High-speed backbones can handle more, but limited by the lower-speed edges



More Reasons...

- Link Multiplexing
 - slower speed technologies share the high-bandwidth uplink
 - multiple IP subnets on one physical link with layer 3 switching (such as to connect Morse, Leavitt and Ocean if we were switched instead of routed)



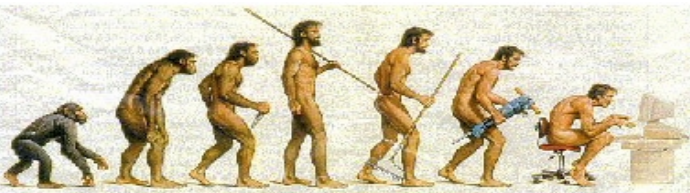
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

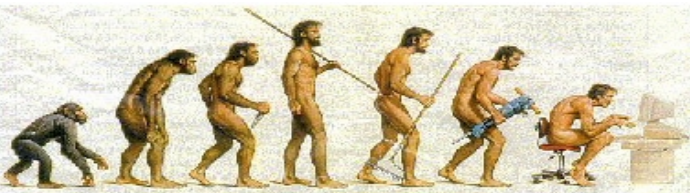
And More Reasons...

- Security
 - **Broadcasts are seen by everyone**
 - **virtual private tunnel**
- Moving end-stations to different ports
- Switching is faster and cheaper than routing, but you still need full routing for some applications (IGMP, VPN)



Basic VLAN Concepts

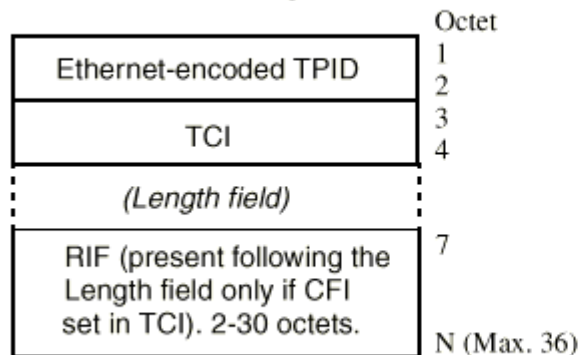
- Port-based VLANs
 - Each port on a switch is in one and only one VLAN (except trunk links)
- Tagged Frames
 - VLAN ID and Priority info is inserted (4 bytes)
- Trunk Links
 - Allow for multiple VLANs to cross one link
- Access Links
 - The edge of the network, where legacy devices attach
- Hybrid Links
 - Combo of Trunk and Access Links
- VID
 - VLAN Identifier



Tagged Frames

- 4 Bytes inserted after Destination and Source Address
- Tagged Protocol Identifier (TPID) = 2 Bytes (x8100)
 - length/type field
- Tagged Control Information (TCI) = 2 Bytes
 - contains VID

Ethernet-encoded Tag Header



SNAP-encoded Tag Header

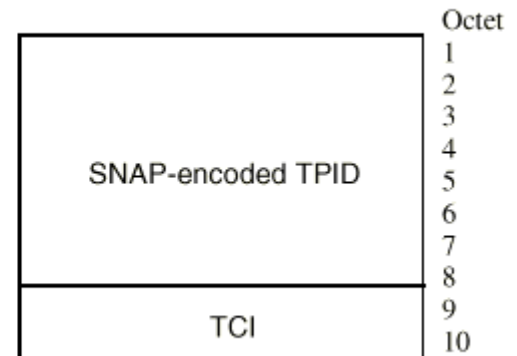


Figure 9-1—Tag Header format

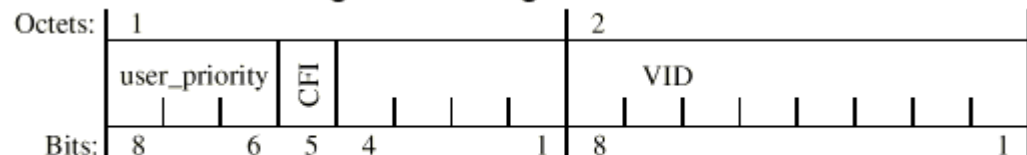


Figure 9-4—Tag Control Information (TCI) format

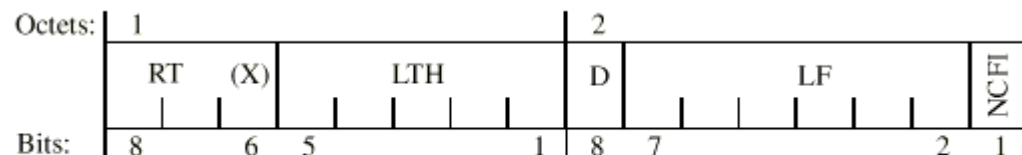
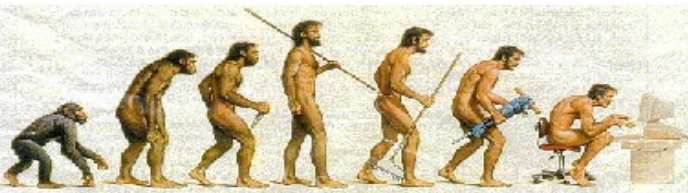


Figure 9-5—Route Control (RC) field



The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Trunk Link

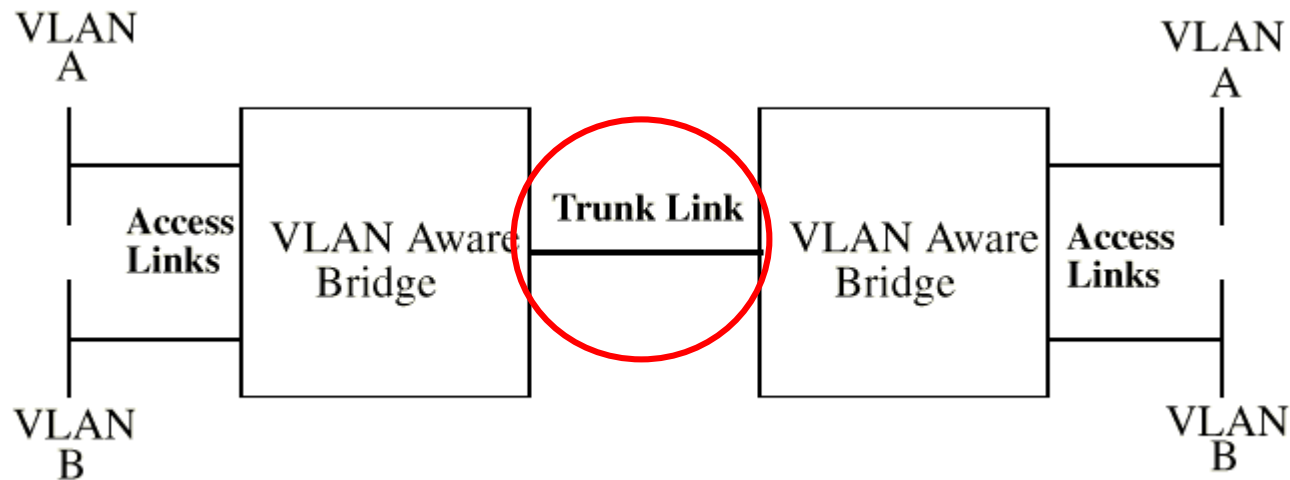
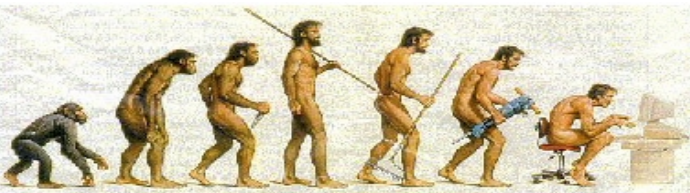


Figure C-1—Port-based VLANs

- Attaches two VLAN switches - carries Tagged frames ONLY.



Access Links

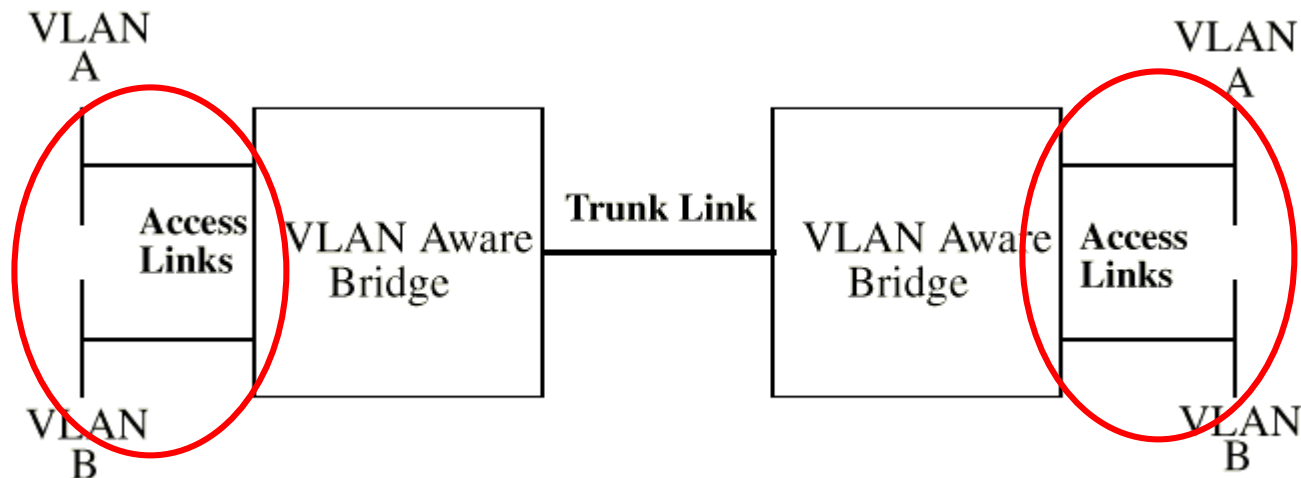
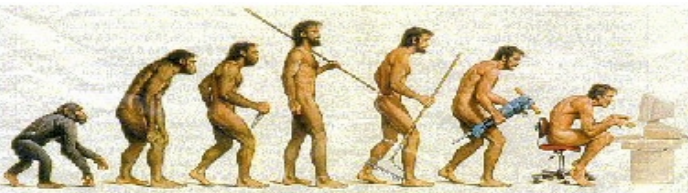


Figure C-1—Port-based VLANs

- Access Links are Untagged for VLAN unaware devices
- the VLAN switch adds Tags to received frames, and removes Tags when transmitting frames.



Hybrid Links

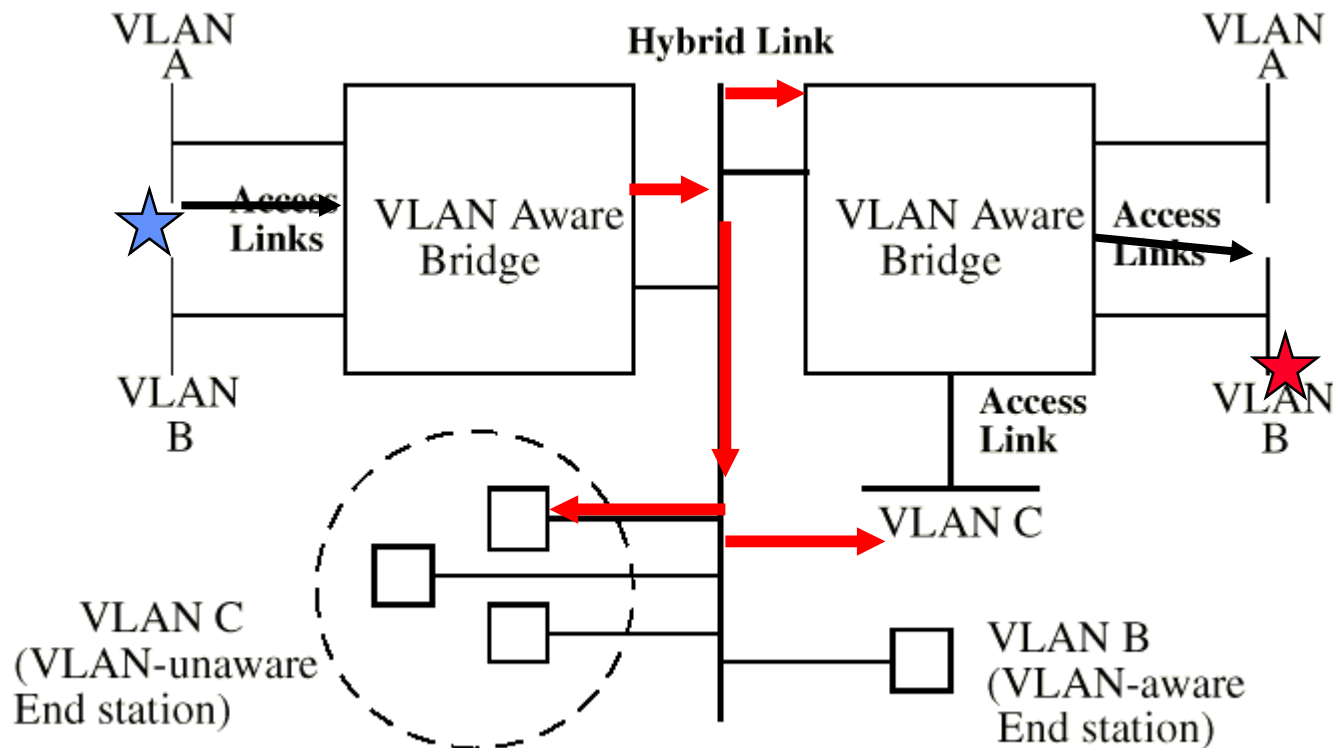
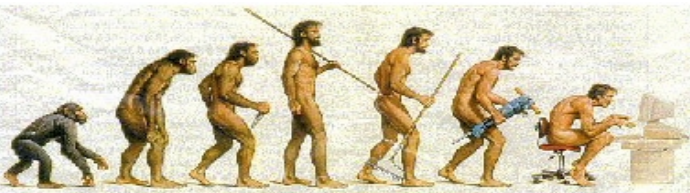


Figure C-2—Hybrid Links

- Hybrid Links - ALL VLAN-unaware devices are in the same VLAN



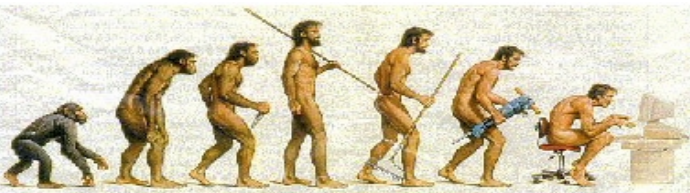
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

So Far So Good...

- So one might ask: “what does the Bridge table look like?”
- Two answers:
 - multiple (distinct) tables: one for each VLAN
 - one mother of all tables, with a VLAN column
- They sound similar, but it turns out they are VERY different

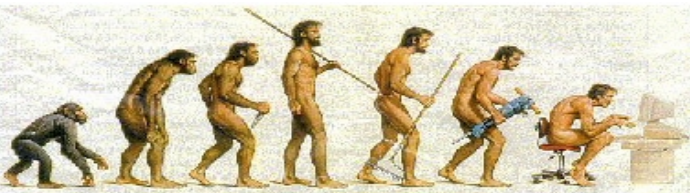


Multiple Tables

Each Table is
for One VLAN

- Called MFD (multiple forwarding databases) or Independent Learning
- Each VLAN learns MAC addresses independently, so duplicate MAC addresses are OK.

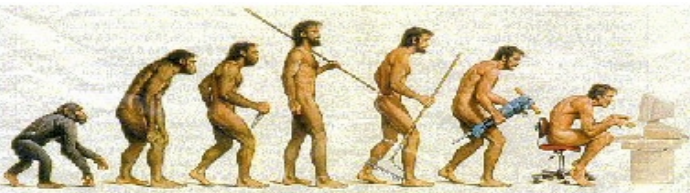
Entry	MAC Addr	Port	active
1	0800900A2580	1	yes
2	002034987AB1	1	yes
3	0500A1987C00	2	yes
4	00503222A001	2	yes
5			
6			
7			
8			
9			
10			
11			
12			



One (Big) Table

- Called SFD (Single Forwarding Database) or Shared Learning
- No duplicate MAC addresses
- Asymmetric VLAN possible

Entry	MAC Addr	Port	active	VLAN
1	0800900A2580	1	yes	2
2	002034987AB1	1	yes	2
3	0500A1987C00	2	yes	2
4	00503222A001	2	yes	2
5	080034090478	3	yes	1
6	049874987AB1	5	yes	1
7	0555A1945600	5	yes	3
8	00503222A023	5	yes	2
9				
10				
11				
12				

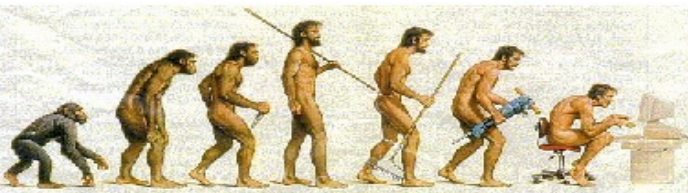
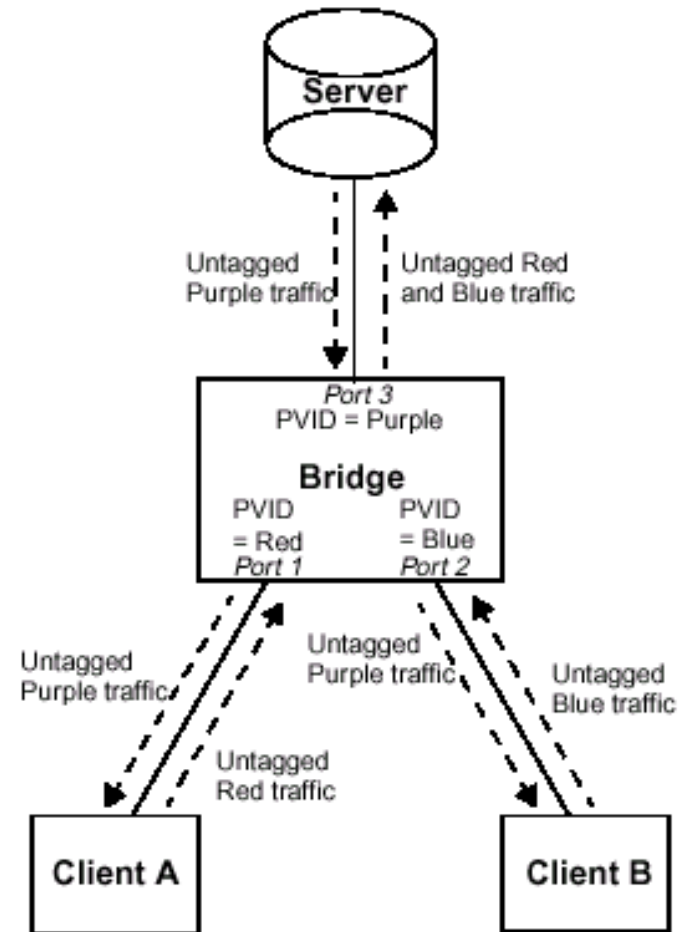


Asymmetric VLANs (also known as “interoperability killer”)

- Legacy server can talk to legacy clients with one physical link
- Legacy clients cannot talk to each other

Member Sets:
Purple - Ports 1 and 2
Red - Port 3
Blue - Port 3

Untagged Sets:
Purple - Ports 1 and 2
Red - Port 3
Blue - Port 3



Independent Learning I

- Legacy router learns MAC addresses from both VLANs
- Requires 2 physical links

Member Sets:
Red - Ports 1, 3
Blue - Ports 2, 4

Untagged Sets:
Red - Ports 1,3
Blue - Ports 2,4

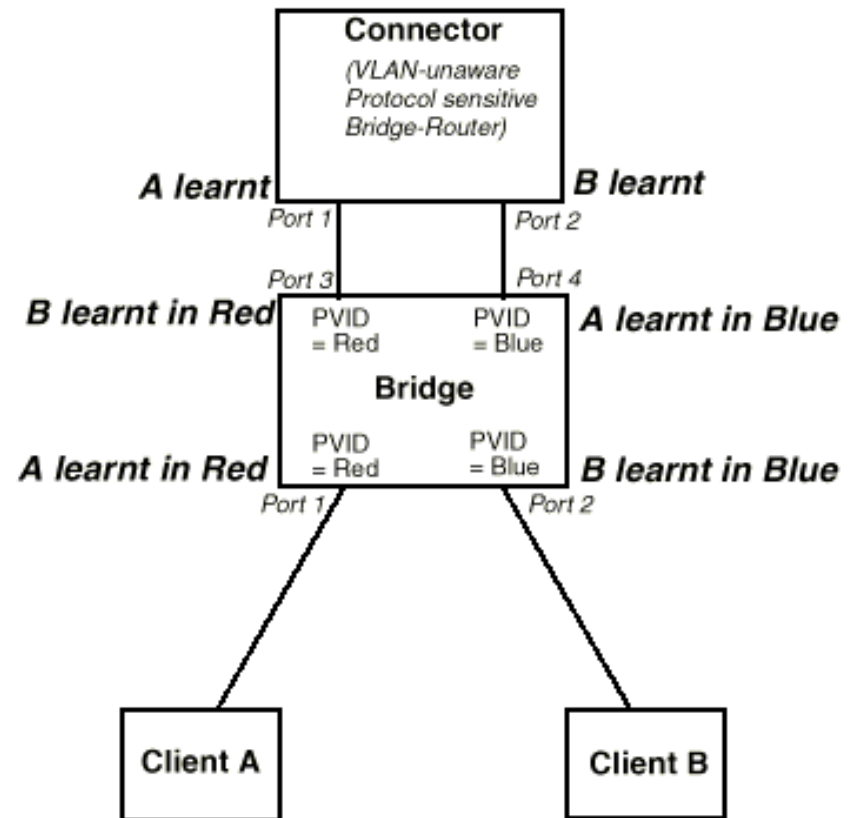
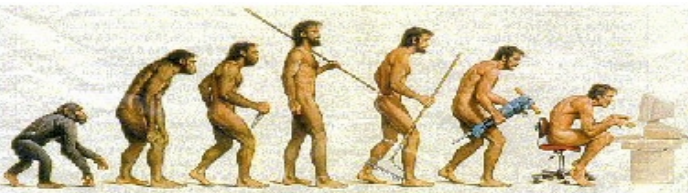


Figure B-2—Connecting independent VLANs - 1



Independent Learning II

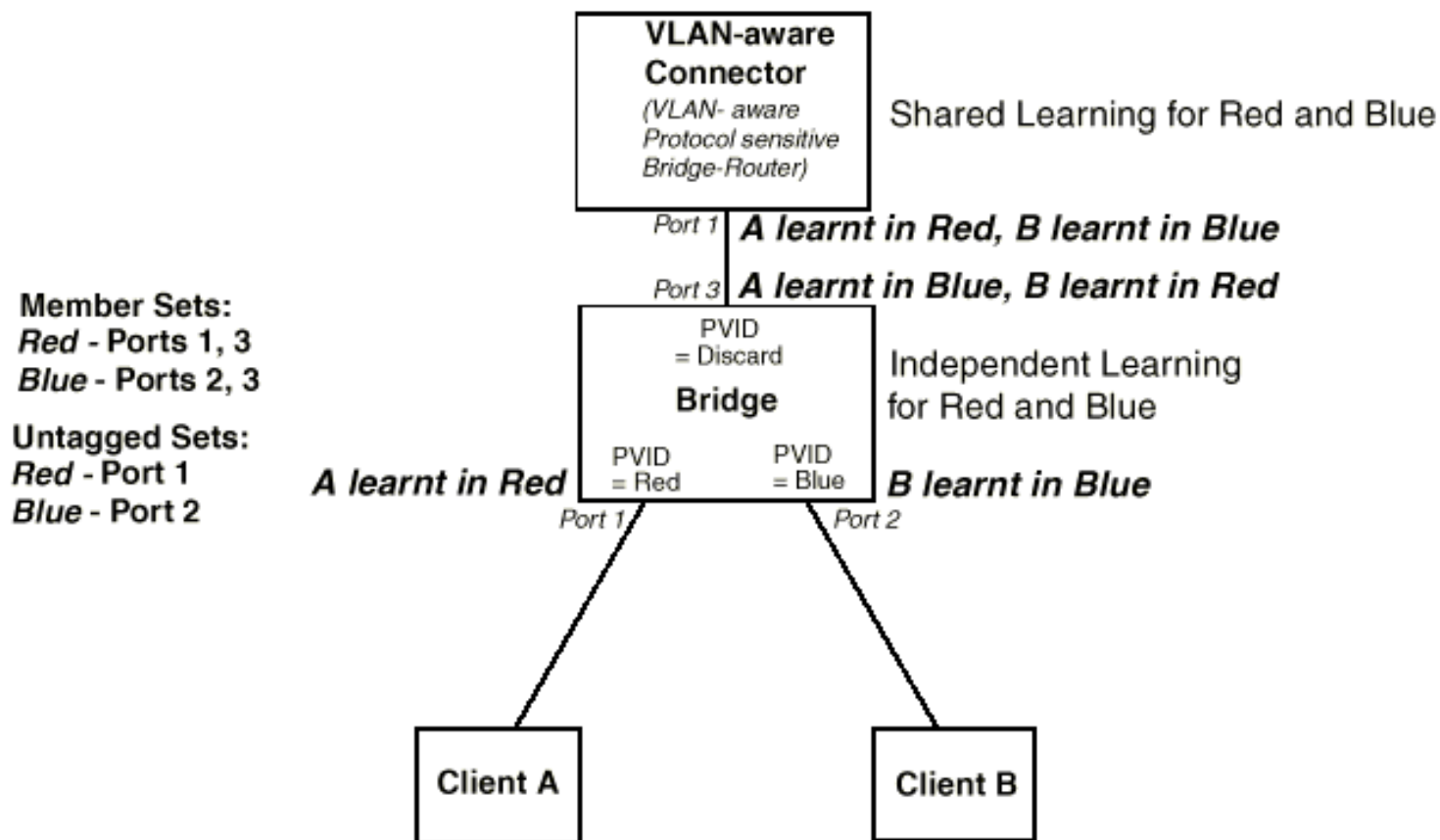
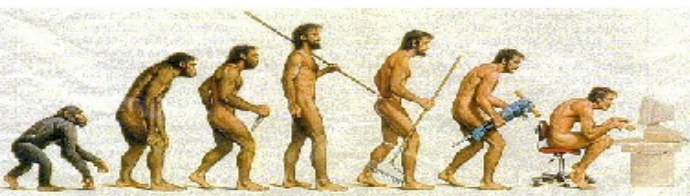


Figure B-3—Connecting independent VLANs - 2

- VLAN-aware router only needs one physical link



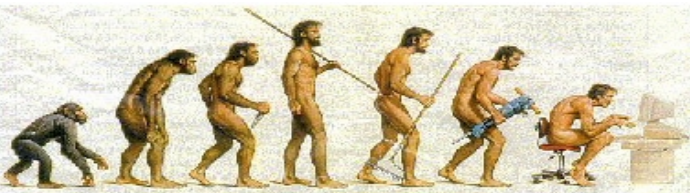
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
 September 25, 2000

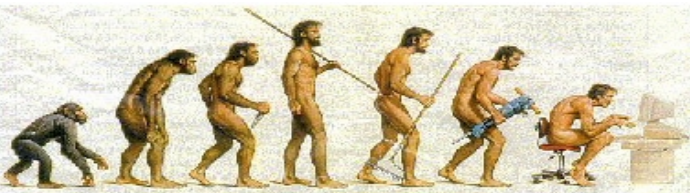
Problems

- Can't combine SFD and MFD switches in one network
- Some switches only do one or the other, and can't be changed
- Hybrids of SFD and MFD makes this tricky



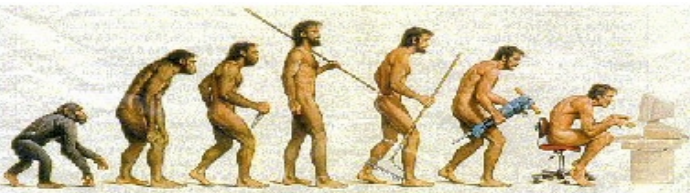
Future Additions

- Layer 3 based VLANs
 - IP traffic on a different VLAN than IPX
- Multiple Spanning Trees (one per VLAN)
 - allows for using the disabled links
- ATM to IEEE VLAN mapping
 - Emulated LANs



GARP (yeah, I know, “the world according to”... that’s a new one!)

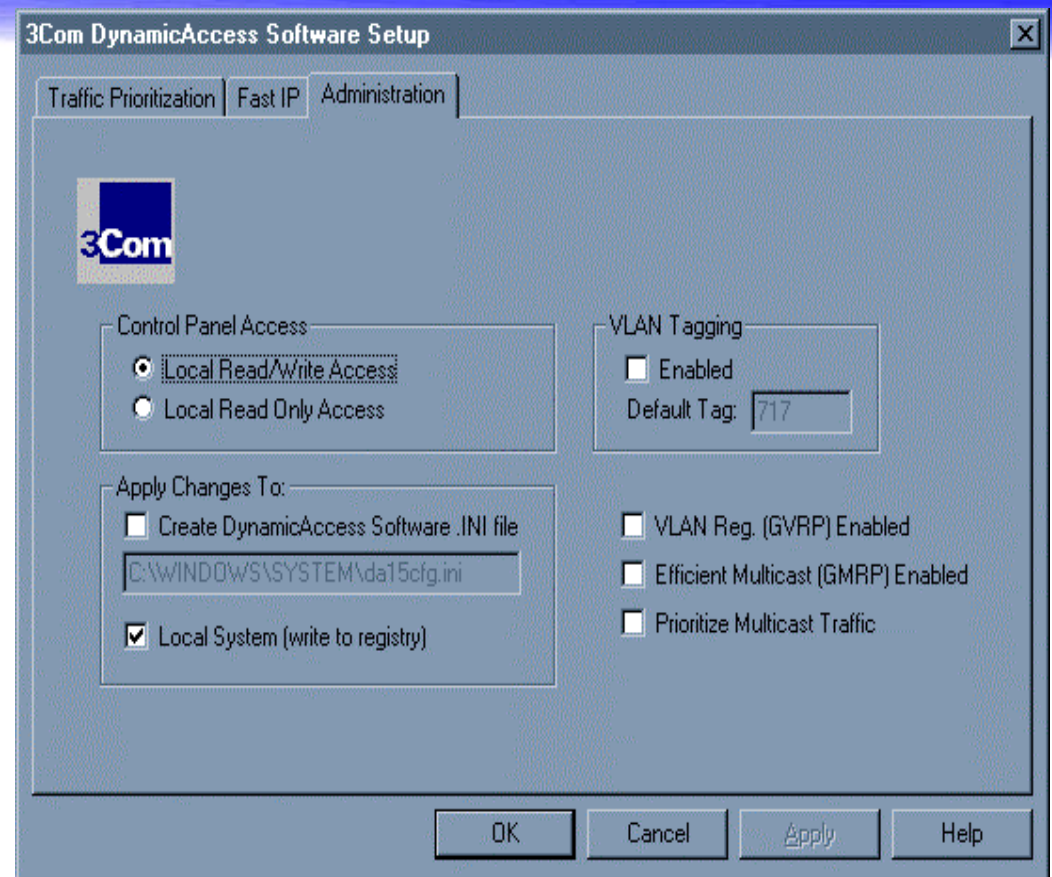
- Generic Attribute Registration Protocol
- Standard Defines:
 - method to declare attributes to other GARP participants
 - frame type to convey GARP messages: Protocol Data Unit (PDU)
 - rules and timers for registering/de-registering attributes
- GVRP: GARP VLAN Registration Protocol
 - GARP based method for VLAN info propagation
- GMRP: GARP Multicast Registration Protocol
 - GARP based method for multicast group propagation



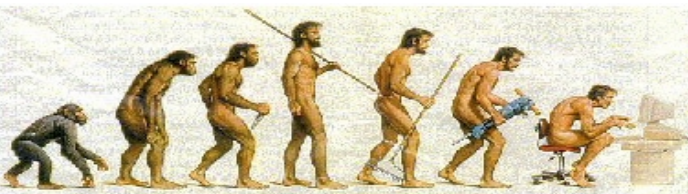
Windows screenshot —>

Vendors (current):
Cisco Systems, 3Com
and Hewlett Packard

Several others are
developing working
implementations also.

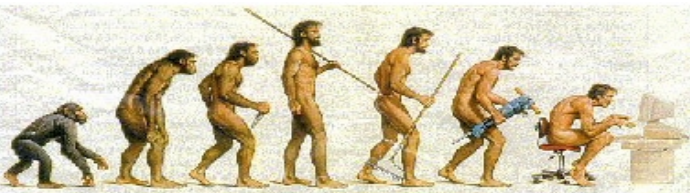


- Industry Implementation Example
 - 3Com manufactures Network Interface Cards that take advantage of GVRP
 - Accessed via the Control Panel (DynamicAccess®)
 - Extremely easy to configure



Priorities/QoS

802.1p



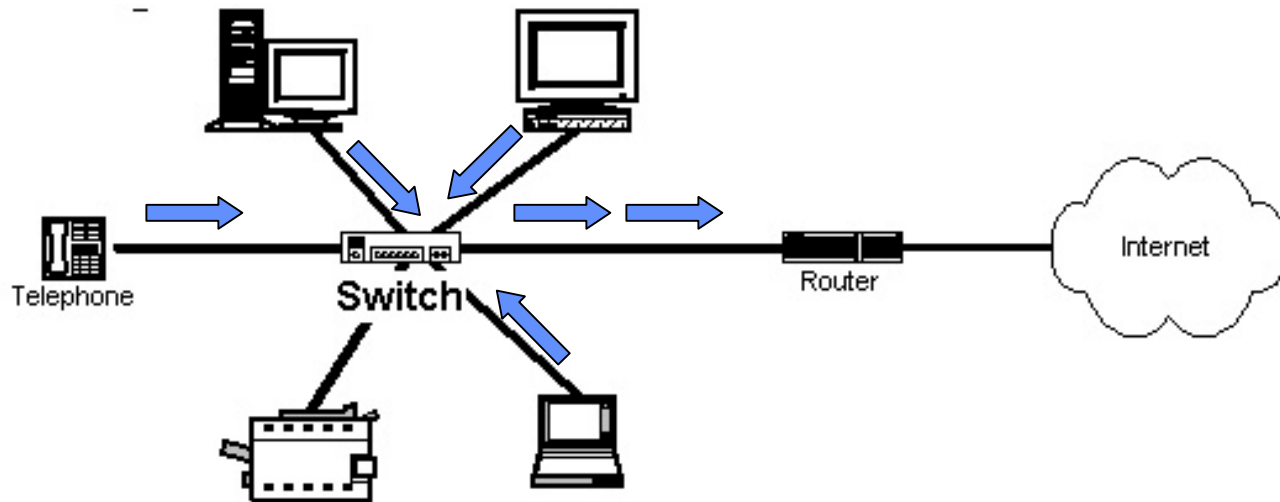
21
0

The Ethernet Evolution

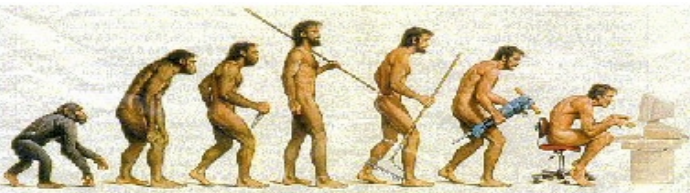
10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

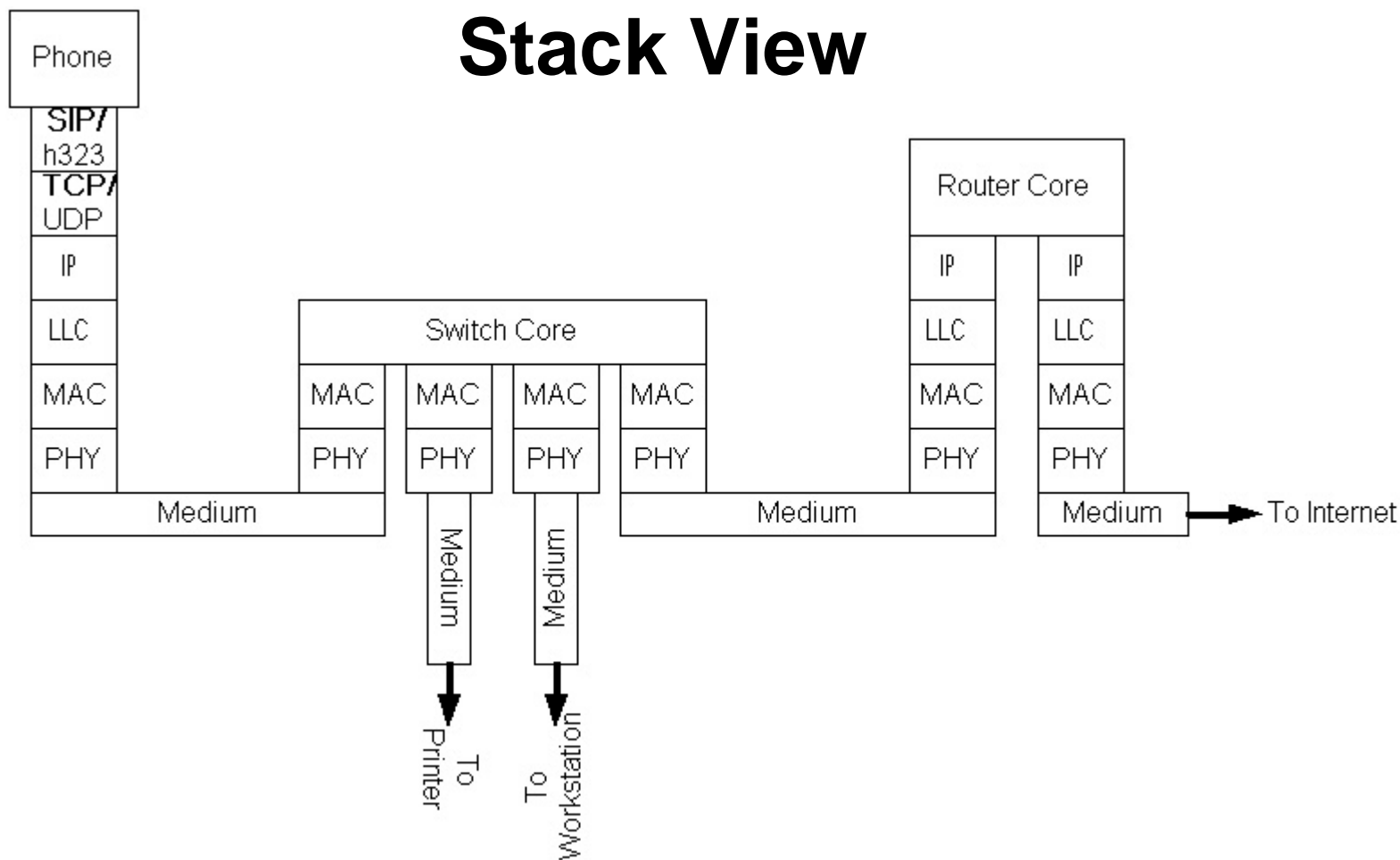
The Problem



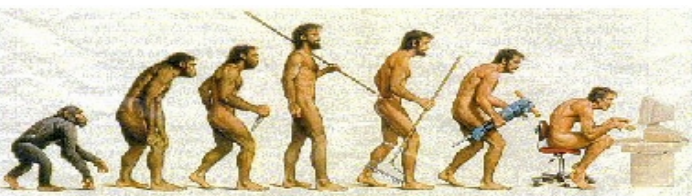
- In a switched network, if more than one device sends packets to the same destination at the same time, the switch will buffer them and send them out based on first-in first-out basis
- So time-sensitive traffic (like voice) has no priority over other types (mail, ftp, web)



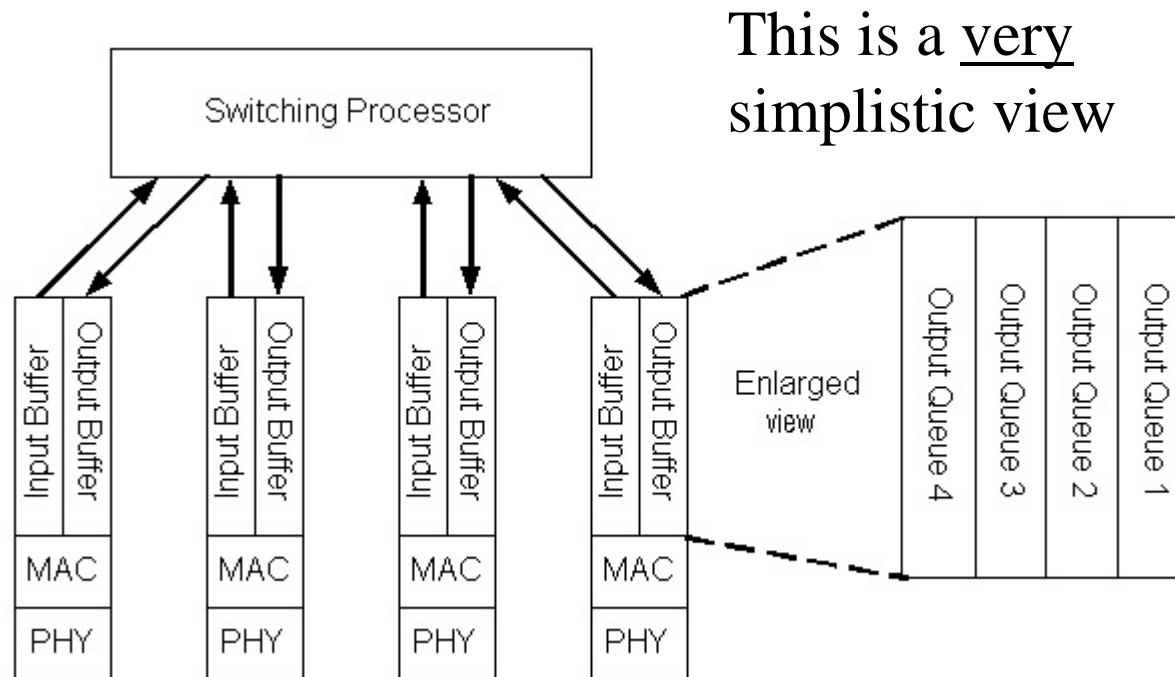
Stack View



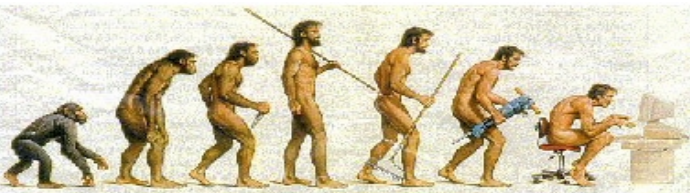
- Same switched network drawn as stack models



Inside the Switch

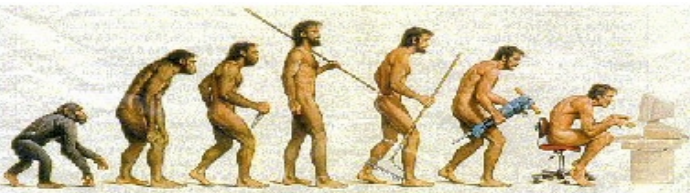


- Buffers can have multiple queues for different priority traffic
- Higher-priority can be sent before lower-priority



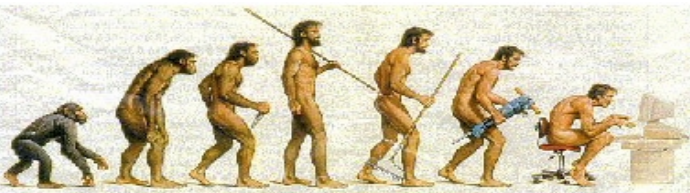
Possible Solution

- Q: How does switch know which traffic to prioritize?
- Possible answer: set priority for each port inside the switch configuration...
- But then how does it tell the next switch along the path what the priority was?
- And not all traffic from the same port should be treated equally - it may come from a router, or a shared link, or another switch



Better Solution

- Tag the priority of each packet with its relative priority value
- Then switch can put higher-priority tagged frames in a higher-priority queue based on the value



Remember VLAN Tagged Frames?

- 4 Bytes inserted after Destination and Source Address
- Tagged Protocol Identifier (TPID) = 2 Bytes (x8100)
 - length/type field
- Tagged Control Information (TCI) = 2 Bytes
 - contains VID

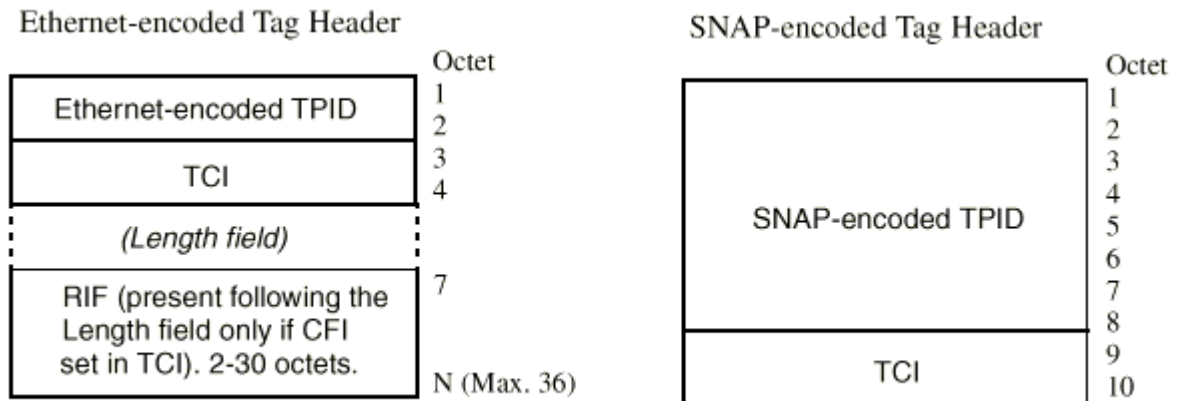


Figure 9-1—Tag Header format

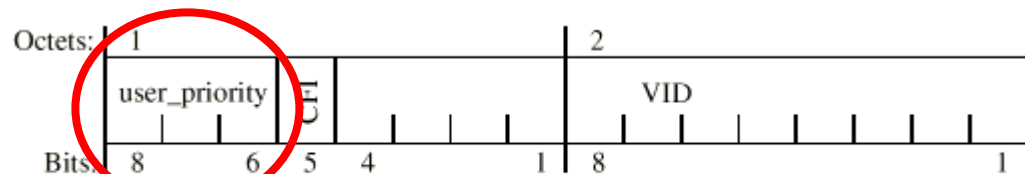


Figure 9-4—Tag Control Information (TCI) format

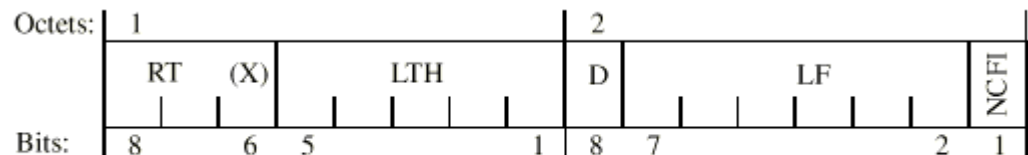
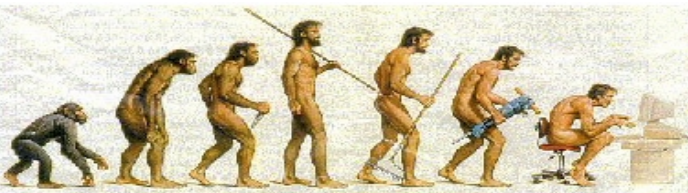


Figure 9-5—Route Control (RC) field



The Ethernet Evolution

10 to 10,000 Mbps

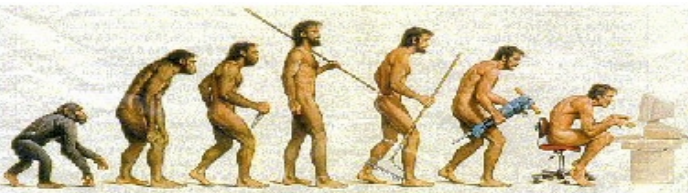
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Priority Tagged Frames



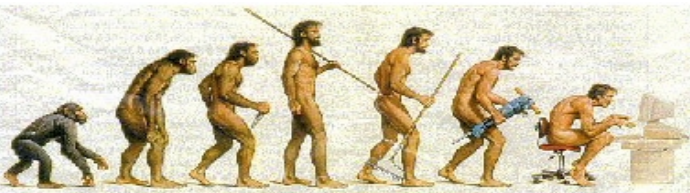
Figure 9-4—Tag Control Information (TCI) format

- Uses `user_priority` field value of 0-7 for priority numbers
- VID may be a real VLAN ID number, or the value of `x000` = “Priority Tagged”, meaning not a VLAN Tagged frame (`x000` is illegal VLAN value)
- Same Ethertype of `x8100`

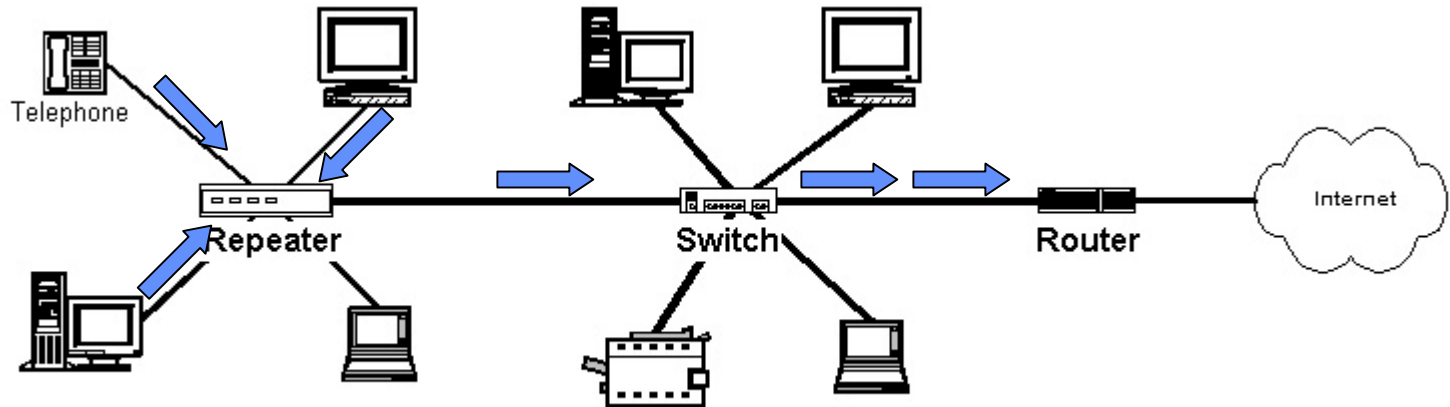


Limitations of 802.1p

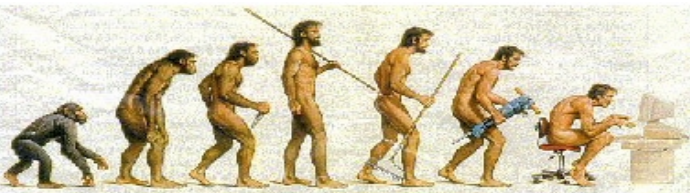
- It is NOT Quality of Service (QoS)
 - There is no guarantee of bandwidth, latency, nor jitter
 - If other devices use the same priority, no difference
 - If higher-priority fills the pipe constantly, lower-priority will never get through (depending on implementation)
- It IS Class of Service (CoS)
- Access to a shared ethernet medium/network isn't helped - it's still under the rules of CSMA/CD



Example



- Even if the phone uses 802.1p tagged frames, it has to share the repeated network with the other devices with the same odds of success
- Once its frames get to the switch, then they will be sent through faster and with better odds than frames from other devices

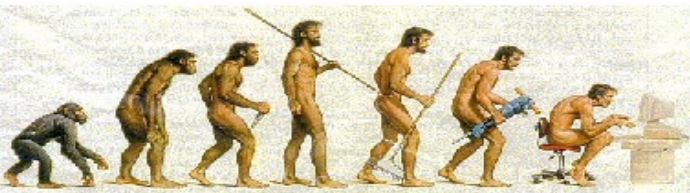


The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Layer-3 Switching



22
0

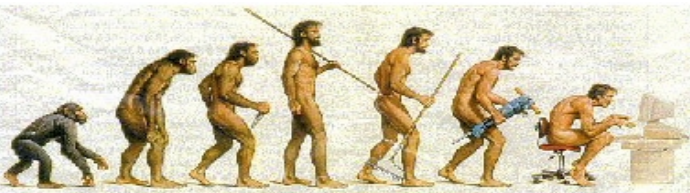
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

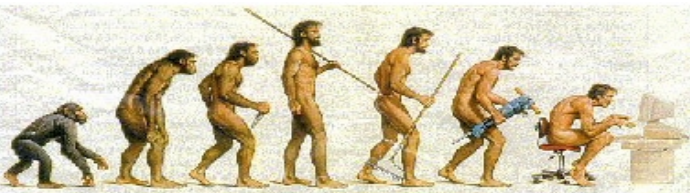
Routing vs. Switching

- Routing is done at layer 3 (IP layer), thereby connecting IP subnets
- Switching is done at layer 2 (MAC layer), thereby connecting Ethernet LAN segments
- Routing is slow because it's done in software
- Switching is fast because it's done in hardware
- So why not Route in hardware?

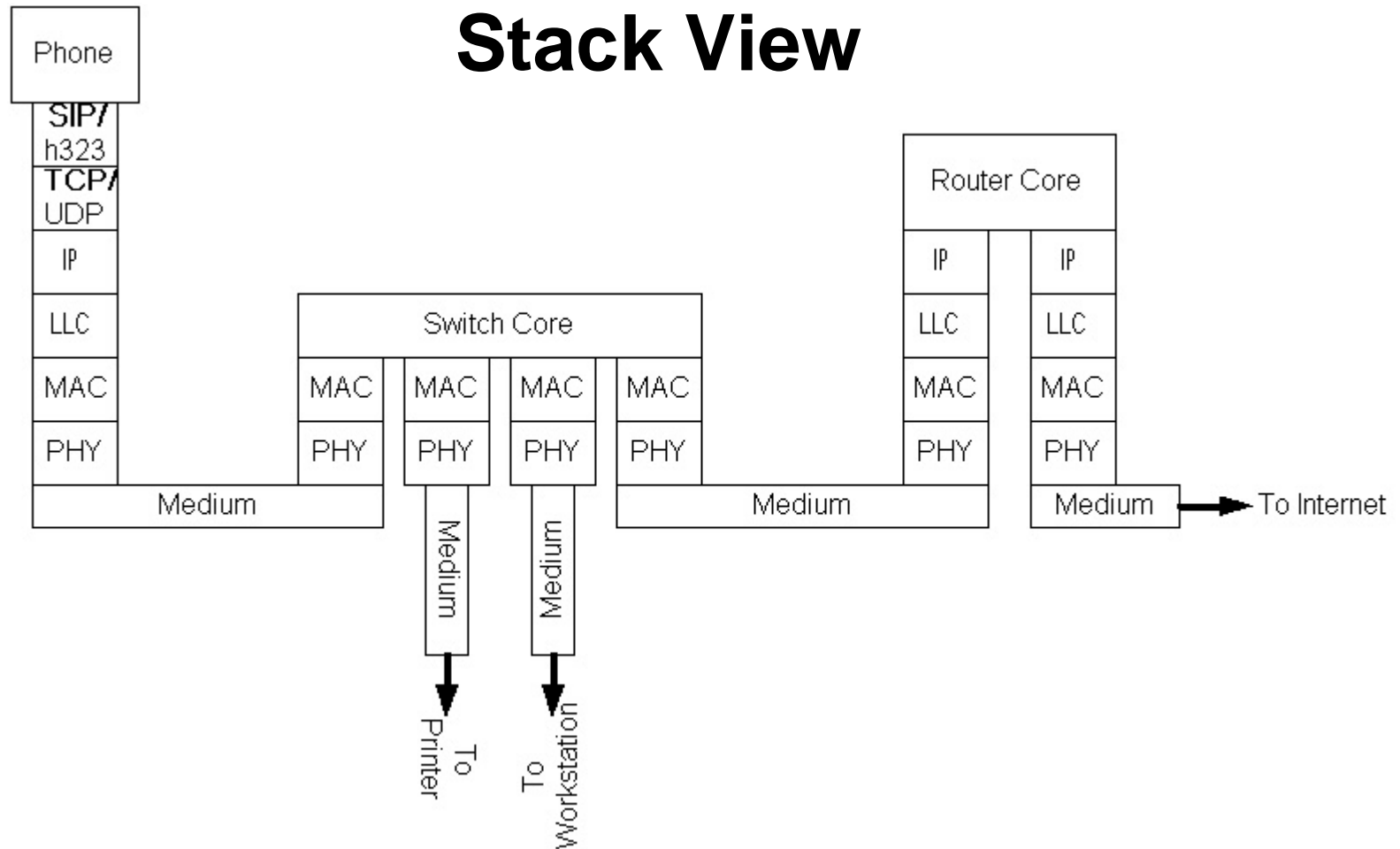


Routing in Hardware

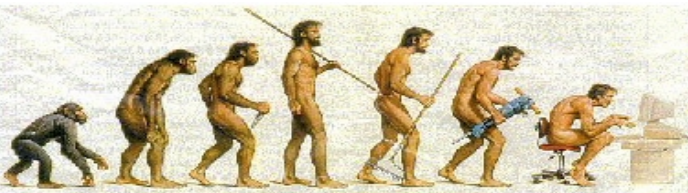
- Put forwarding engine into hardware, leave routing protocols/database in software
- Remove unnecessary or seldom-used protocols (IPX, Appletalk, etc.)
- Don't route to non-Ethernet technologies (Token Ring, FDDI, Frame Relay, ATM, etc.)
- Lose advanced functions (statefull firewall, NAT, IP header compression, etc.)



Stack View



Remember this?

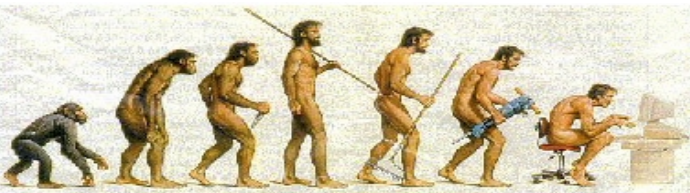
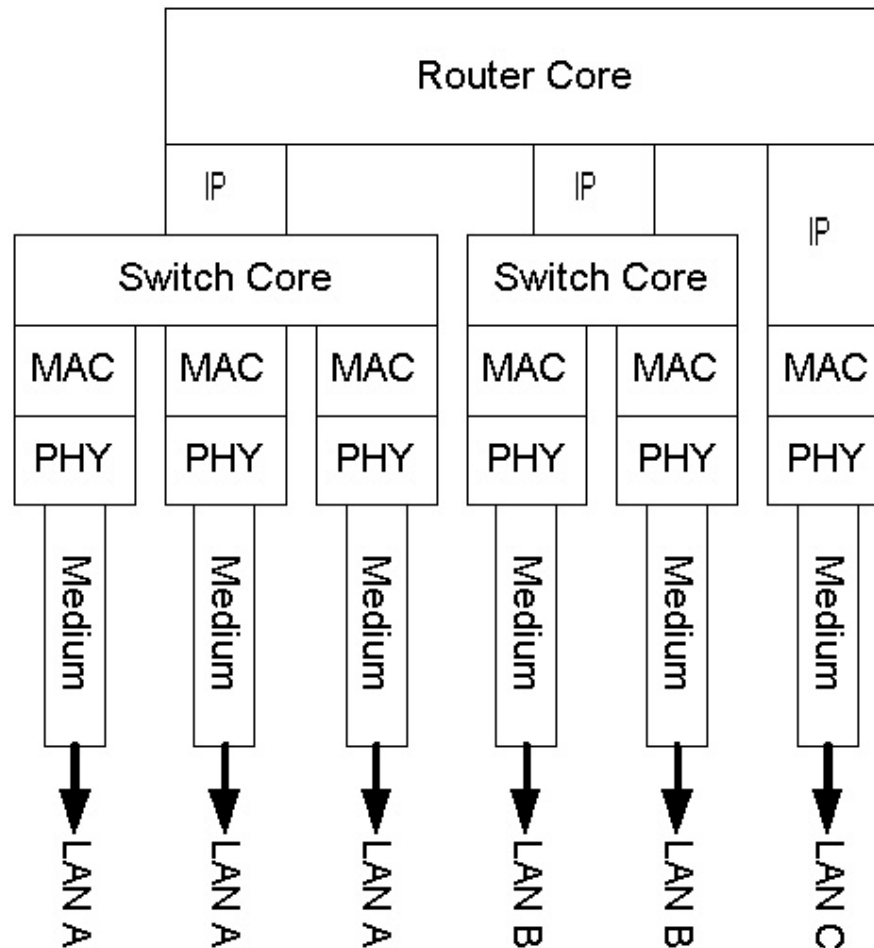


The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Layer-3 Switch Stack View



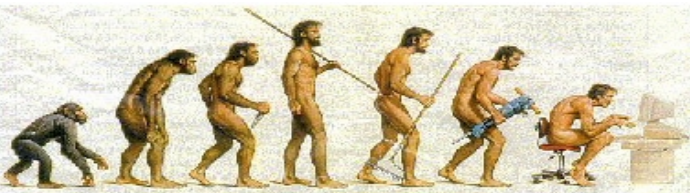
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

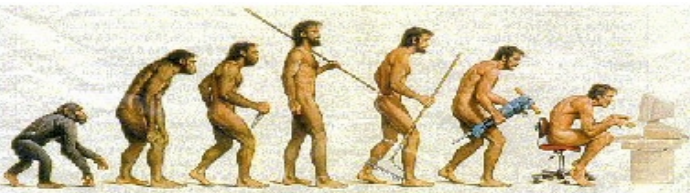
Layer-3 Switch Implementations

- Actual implementations vary depending on architecture
- Most have fast-forwarding engines per blade which can forward based on IP Address
- Usually, there is one or more central CPUs to handle exception cases, routing updates and management
- Some L3 Switches now offer advanced features:
 - filtering based on address, ports, etc.
 - IP multicast in the fast path
 - IPX, Appletalk, etc.
 - Advanced routing protocols such as BGP



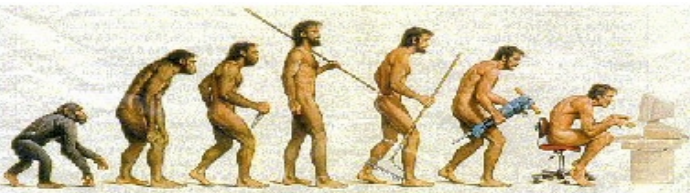
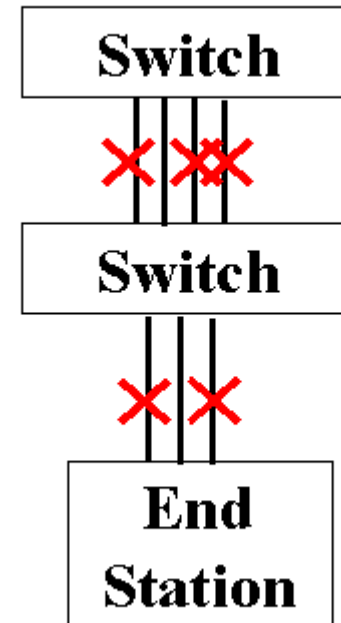
Link Aggregation

802.3ad



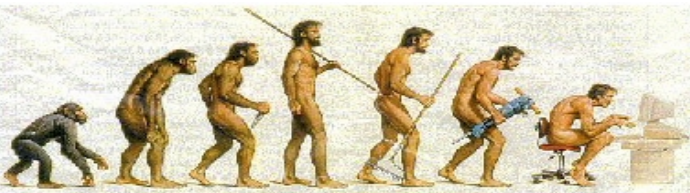
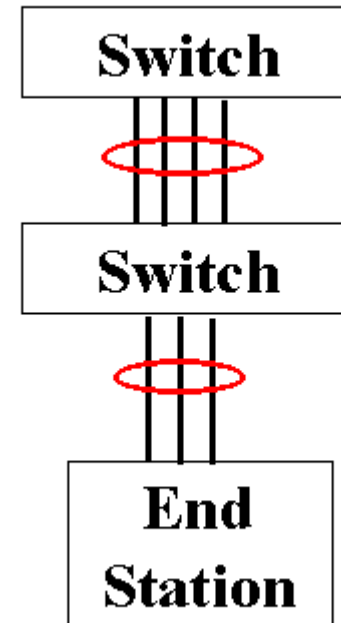
Basic Concept

- Originally only one of many separate physical links could be used (else loops)
- Spanning tree used to disable all but one
- Did provide redundancy, but not efficient use of links



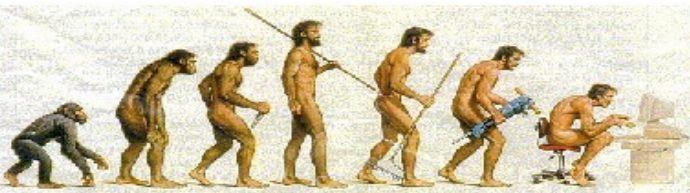
Enter Trunking

- Multiple physical links joined into one logical link
- Only one MAC address used for trunk group
- All links in same group must be same speed and full-duplex



Proprietary Today

- Very popular with users, but proprietary
- EtherChannel, Fast EtherChannel, Gig
 - Cisco's protocol (routers, switches)
 - Also licensed by Adaptec, Intel, Compaq, Sun, HP, ZNYX, Phobos, Auspex (all NICs)
- MultiLink Trunking (MLT)
 - Nortel's protocol (switches)
- Virtual Link Trunking (VLT)
 - 3Com's protocol (switches)



802.3ad - Link Aggregation

- Study group started in November 1997
- Working group started in July 1998
- Expected standard in Summer 1999 (wow!)
- website:
<http://grouper.ieee.org/groups/802/3/ad/>

P802.3ad/D1.0
February 8, 1999

IEEE Draft P802.3ad/D1.0

Supplement to Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method & Physical Layer Specifications:

Link Aggregation

Sponsor

LAN MAN Standards Committee
of the
IEEE Computer Society

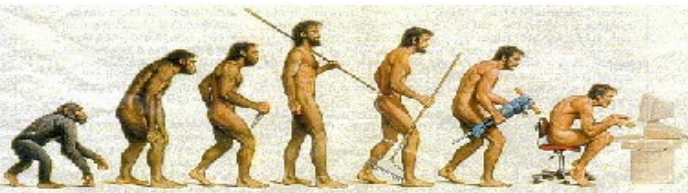
This Draft was Prepared by Tony Jeffree, David Law, and Rich Seifert for Task Force Ballot, as directed at the interim meeting held in January 1999 in Miami Beach, FL. This draft expires on May 1, 1999.

Copyright © 1999 by the Institute of Electrical and Electronics Engineers, Inc.
345 East 47th Street
New York, NY 10017, USA
All rights reserved.

This is an unapproved draft of a proposed IEEE standard, subject to change. Permission is hereby granted for IEEE Standards Committee participants to reproduce this document for the purposes of IEEE standardization activities. If this document is submitted to ISO or IEC, authorization shall be given in the IEEE Copyright Administration. Permission is also granted for member bodies and technical committees of ISO and IEC to reproduce this document for the purposes of developing a national position. Other entities seeking permission to reproduce portions of this document for those or other uses, must contact the IEEE Standards Department for the appropriate license. Use of information contained in this unapproved draft is at your own risk.

IEEE Standards Department
Copyright and Permissions
400 Ross Lane, P.O. Box 1331
Piscataway, NJ 08855-1331, USA

Copyright © 1999 IEEE. All rights reserved. This is an unapproved IEEE Standards Draft, subject to change.

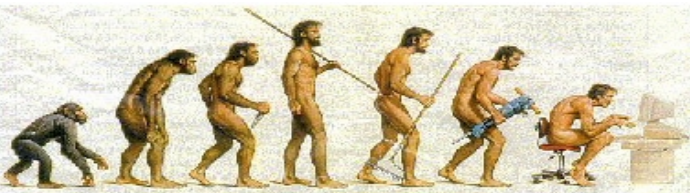


23
0

10 to 10,000 Mbps
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Benefits and Goals

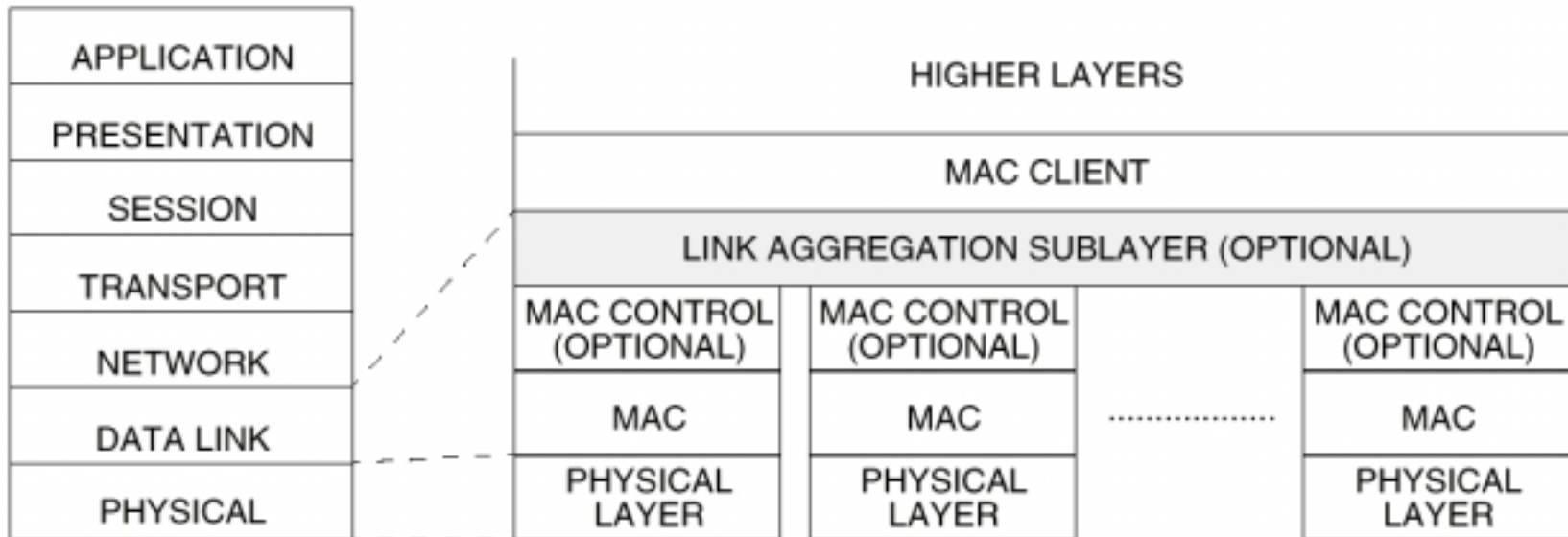
- Increased Bandwidth (duh)
- Incremental bandwidth increases - smaller jumps than 10 to 100 to 1000 - not perfect scaling, but better than one link's worth
- Failure protection/redundancy (maybe)
- Automatic configuration - if it can aggregate, it will
- Rapid reconfiguration
- Deterministic behavior - configuration will not be dependent on the order in which events occurred
- NOTE: some of these goals are inconsistent (see later)



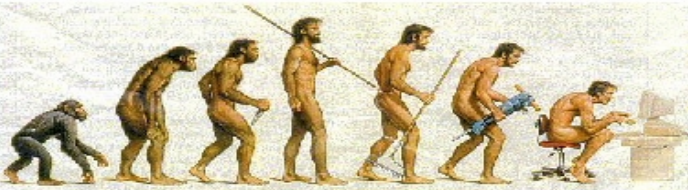
Layers Involved

OSI
REFERENCE MODEL
LAYERS

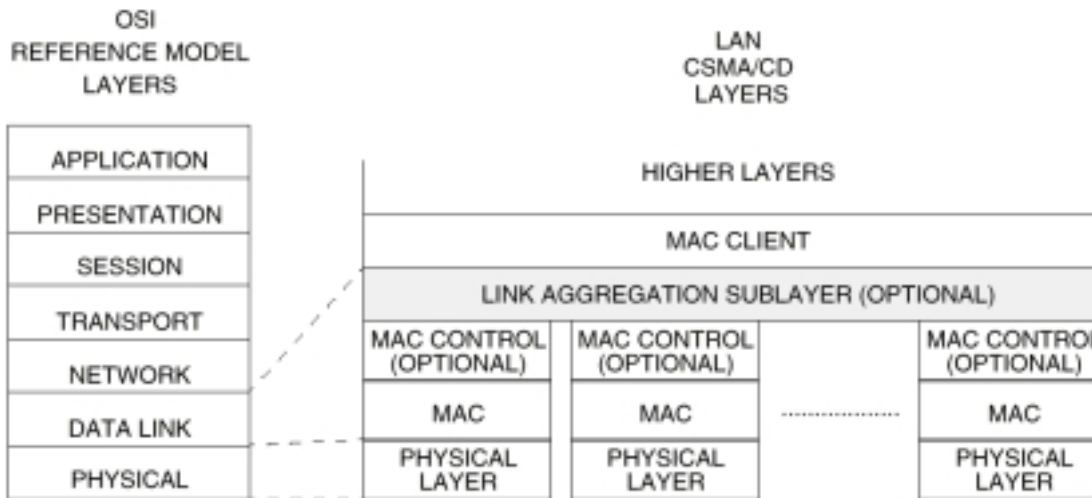
LAN
CSMA/CD
LAYERS



- IEEE Standard - adds to 802.3 CSMA/CD Standard
- Works with 10/100/1000 - portable to other MACs



Layers Involved (cont.)



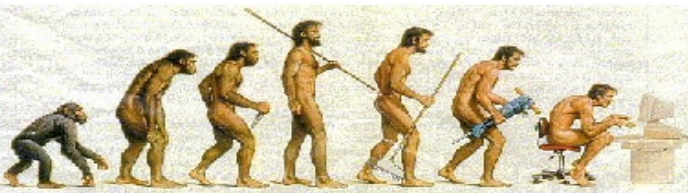
- MAC Client is Bridge layer, LLC, etc.
- MAC Control (optional) is for Flow control (802.3x)
- MAC is CSMA/CD, CRC32, Frame Encapsulation

This means Inter-Packet Gap and encapsulation is enforced by the individual link MACs (called physical MACs), but FCS and source MAC address can be sent from the single MAC above (e.g., if relayed)

No CSMA/CD since full-duplex required for link aggregation

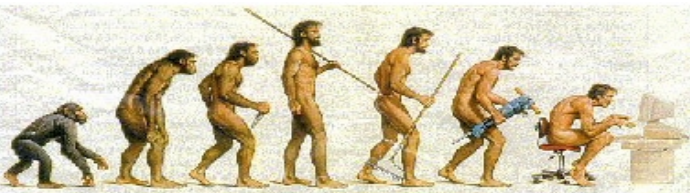
Notice the split happens at the traditional software layer, so changes need only be done in software (usually)

Flow control operates independent of the aggregation - it could pause link aggregation control frames (this is good, me thinks)



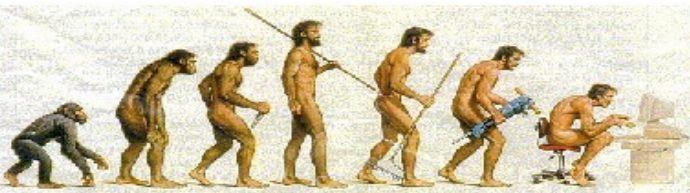
The diagram illustrates a network topology with three LANs and three bridges. The top LAN is connected to two bridges. The middle LAN is connected to one bridge. The bottom LAN is connected to one bridge. A red 'X' is placed over the connection between the top LAN and the middle bridge, indicating a blocked path. A red circle with a diagonal line is placed over the connection between the middle LAN and the bottom bridge, indicating a blocked path. The text 'Spanning Tree' is written in the top left corner.

- So Spanning Tree would disable the link - thus all physical links disabled
This is a bad example because it's impossible for a bridge to be trunk linked to a shared medium (gotcha!)

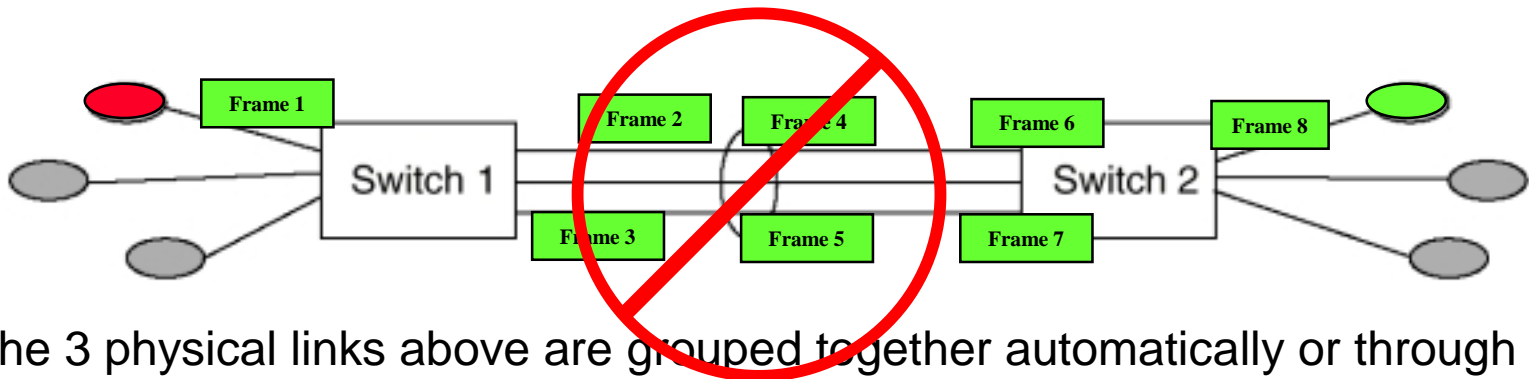


Quick Overview

- Ports constantly send out packets
- When one or more cables are linked, they auto-configure an aggregate group link
- (yes, I do mean one or more)
- Once aggregated, they still send each other packets as a keep-alive



How it works (or doesn't)



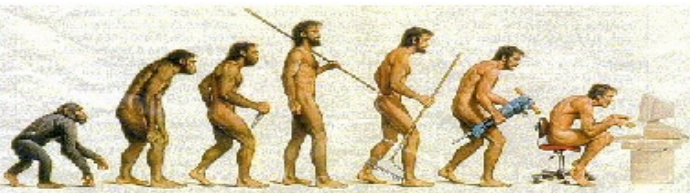
- The 3 physical links above are grouped together automatically or through configuration to be one logical link
- The green user on the right sends frames to the red user

Each frame is NOT broken up/spread across the wires (we're still stuck with a normal MAC, after all)

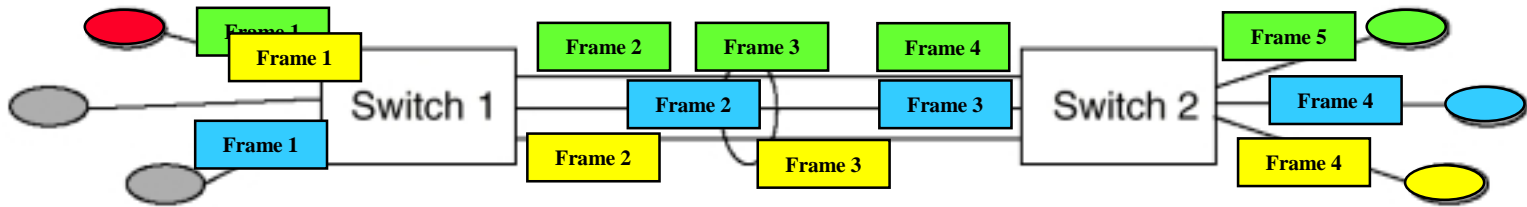
Instead, they are each sent whole - on one wire.

Also, we CANNOT send one frame on one wire, another on a second wire, etc. (the way shown above) Why not?

Because they could arrive out of order - that would be BAD.

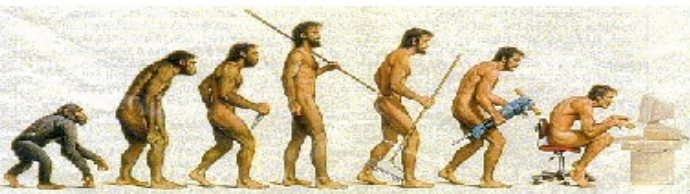


How it really works



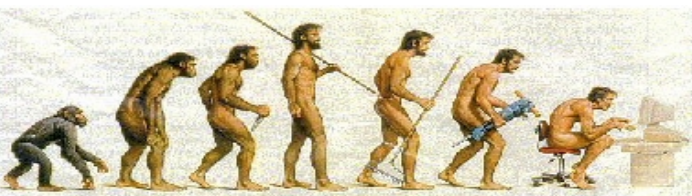
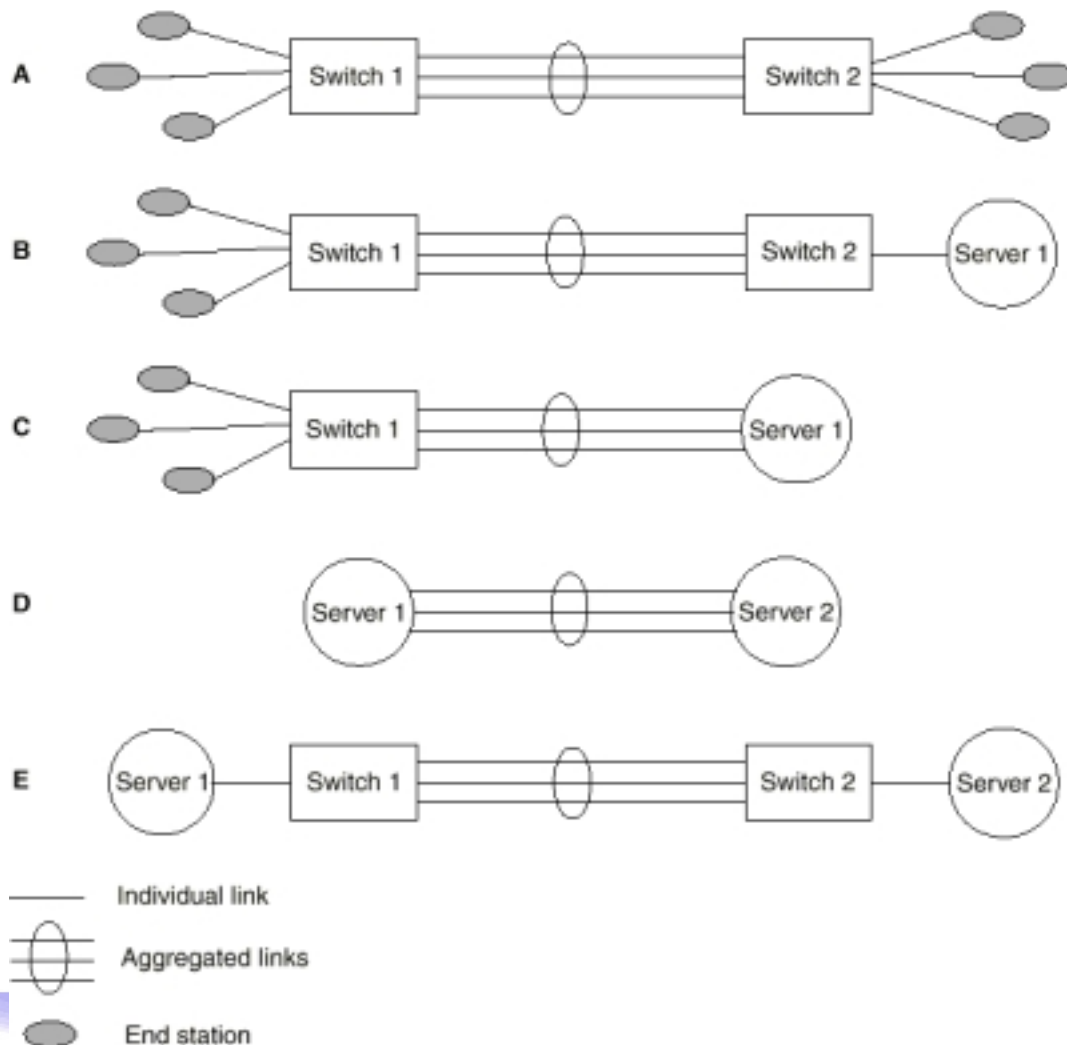
- So to protect against out of order delivery, the standard requires that the transmitting side (Switch 2) keep “conversations” together on one wire.
- A conversation is a one-way sequence of frames with same dest/src address for the same application. (i.e., anything that would be hurt by mis-ordering)
- In the case above, Switch 2 could keep all frames from station green to any other station on the same wire. (src. addr. based) So even if green sends frames to someone else on the left, it would still go through the same wire.

Likewise, any frames from other stations could be kept together on a separate wire. For example if blue and yellow sent frames, each of their streams could be on a separate wire. This is called a distribution algorithm, and is totally up to Switch 2. The standard only requires that a conversation not be misordered.



Reference Topologies

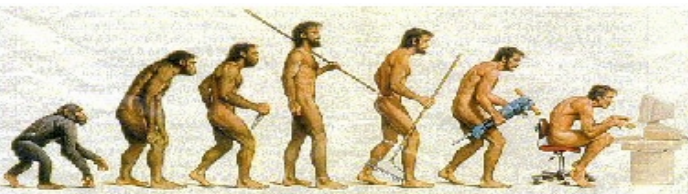
- Notice that repeaters cannot participate
- Each trunk link is one MAC layer to one MAC layer.
- All the trunks are MAC-to-MAC, but network A is many MAC addresses to many others, whereas B and C are many to one, and D and E are one to one. Thus the distribution algorithms will be different for them: based on src. addr., src & dest, src & dest & type, etc.



LACPDU

- Link Aggregation Control (LAC) entities send each other configuration info using LACPDUs (LAC Protocol Data Units)
- The destination address of the LACPDU is a well-known multicast address
- If the link partner doesn't support LA, then it discards them (doesn't forward)
- Repeaters forward them like normal frames (remember, repeaters are layer 1 - they don't even know what a "frame" is!) - BUT if there's a repeater, then the link should be half-duplex and thus non-aggregatable (is that a word?)

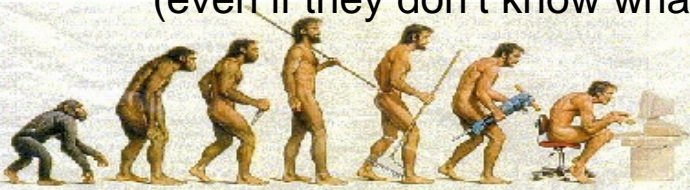
	Octets
Destination Address	6
Source Address	6
Length/Type	2
Subtype = LACP	1
Version Number	1
TLV_type = Actor Information	1
Length = 16	1
Actor_Port_Priority	1
Actor_Port	2
Actor_System_Priority	2
Actor_System	6
Actor_Key	2
Actor_State	1
TLV_type = Partner Information	1
Length = 16	1
Partner_Port_Priority	1
Partner_Port	2
Partner_System_Priority	2
Partner_System	6
Partner_Key	2
Partner_State	1
TLV_type = Terminator	1
Length = 0	1
Reserved	74
FCS	4



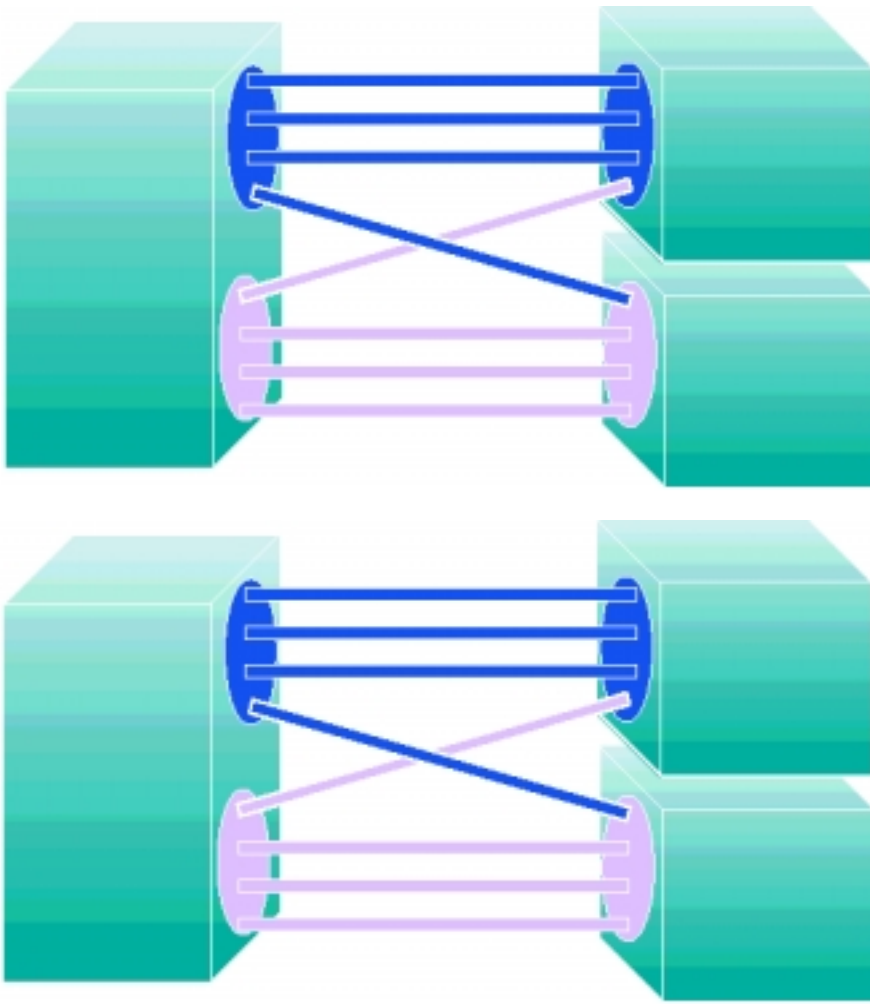
Destination Address

Assignment	Value
Bridge Group Address	01-80-C2-00-00-00
Reserved for future standardization	01-80-C2-00-00-01
Reserved for future standardization	01-80-C2-00-00-02
Reserved for future standardization	01-80-C2-00-00-03
Reserved for future standardization	01-80-C2-00-00-04
Reserved for future standardization	01-80-C2-00-00-05
Reserved for future standardization	01-80-C2-00-00-06
Reserved for future standardization	01-80-C2-00-00-07
Reserved for future standardization	01-80-C2-00-00-08
Reserved for future standardization	01-80-C2-00-00-09
Reserved for future standardization	01-80-C2-00-00-0A
Reserved for future standardization	01-80-C2-00-00-0B
Reserved for future standardization	01-80-C2-00-00-0C
Reserved for future standardization	01-80-C2-00-00-0D
Reserved for future standardization	01-80-C2-00-00-0E
Reserved for future standardization	01-80-C2-00-00-0F

- LACPDU Destination Address is a multicast of the type not forwarded by bridges (even if they don't know what it's for, they won't forward it)



Some Background on Selection



- Wanted auto-config to avoid user mistakes, as well as be flexible
- For instance, when the user crosses wires from 2 systems (shown left)
- By using a System ID Key, this can be detected and fixed



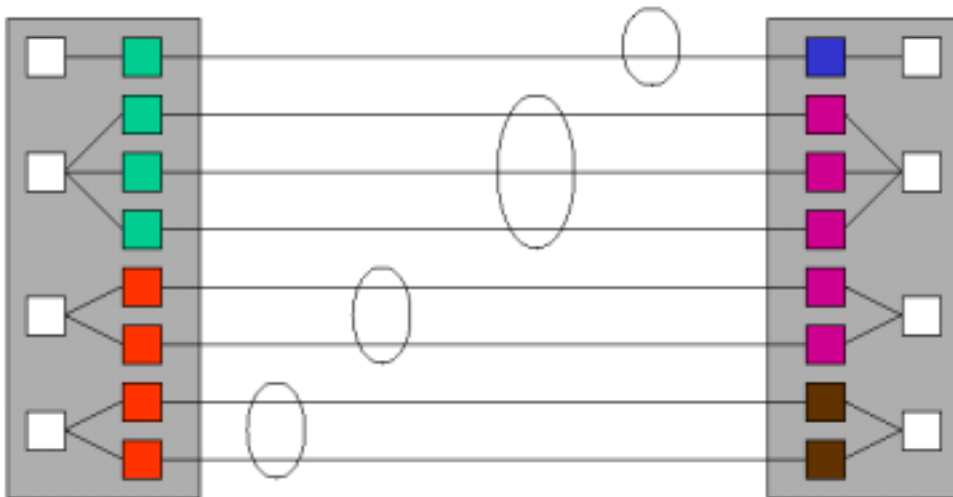
More Goals



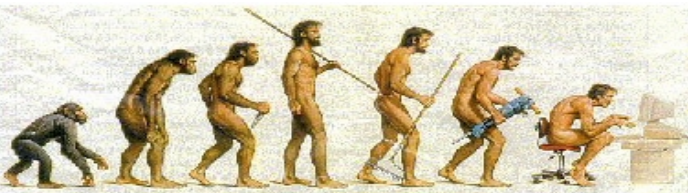
- Can also detect loopbacks and disable ports
- Plus need Port ID Keys to detect which ports can be aggregated together

System ID = A

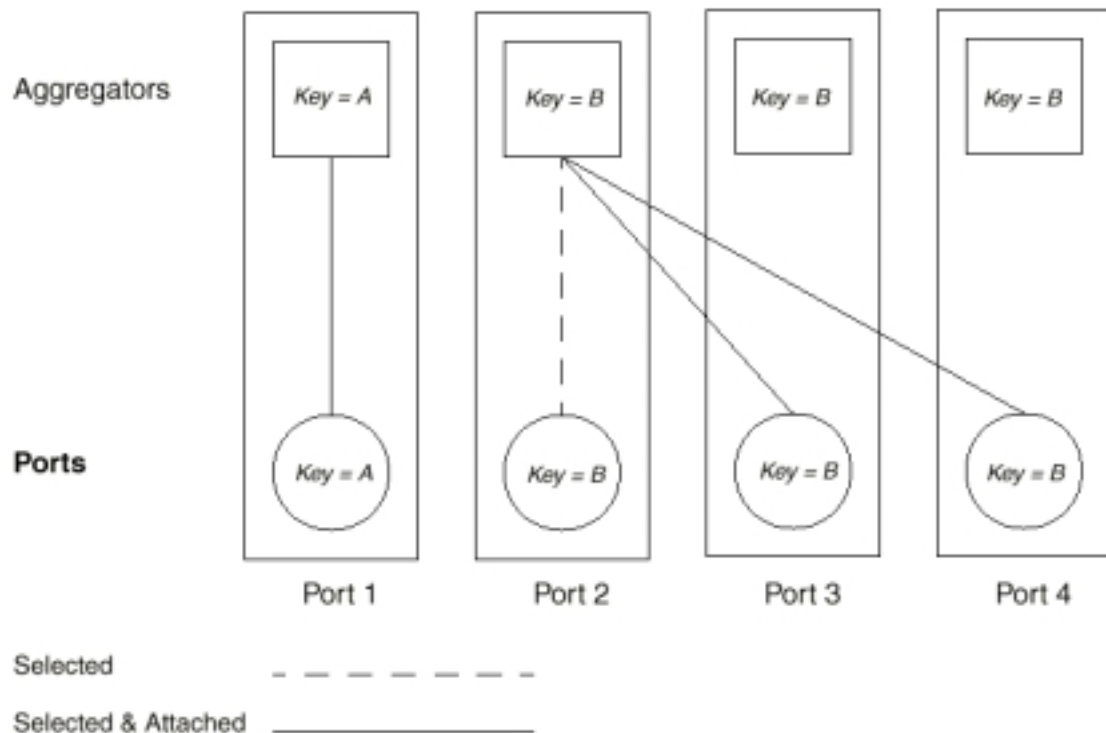
System ID = B



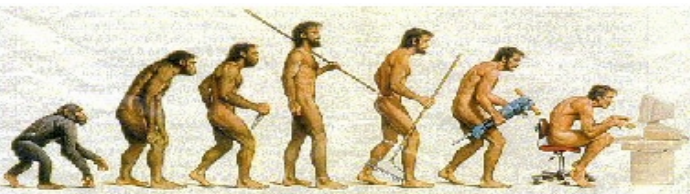
So for two systems A and B with ports 1 and 2, we call it {A1,B2}



One more set of Keys

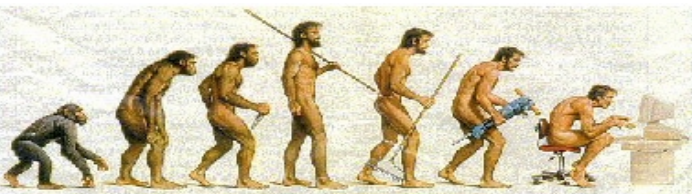


- For manual config, and alternative selection algorithms, the Aggregators need a Key
- This way the Agg Group MAC address can be assigned to the ports
- Port selects Agg with same Key as itself
- This Key is internal - not transmitted - only port key and sys ID transmitted



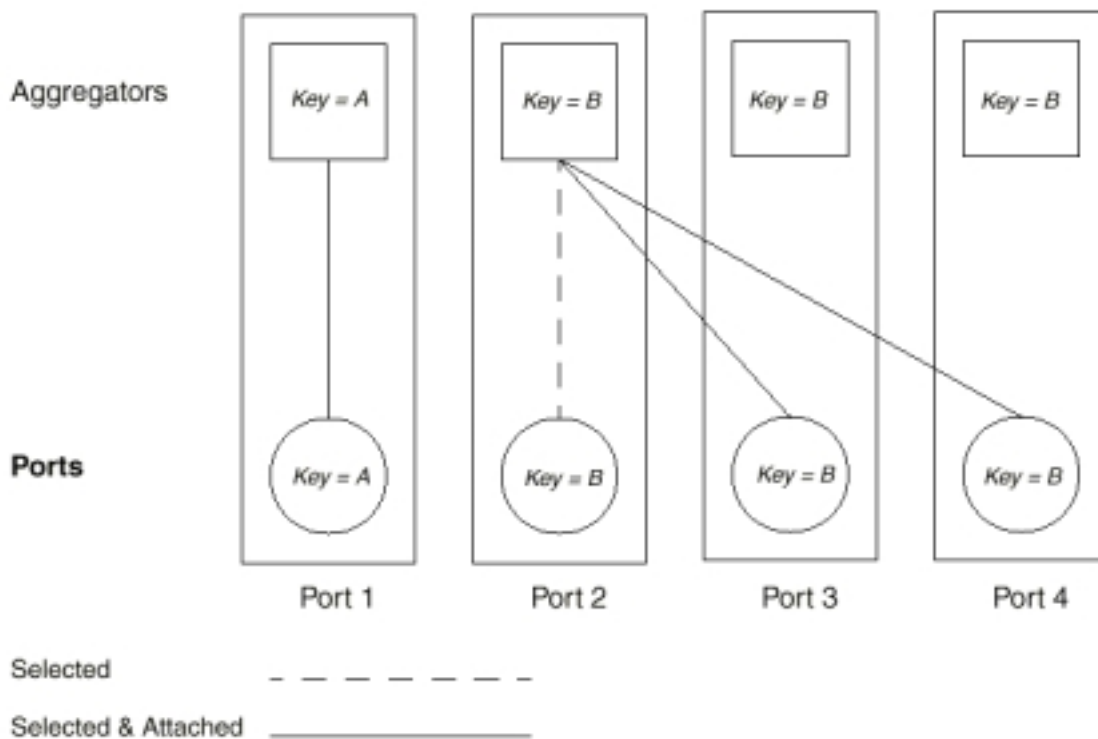
Selection Logic

- IEEE wanted differing implementations of selection, as long as the following are kept:
 - Each port assigned a Key - ports that can agg together have the same key - else unique Key
 - Each Aggregator assigned a Key and MAC address
 - Ports only select Aggs with same Key
 - Ports in the same LAG use the same Aggregator
 - Individual ports each have their own Agg
 - If a port can't select an Agg due to rules above, then it can't be attached to one
 - MAC Client won't see the port until it's selected and attached
- That's it. This leaves a lot open.

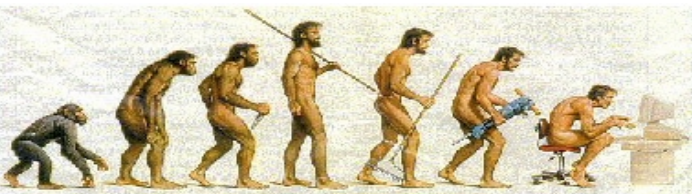


Default Selection Operation

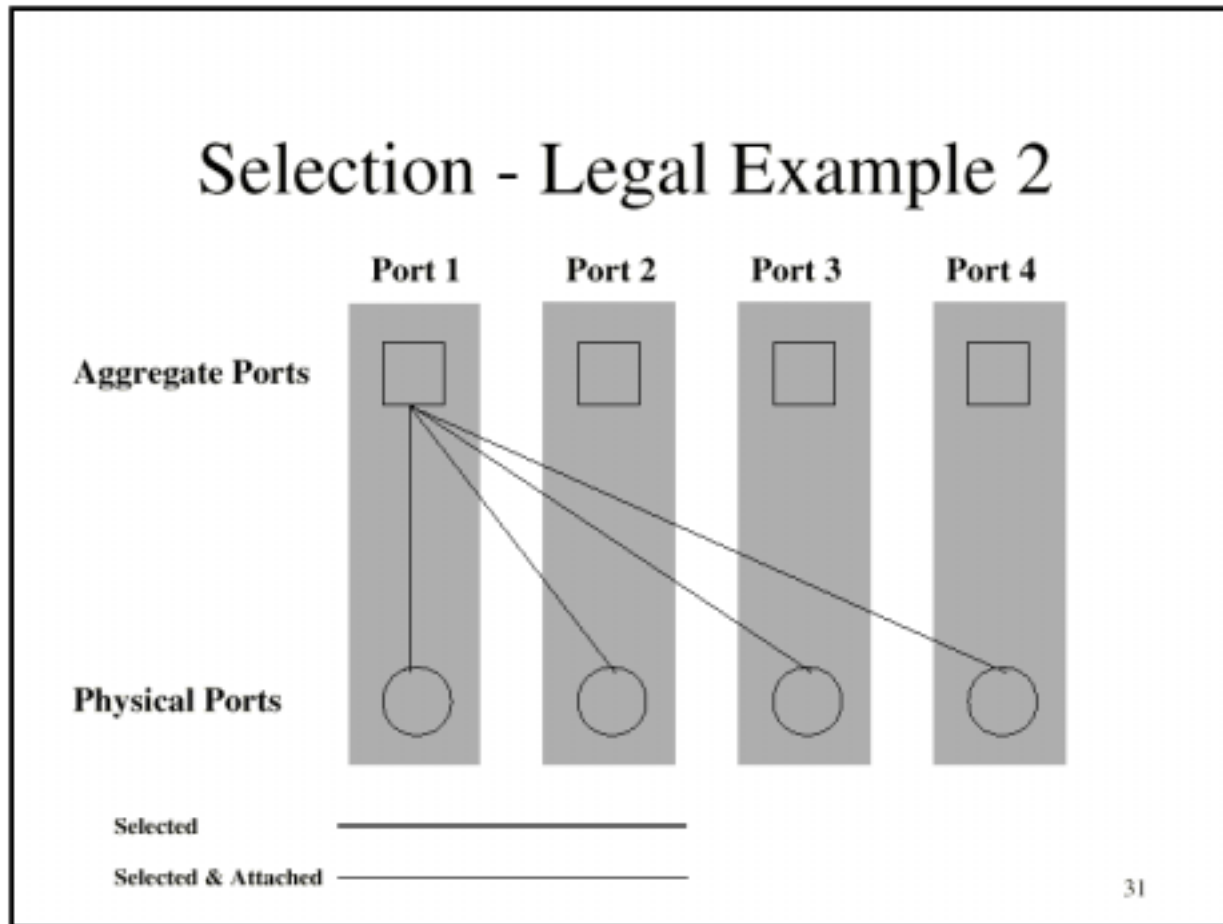
- When multiple ports in an aggregation, select the lowest number Aggregator of the ports in the aggregation
- That port may be just selected but not attached



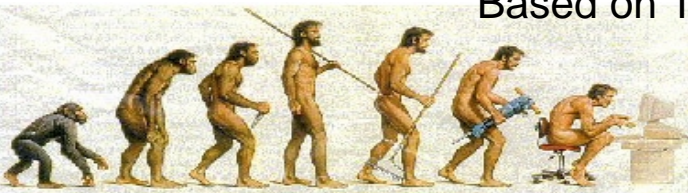
This algorithm has a flaw - can you see it?



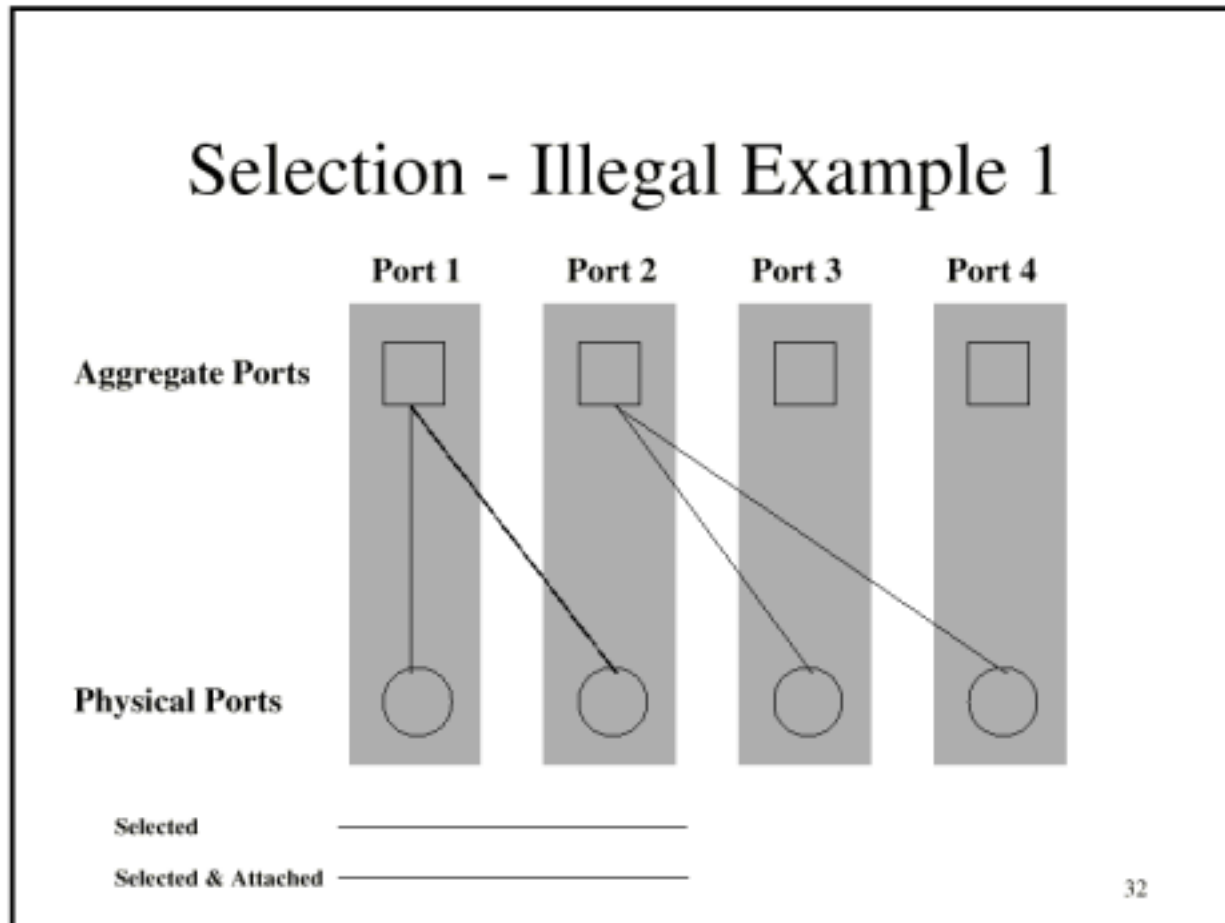
Default Examples



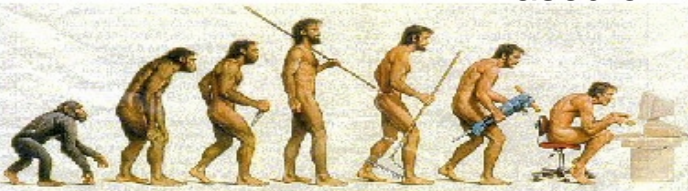
Based on Tony Jeffree's presentation to IEEE, 11/98



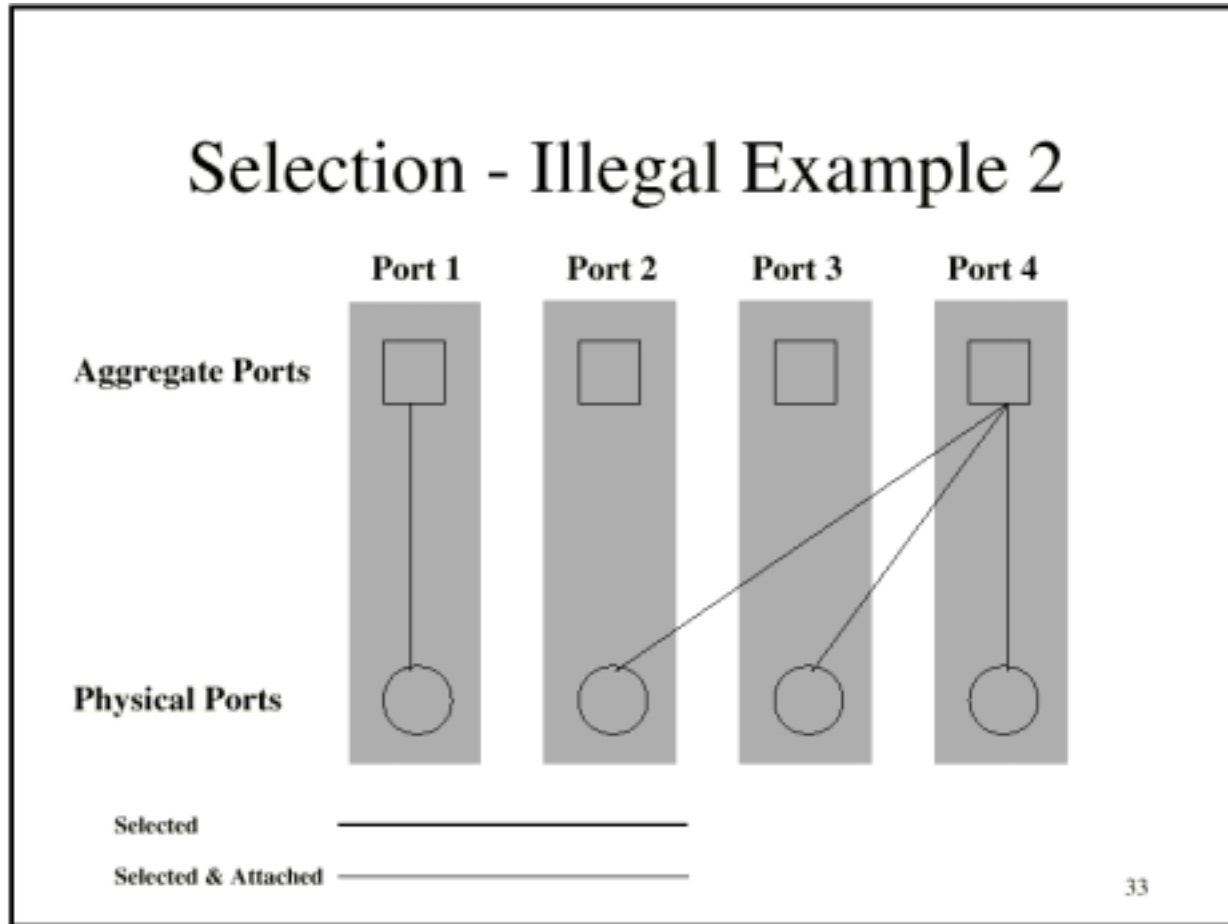
Illegal Examples



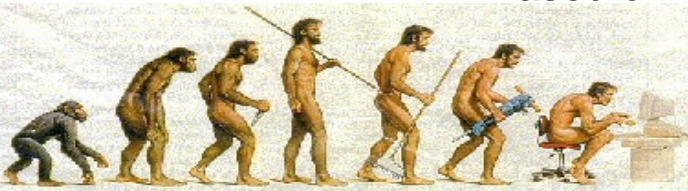
Based on Tony Jeffree's presentation to IEEE, 11/98



More Illegal Examples

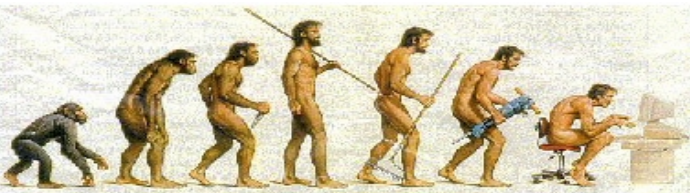


Based on Tony Jeffree's presentation to IEEE, 11/98



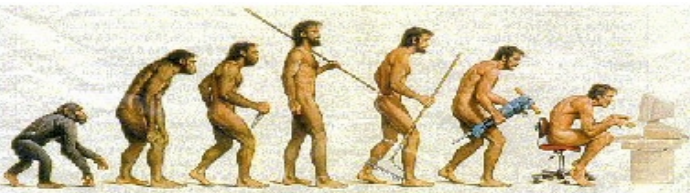
Problems

- Not necessarily deterministic
- If deterministic, then a failover can take the net down for a LONG time (due to spanning tree)
- A few too many options for users
- Doesn't support full-duplex repeaters (yes, there is a such a thing)



Multiple Spanning Trees

802.1s



25
0

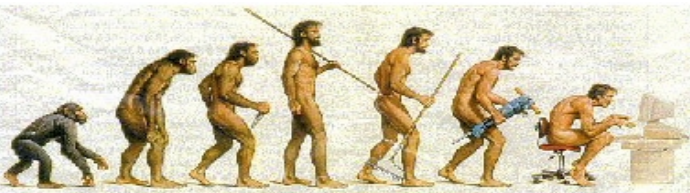
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

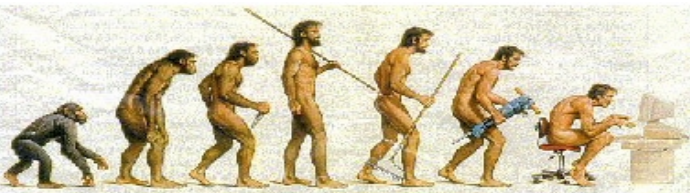
Multiple Spanning Trees

- If multiple VLANs are run on a network, they all run on the same spanning tree
- This means unused (blocked) links don't carry any traffic
- But traffic from separate VLANs can be sent on different spanning trees, thereby utilizing the blocked ports
- So now one can assign VLANs to separate spanning trees



Rapid Reconfiguration

802.1w



25
2

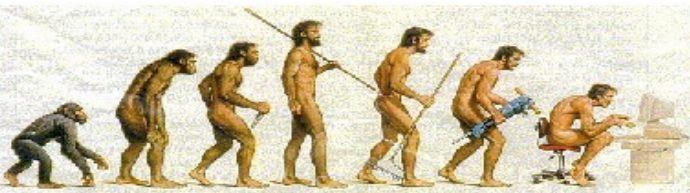
The Ethernet Evolution

10 to 10,000 Mbps

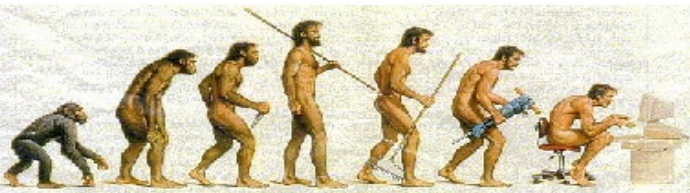
prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Rapid Reconfiguration

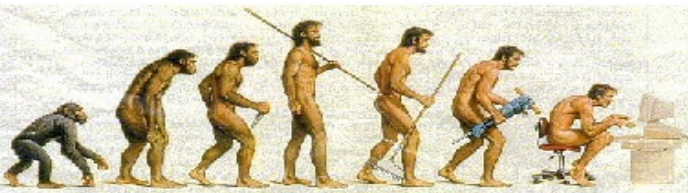
- A broken spanning tree can take 50 seconds to re-converge
- A new algorithm, using new BPDU frames, can decrease the time down to 10 milliseconds (best case)
- It's backwards compatible with legacy 802.1D bridges (they won't converge faster, but they will work)



Wrap-Up



The Future



25
5

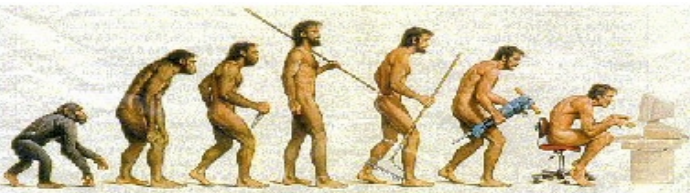
The Ethernet Evolution

10 to 10,000 Mbps

prepared by Hadriel Kaplan & Bob Noseworthy
September 25, 2000

Acknowledgements

- Some of the slides, drawings, and animations were ~~stolen~~ borrowed from:
 - Gerry Nadeau
 - Benjamin Schultz
 - Adam Healey
 - Rupert Dance
 - Gary Pressler
 - Eric Lynskey
 - The IEEE standards and website



Websites

- To get the most recent version of this presentation, go to:

<http://www.iol.unh.edu/training/ethernet.html>

- IEEE website: <http://grouper.ieee.org/groups/802>
- IETF website: <http://www.ietf.org>
- Our email addresses:
 - hadriel@nortelnetworks.com
 - ren@iol.unh.edu

