



The Ethernet Evolution

The 180 Degree Turn

(C) Herbert Haas 2010/02/15



*“Use common sense in routing cable.
Avoid wrapping coax around sources
of strong electric or magnetic fields.
Do not wrap the cable around
flourescent light ballasts or
cyclotrons, for example.”*

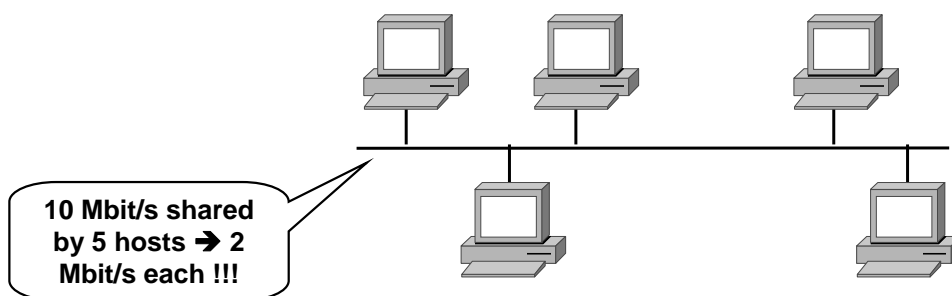


Ethernet Headstart Product Information and Installation Guide,
Bell Technologies, pg. 11

History: Initial Idea



- Shared media → CSMA/CD as access algorithm
- COAX Cables
- Half duplex communication
- Low latency → No networking nodes (except repeaters)
- One collision domain and also one broadcast domain



(C) Herbert Haas 2010/02/15

3

The initial idea of Ethernet was completely different than what is used today under the term "Ethernet". The original new concept of Ethernet was the use of a shared media and an Aloha based access algorithm, called Carrier Sense Multiple Access with Collision Detection (CSMA/CD). Coaxial cables were used as shared medium, allowing a simple coupling of station to bus-like topology.

Coax-cables were used in baseband mode, thus allowing only unicast transmissions. Therefore, CSMA/CD was used to let Ethernet operate under the events of frequent collisions.

Another important point: No intermediate network devices should be used in order to keep latency as small as possible. Soon repeaters were invented to be the only exception for a while.

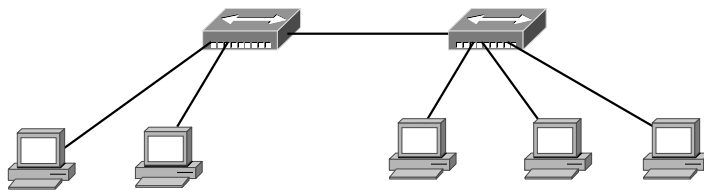
An Ethernet segment is a coax cable, probably extended by repeaters. The segment constitutes one collision domain (only one station may send at the same time) and one broadcast domain (any station receives the current frame sent). Therefore, the total bandwidth is shared by the number of devices attached to the segment. For example 10 devices attached means that each device can send 1 Mbit/s of data on average.

Ethernet technologies at that time (1975-80s): 10Base2 and 10Base5

History: Multiport Repeaters



- **Demand for structured cabling (voice-grade twisted-pair)**
 - ♦ 10BaseT (Cat3, Cat4, ...)
- **Multiport repeater ("Hub") created**
- **Still one collision domain ("CSMA/CD in a box")**



(C) Herbert Haas 2010/02/15

4

Later, Ethernet devices supporting structured cabling were created in order to reuse the voice-grade twisted-pair cables already installed in buildings. 10BaseT had been specified to support Cat3 cables (voice grade) or better, for example Cat4 (and today Cat5, Cat6, and Cat7).

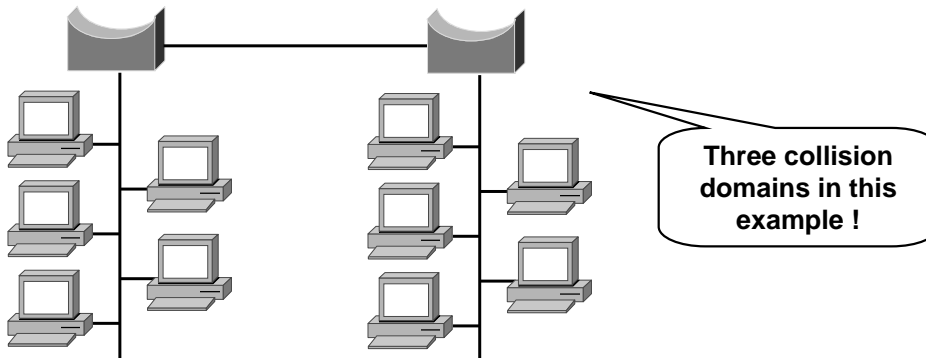
Hub devices were necessary to interconnect several stations. These hub devices were basically multi-port repeaters, simulating the half-duplex coax-cable, which is known as "CSMA/CD in a box". Logically, nothing has changed, we have still one single collision and broadcast domain.

Note that the Ethernet topology became star-shaped.

History: Bridges



- Store and forwarding according destination MAC address
- Separated collision domains
- Improved network performance
- Still one broadcast domain



(C) Herbert Haas 2010/02/15

5

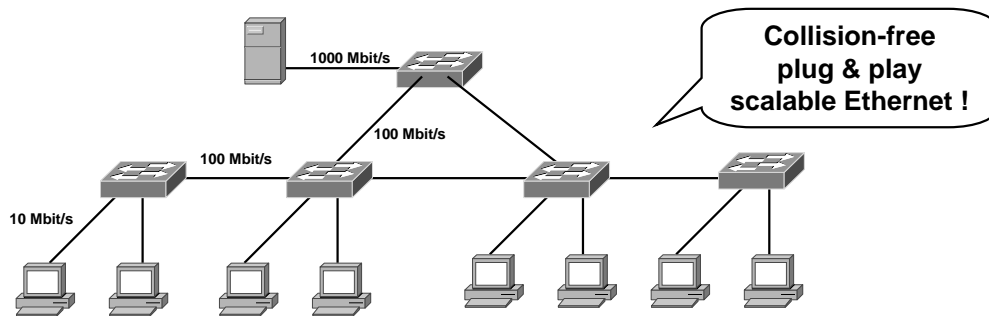
Bridges were invented for performance reasons. It seemed to be impractical that each additional station reduces the average per-station bandwidth by $1/n$. On the other hand the benefit of sharing a medium for communication should be still maintained (which was expressed by Metcalfe's law).

Bridges are store and forwarding devices (introducing significant delay) that can filter traffic based on the destination MAC addresses to avoid unnecessary flooding of frames to certain segments. Thus, bridges segment the LAN into several collision domains. Broadcasts are still forwarded to allow layer 3 connectivity (ARP etc), so the bridged network is still a single broadcast domain.

History: Switches



- **Switch = Multiport Bridges with HW acceleration**
- **Full duplex → Collision-free Ethernet → No CSMA/CD necessary anymore**
- **Different data rates at the same time supported**
 - ♦ Autonegotiation
- **VLAN splits LAN into several broadcast domains**



(C) Herbert Haas 2010/02/15

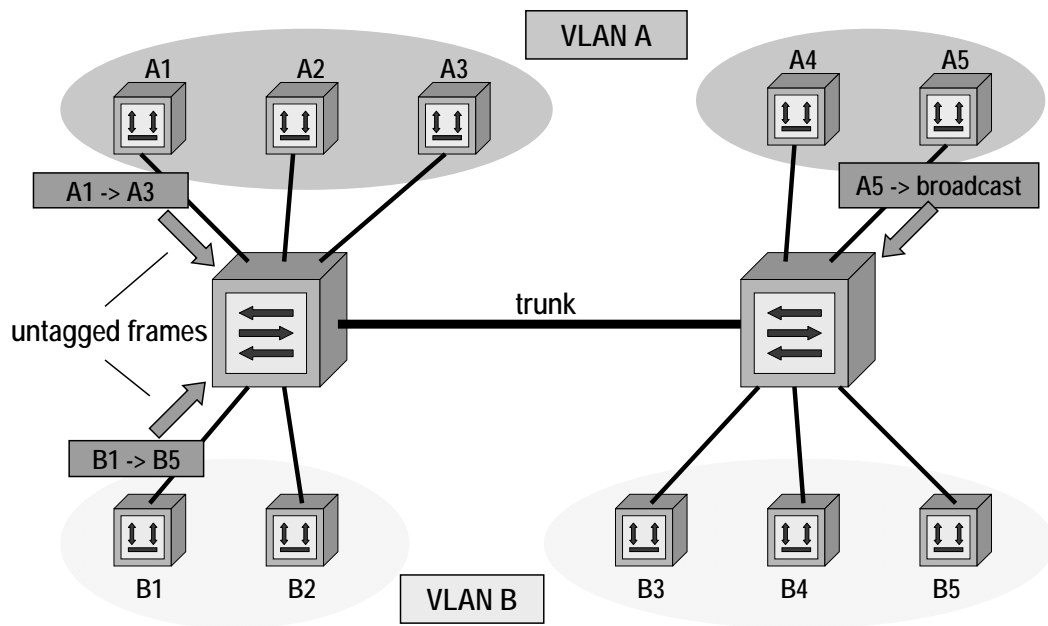
6

Several vendors built advanced bridges, which are partly or fully implemented in hardware. The introduced latency could be dramatically lowered and furthermore other features were introduced, for example full duplex communication on twisted pair cables, different frame rates on different ports, special forwarding techniques (e.g. cut through or fragment free), Content Addressable Memory (CAM) tables, and much more. Of course marketing rules demand for another designation for this machine: the switch was born.

Suddenly, a collision free plug and play Ethernet was available. Simply use twisted pair cabling only and enable autonegotiation to automatically determine the line speed on each port (of course manual configurations would also do). This way, switched Ethernet become very scalable.

Furthermore, Virtual LANs (VLANs) were invented to split the LAN into several broadcast domains. VLANs improve security, utilization, and allows for logical borders between workgroups.

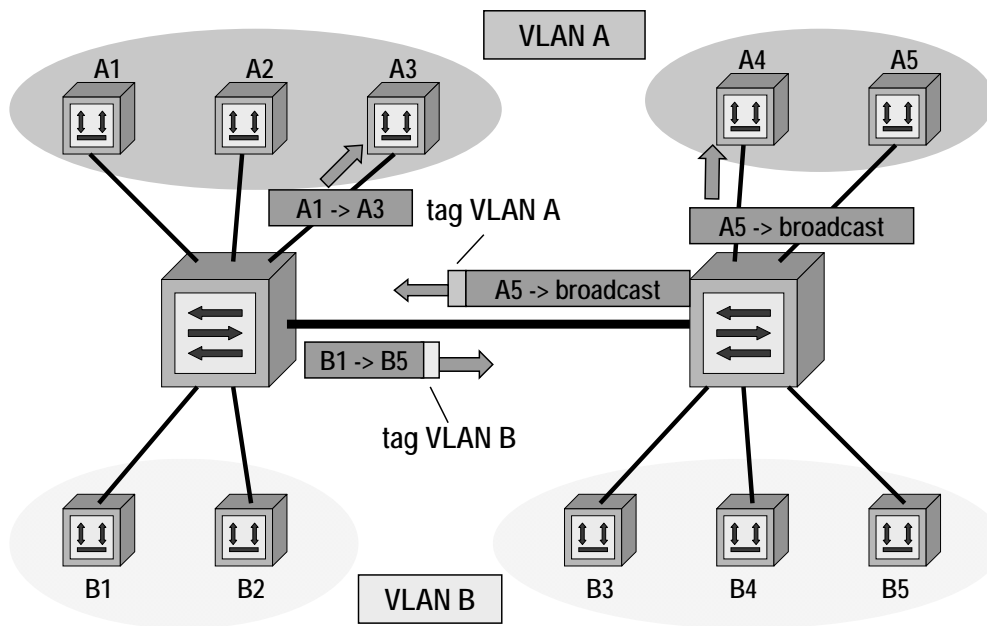
VLAN Operation (1)



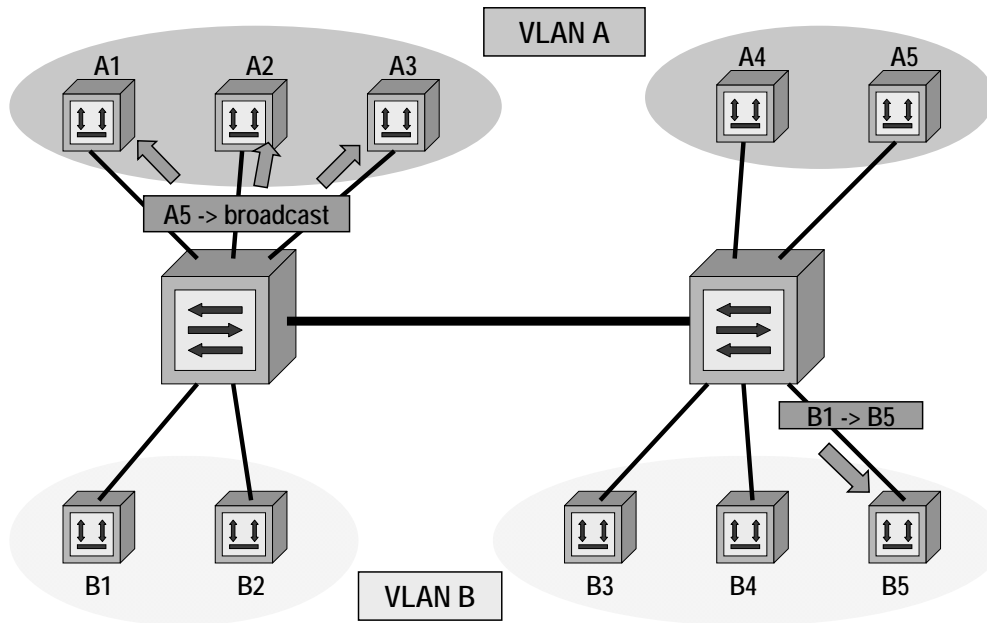
2010/02/15

7

VLAN Operation (2)



VLAN Operation (3)



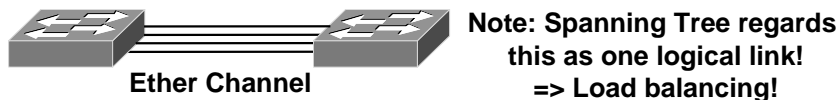
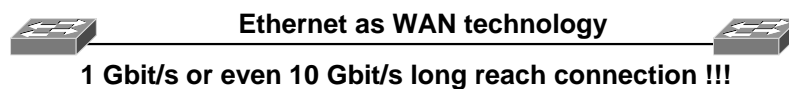
2010/02/15

9

Today



- **No collisions → no distance limitations !**
- **Gigabit Ethernet becomes WAN technology !**
 - ♦ Over 100 km link span already
- **Combine several links to "Etherchannels"**
 - ♦ Link Aggregation Control Protocol (LACP, IEEE 802.3ad)
 - ♦ Cisco proprietary: Port Aggregation Protocol (PAgP)
 - ♦ HP: Mesh (like L2-routing over 5-8 hops)



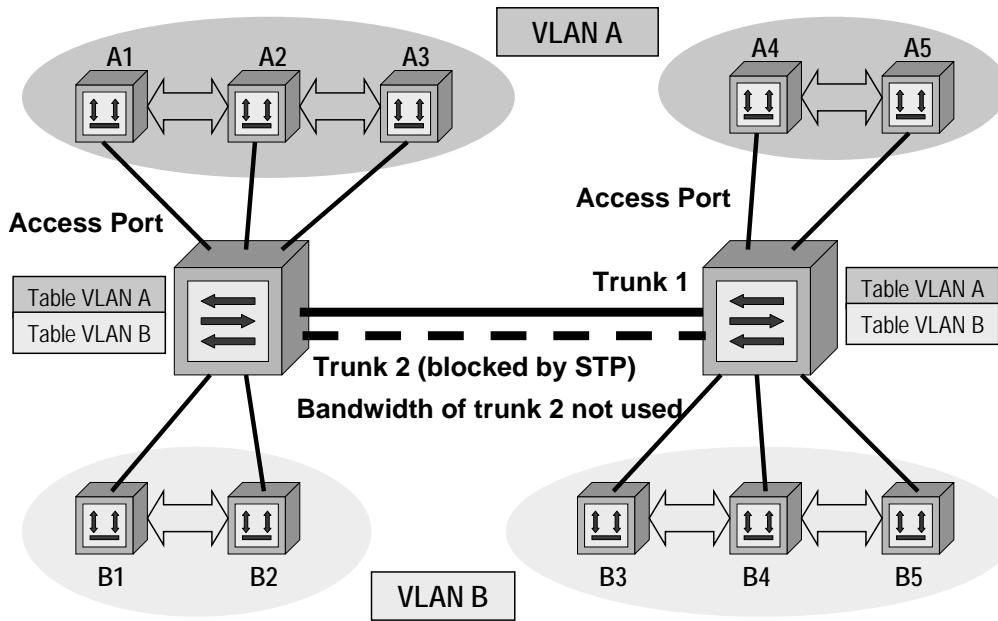
Today, Gigabit and even 10 Gigabit Ethernet is available. Only twisted pair and more and more fiber cables are used between switches, allowing full duplex collision-free connections. Since collisions cannot occur anymore, there is no need for a collision window anymore! From this it follows, that there is virtually no distance limit between each two Ethernet devices.

Recent experiments demonstrated the interconnection of two Ethernet Switches over a span of more than 100 km! Thus Ethernet became a WAN technology! Today, many carriers use Ethernet instead of ATM/SONET/SDH or other rather expensive technologies. GE and 10GE is relatively cheap and much simpler to deploy. Furthermore it easily integrates into existing low-rate Ethernet environments, allowing a homogeneous interconnection between multiple Ethernet LAN sites. Basically, the deployment is plug and play.

If the link speed is still too slow, so-called "Etherchannels" can be configured between each two switches by combining several ports to one logical connection. Note that it is not possible to deploy parallel connections between two switches without an Etherchannel configuration because the Spanning Tree Protocol (STP) would cut off all redundant links.

Depending on the vendor, up to eight ports can be combined to constitute one "Etherchannel".

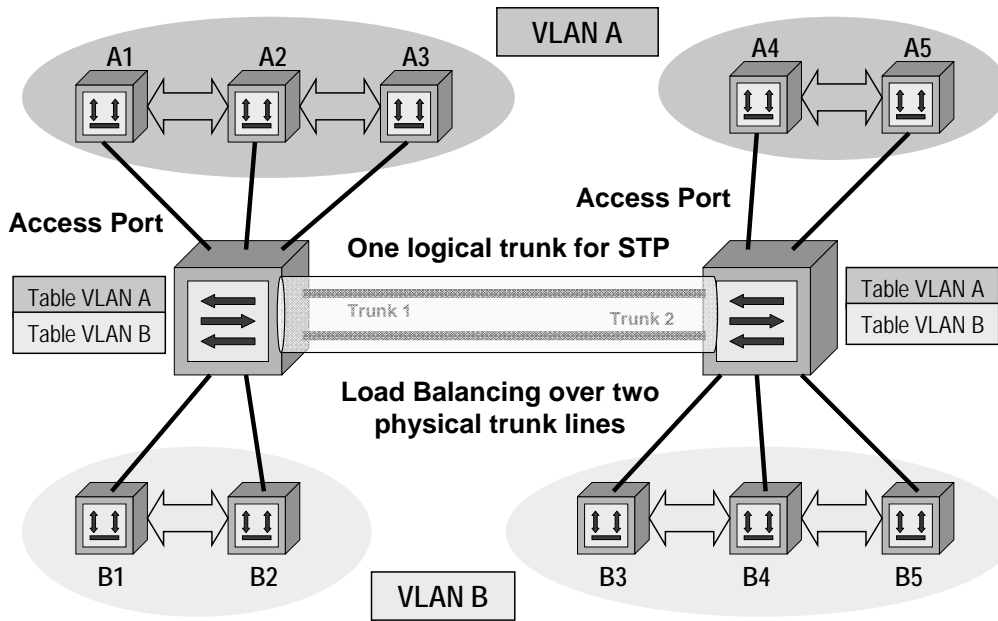
Trunking without LACP / FEC / GEC



2010/02/15

11

Trunking with LACP / FEC / GEC



2010/02/15

12

What About Gigabit Hubs?



- **Would limit network diameter to 20-25 meters (Gigabit Ethernet)**
- **Solutions**
 - ◆ **Frame Bursting**
 - ◆ **Carrier Extension**
- **No GE-Hubs available on the market today → forget it!**
- **No CSMA/CD defined for 10GE (!)**

Remember: Hubs simulate a half-duplex coaxial cable inside, hence limiting the total network diameter. For Gigabit Ethernet this limitation would be about 25 meters, which is rather impracticable for professional usage. Although some countermeasures had been specified in the standard, such as frame bursting and carrier extension, no vendor developed an GE hub as for today. Thus: Forget GE Hubs!

The 10 GE specification does neither consider copper connections nor hubs. 10 GE can only run over fiber.

At this point please remember the initial idea in the mid 1970s: Bus, CSMA/CD, short distances, no network nodes.

Today: Structured cabling (point-to-point or star), never CSMA/CD, WAN capabilities, sophisticated switching devices in between.

CSMA/CD Restrictions (Half Duplex Mode)

- **Solutions to increase the maximal net expansion:**

- Carrier Extension:

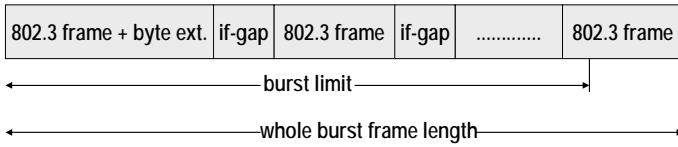
- extension bytes appended to (and removed from) the Ethernet frame by the physical layer
- frame exists a longer period of time on the medium

- Frame Bursting:

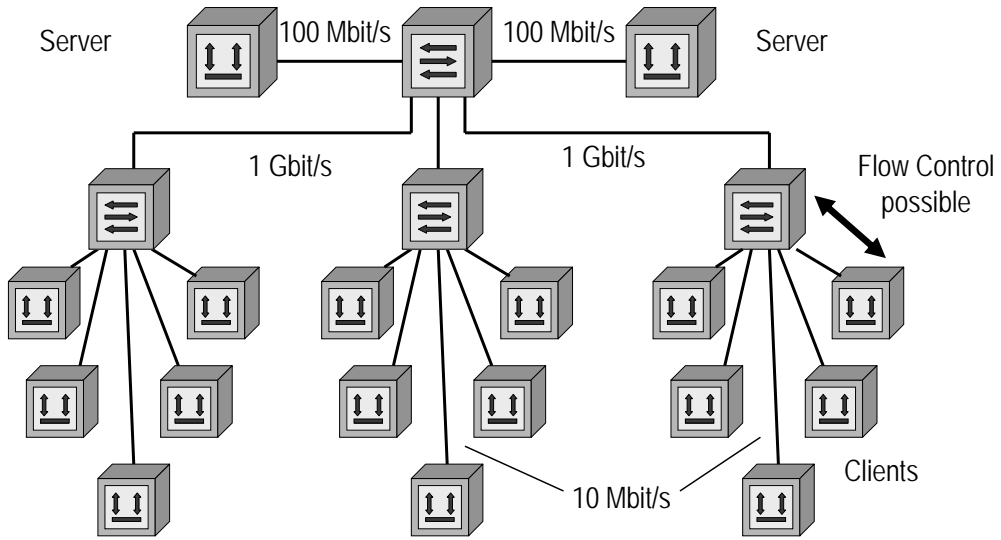
- to minimize the extension bytes overhead, station may chain several frames together and transmit them at once ("burst").

- **With both methods the minimal frame length is increased from 512 to 4096 bits**
 - = 512 bytes
 - The corresponding time is called slottime
- **If a station decides to chain several frames to a burst frame, the first frame inside the burst frame must have a length of at least 512 bytes**
 - By using extension bytes if necessary
- **The next frames (inside the burst frame) can have normal length (i.e. at least 64 bytes)**

- **Station may chain frames up to 8192 bytes (=burst limit)**
 - Also may finish the transmission of the last frame even beyond the burst limit
- **So the whole burst frame length must not exceed 8192+1518 bytes**
 - Incl. interframe gap of 0.096 μ s = 12 bytes



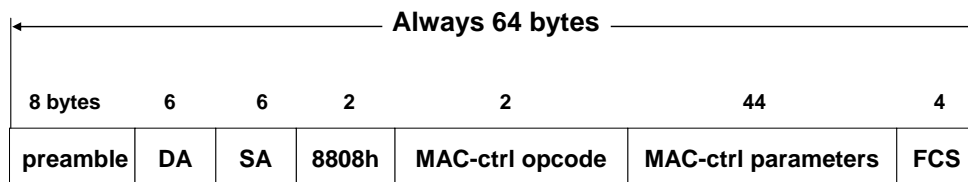
Ethernet Switching <-> Flow Control



MAC Control Frames



- **Additional functionality easily integrated**
- **Currently only Pause-Frame supported**



MAC-ctrl opcode Defines function of control frame

MAC-ctrl parameters control parameter data (always filled up to 44 bytes)

Different data rates between switches (and different performance levels) often lead to congestion conditions, full buffers, and frame drops. Traditional Ethernet flow control was only supported on half-duplex links by enforcing collisions to occur and hereby triggering the truncated exponential backoff algorithm. Just let a collision occur and the aggressive sender will be silent for a while.

A much finer method is to send some dummy frames just before the backoff timer allows sending. This way the other station never comes to send again.

Both methods are considered as ugly and only work on half duplex lines.

Therefore the MAC Control frames were specified, allowing for active flow control. Now the receiver sends this special frame, notifying the sender to be silent for N slot times.

The MAC Control frame originates in a new Ethernet layer—the MAC Control Layer—and will support also other functionalities, but currently only the "Pause" frame has been specified.

- **on receiving the pause command**
 - station stops sending normal frames for a given time which is specified in the MAC-control parameter field
- **this pause time is a multiple of the slot time**
 - 4096 bit-times when using Gigabit Ethernet or 512 bit-times with conventional 802.3
- **paused station waits**
 - until pause time expires or an additional MAC-control frame arrives with pause time = 0
 - note: paused stations are still allowed to send MAC-control-frames (to avoid blocking of LAN)

- **destination address is either**
 - address of destination station or
 - broadcast address or
 - special multicast address 01-80-C2-00-00-01
- **this special multicast address prevents bridges to transfer associated pause-frames to not concerned network segments**
- **hence flow-control (with pause commands) affects only the own segment**

Auto Negotiation



- **Enables each two Ethernet devices to exchange information about their capabilities**
 - ♦ **Signal rate, CSMA/CD, half- or full-duplex**
- **Using Link-Integrity-Test-Pulse-Sequence**
 - ♦ **Normal-Link-Pulse (NLP) technique is used in 10BaseT to check the link state (green LED)**
 - ♦ **10 Mbit/s LAN devices send every 16.8 ms a 100ns lasting NLP, no signal on the wire means disconnected**

Several Ethernet operating modes had been defined, which are incompatible to each other, including different data rates (10, 100, 1000 Mbit/s), half or full duplex operation, MAC control frames capabilities, etc.

Original Ethernet utilized so-called Normal Link Pulses (NLPs) to verify layer 2 connectivity. NLPs are single pulses which must be received periodically between regular frames. If NLPs are received, the green LED on the NIC is turned on.

Newer Ethernet cards realize auto negotiation by sending a sequence of NLPs, which is called a Fast Link Pulse (FLP) sequence.

Fast Link Pulses

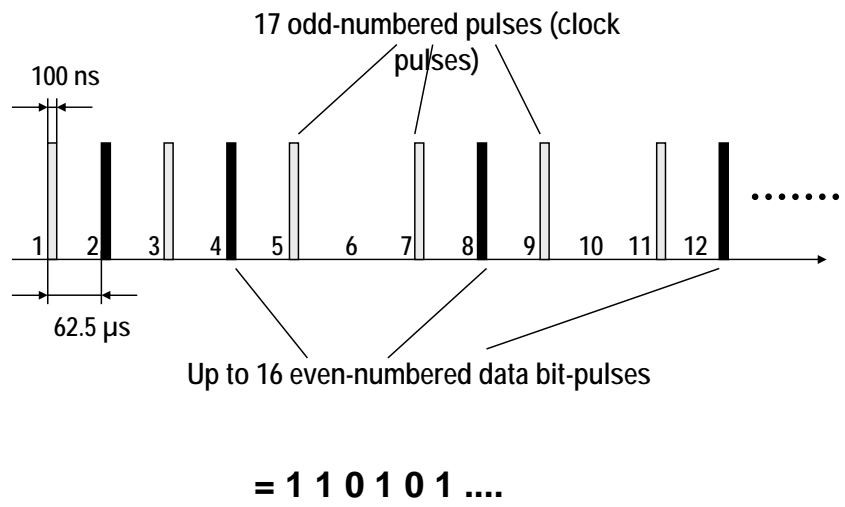


- **Modern Ethernet NICs send bursts of Fast-Link-Pulses (FLP) consisting of 17-33 NLPs for Autonegotiation signalling**
- **Each representing a 16 bit word**
 - ◆ **GE sends several "pages"**

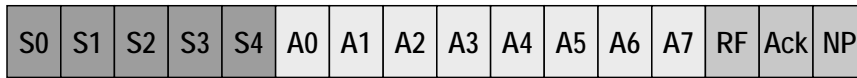
A series of FLPs constitute an autonegotiation frame. The whole frame consists of 33 timeslots, where each odd numbered timeslot consists of a real NLP and each even timeslot is either a NLP or empty, representing 1 or 0. Thus, each FLP sequence consists of a 16 bit word.

Note that GE Ethernet sends several such "pages".

FLP Burst Coding



Base Page



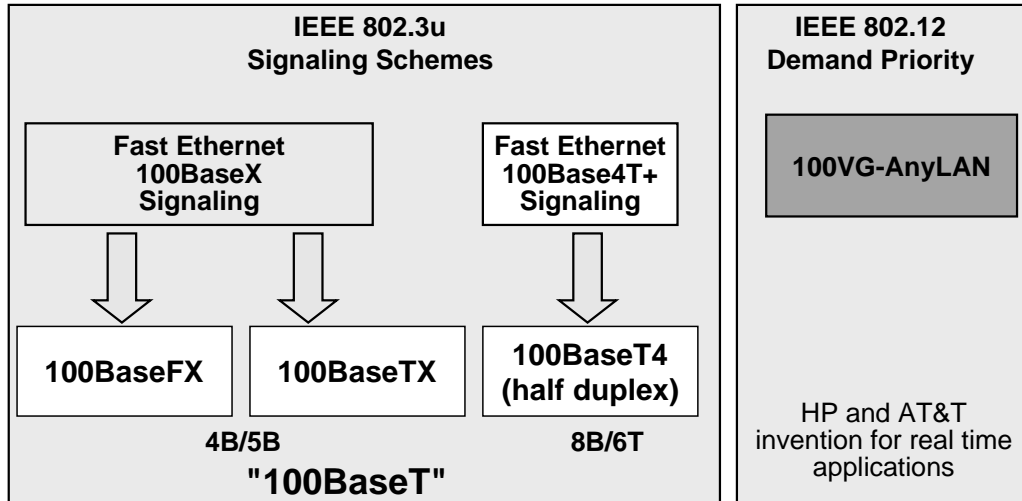
Selector field

Technology ability field

provides selection of up to 32 different message types; currently only 2 selector codes available:
 10000....IEEE 802.3
 01000....IEEE 802.9 (ISLAN-16T) (ISO-Ethernet)

Bit	Technology
A0	10BaseT
A1	10BaseT-full duplex
A2	100BaseTx
A3	100BaseTx-full duplex
A4	100BaseT4
A5	Pause operation for full duplex links
A6	reserved
A7	reserved

100 Mbit Ethernet Overview



(C) Herbert Haas 2010/02/15

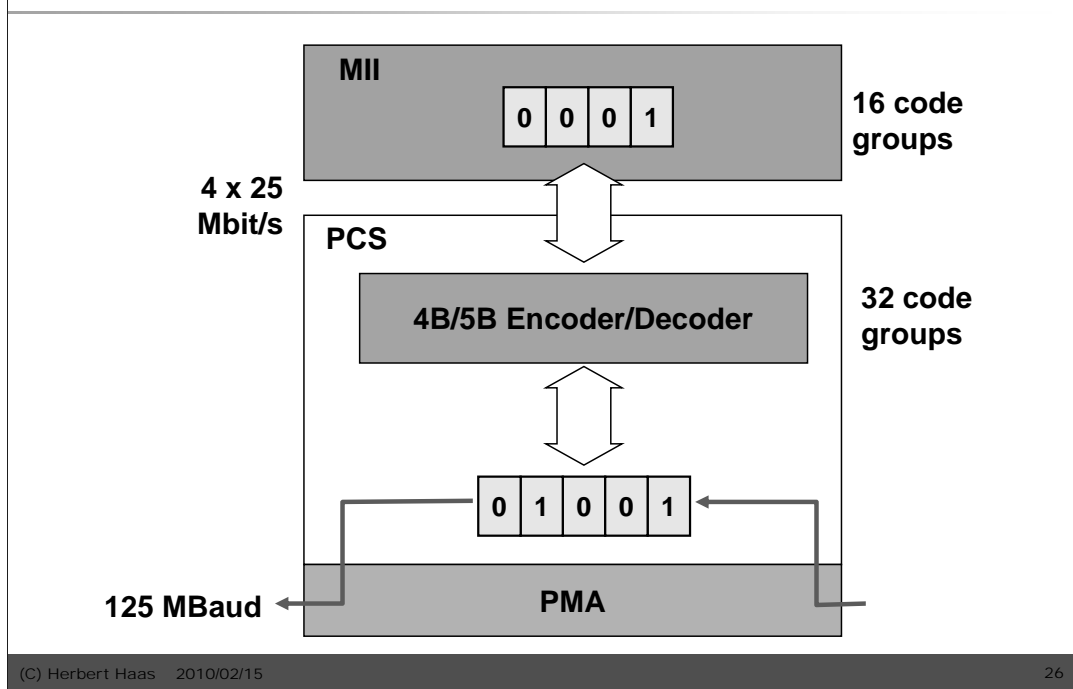
25

The diagram above gives an overview of 100 Mbit/s Ethernet technologies, which are differentiated into IEEE 802.3u and IEEE 802.12 standards. The IEEE 802.3u defines the widely used Fast Ethernet variants, most importantly those utilizing the 100BaseX signaling scheme. The 100BaseX signaling consists of several details, but basically it utilizes 4B5B block coding over only two pairs of regular Cat 5 twisted pair cables or two strand 50/125 or 62.5/125- μ m multimode fiber-optic cables.

100Base4T+ signaling has been specified to support 100 Mbit/s over Cat3 cables. This mode allows half duplex operation only and uses a 8B6T code over 4 pairs of wires; one pair for collision detection, three pairs for data transmission. One unidirectional pair is used for sending only and two bi-directional pairs for both sending and receiving.

The 100VG-AnyLAN technology had been created by HP and AT&T in 1992 to support deterministic medium access for realtime applications. This technology was standardized by the IEEE 802.12 working group. The access method is called "demand priority". 100VG-AnyLAN supports voice grade cables (VG) but requires special hub hardware. The 802.12 working group is no longer active.

4B/5B Coding



(C) Herbert Haas 2010/02/15

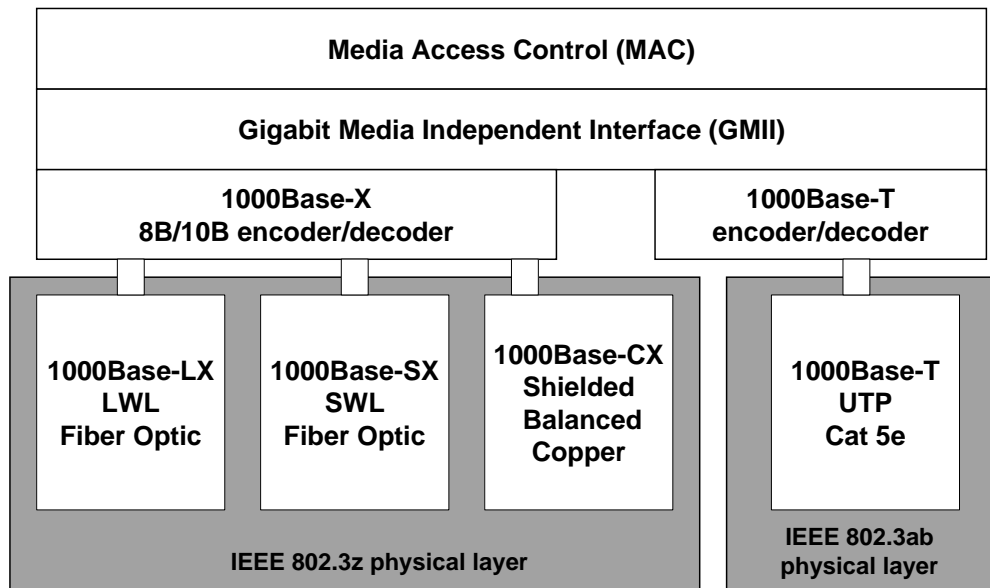
26

The diagram above shows the basic principle of the 4B5B block coding principle, which is used by 802.3u and also by FDDI. The basic idea is to transform any arbitrary 4 bit word into a (relatively) balanced 5 bit word. This is done by a fast table lookup.

Balancing the code has many advantages: better bandwidth utilization, better laser efficiency (constant temperature), better bit-synchronization (PLL), etc.

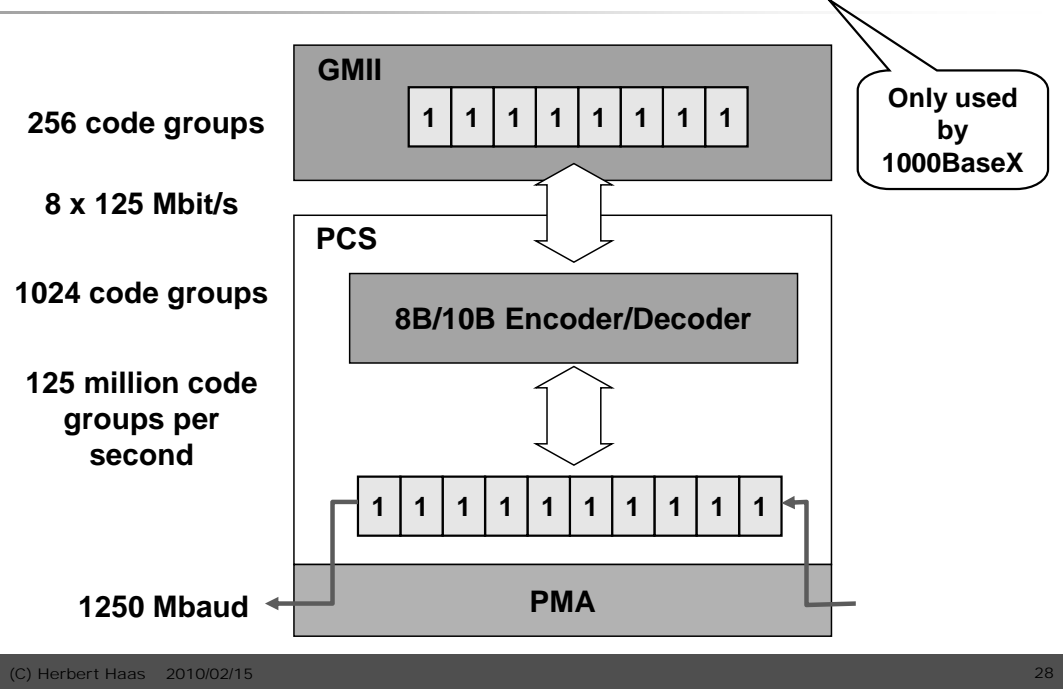
Note that the signaling overhead is $5/4 \rightarrow 12.5\%$.

Gigabit Ethernet



Gigabit Ethernet has been defined in March 1996 by the working group IEEE 802.3z. The GMII represents a abstract interface between the common Ethernet layer 2 and different signaling layers below. Two important signaling techniques had been defines: The standard 802.3z defines 1000Base-X signaling which uses 8B10B block coding and the 802.3ab standard uses 1000Base-T signaling. The latter is only used over twisted pair cables (UTP Cat 5 or better), while 1000BaseX is only used over fiber, with one exception, the twinax cable (1000BaseCX), which is basically a shielded twisted pair cable.

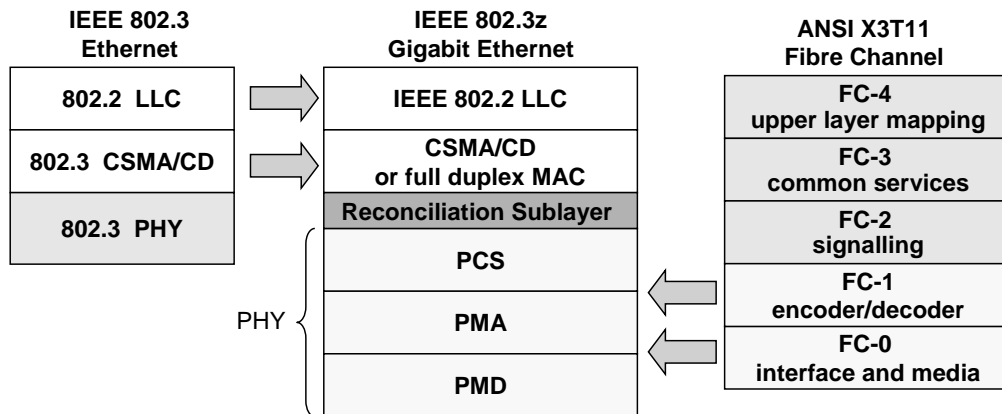
GE 8B/10B Coding



8B10B block coding is very similar to 4B5B block coding but allows fully balanced 10-bit codewords. Actually, there are not enough balanced 10-bit codewords available. Note that there are 256 8-bit codewords which need to be mapped on 1024 10-bit codewords. But instead of using a fully balanced 10-bit codeword for each 8-bit codeword, some 8-bit codewords are represented by two 10-bit codewords, which are sent in an alternating manner. That is, both associated 10-bit words are bit-complementary.

Again, the signaling overhead is 12.5%, that is 1250 Mbaud is necessary to transmit a bit stream of 1000 Mbit/s.

GE Signaling



Gigabit Ethernet layers have been defined by adaptation of the LLC and MAC layers of classical Ethernet and the physical layers of the ANSI Fiber Channel technology. A so-called reconciliation layer is used in between for seamless interoperation. The physical layer of the Fiber Channel technology uses 8B10B block coding.

1000BaseX



- **Two different wavelengths supported**
- **Full duplex only**
 - ♦ **1000Base-SX: short wave, 850 nm MMF, up to 550m max. distance**
 - ♦ **1000Base-LX: long wave, 1300 nm MMF or SMF, up to 5km max. distance**
- **1000Base-CX:**
 - ♦ **Twinax Cable (high quality 150 Ohm balanced shielded copper cable)**
 - ♦ **About 25 m distance limit, DB-9 or the newer HSSDC connector**

Gigabit Ethernet can be transmitted over various types of fiber. Currently (at least) two types are specified, short and long wave transmissions, using 850 nm and 1300 nm respectively. The long wave can be used with both single mode (SMF) and multimode fibers (MMF). Only SMF can be used for WAN transmissions because of the much lower dispersion effects.

Note that there are several other implementations offered by different vendors, such as using very long wavelengths at 1550 nm together with DWDM configurations.

The twinax cable is basically a shielded twisted pair cable.

1000BaseT

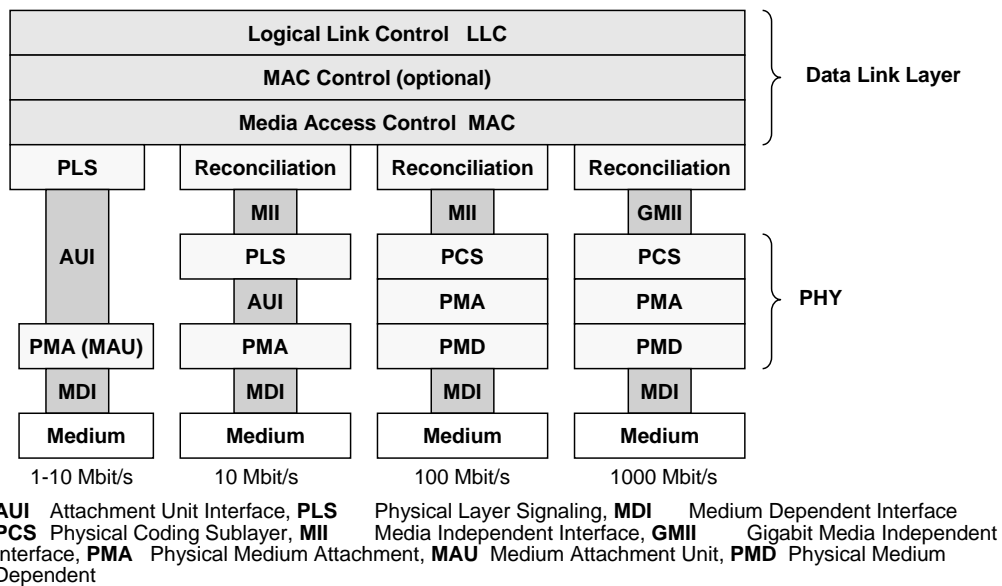


- **Defined by 802.3ab task force**
- **UTP**
 - ♦ **Uses all 4 line pairs simultaneously for duplex transmission! (echo cancellation)**
 - ♦ **5 level PAM coding**
 - 4 levels encode 2 bits + extra level used for Forward Error Correction (FEC)
 - ♦ **Signal rate: 4 x 125 Mbaud = 4 x 250Mbit/s data rate**
 - Cat. 5 links, max 100 m; all 4pairs, cable must conform to the requirements of ANSI/TIA/EIA-568-A
 - ♦ **Only 1 CSMA/CD repeater allowed in a collision domain**
 - ♦ **up to 100m max. distance**

It is very difficult to transmit Gigabit speeds over unshielded twisted pair cables. Only a mix of multiple transmission techniques ensure that this high data rate can be transmitted over a UTP Cat5 cable. For example all 4 pairs are used together for both directions. Echo cancellation ensures that the sending signal does not confuse the received signal. 5 level PAM is used for encoding instead of 8B10B because of its much lower symbol rate. Now we have only 125 Mbaud x 4 instead of 1250 Mbaud.

The interface design is very complicated and therefore relatively expensive. Using Cat 6 or Cat 7 cables allow 500 Mbaud x 2 pairs, that is 2 pairs are designated for TX and the other 2 pairs are used for RX. This dramatically reduces the price but requires better cables, which are not really expensive but slightly thicker. Legacy cable ducts might be too small in diameter.

Several Physical Media Supported



The diagram above shows various physical media designs supported by the official GE standard. Each modern GE card could theoretically support the old 10 Mbit/s standard as well. However many vendors create GE NICs that only support GE or GE and FE—who would connect a precious GE interface with another interface, which is 100 times slower?

10 Gigabit Ethernet / IEEE 802.3ae



- **Only optical support**
 - ◆ 850nm (MM) / 1310nm / 1550 nm (SM only)
 - ◆ No copper PHY anymore !
- **Different implementations at the moment – standardization not finished!**
- **8B/10B (IBM), SONET/SDH support, ...**
- **XAUI ("Zowie") instead of GMII**

10 GE only supports optical links. Note that GE is actually a synchronous protocol! There is no statistical multiplexing done at the physical layer anymore, because optical switching at that bit rate only allows synchronous transmissions.

The GMII has been replaced (or enhanced) by the so-called XAUI, known as "Zowie".

Note: At the time of writing this module, the 10 GE standard was not fully finished. Though, some vendors already offer 10 GE interface cards for their switches.

These interfaces are very expensive but the investment ensures backward compatibility to lower Ethernet rates and at the same time provides a very high speed WAN interface.

An alternative technology would be OC192, which requires a very expensive and complex SONET/SDH environment.

10 Gigabit Ethernet (IEEE 802.3ae)

- **Preserves Ethernet framing**
- **Maintains the minimum and maximum frame size of the 802.3 standard**
- **Supports only full-duplex operation**
 - CSMA/CD protocol was dropped
- **Focus on defining the physical layer**
 - Four new optical interfaces (PMD)
 - To operate at various distances on both single-mode and multi-mode fibers
 - Two families of physical layer specifications (PHY) for LAN and WAN support
 - Properties of the PHY defined in corresponding PCS
 - Encoding and decoding functions

PMDs

- **10GBASE-L**
 - SM-fiber, 1300nm band, maximum distance 10km
- **10GBASE-E**
 - SM-fiber, 1550nm band, maximum distance 40km
- **10GBASE-S**
 - MM-fiber, 850nm band, maximum distance 26 – 82m
 - With laser-optimized MM up to 300m
- **10GBASE-LX4**
 - For SM- and MM-fiber, 1300nm
 - Array of four lasers each transmitting 3,125 Gbit/s and four receivers arranged in WDM (Wavelength-Division Multiplexing) fashion
 - Maximum distance 300m for legacy FDDI-grade MM-fiber
 - Maximum distance 10km for SM-fiber

WAN PHY / LAN PHY and their PCS

- **LAN-PHY**

- 10GBASE-X
- 10GBASE-R
 - 64B/66B coding running at 10,3125 Gbit/s

- **WAN-PHY**

- 10GBASE-W
 - 64B/66B encoded payload into SONET concatenated STS192c frame running at 9,953 Gbit/s
 - Adaptation of 10Gbit/s to run over traditional SDH links

IEEE 802.3ae PMDs, PHYs, PCSs

		PCS		
PMD	10GBASE-E	10GBASE-ER		10GBASE-EW
	10GBASE-L	10GBASE-LR		10GBASE-LW
	10GBASE-S	10GBASE-SR		10GBASE-SW
	10GBASE-L4		10GBASE-LX4	
		LAN PHY		WAN PHY

10 Gigabit Ethernet over Copper

- **IEEE 802.3ak defined in 2004**
 - 10GBASE-CX4
 - Four pairs of twin-axial copper wiring with IBX4 connector
 - Maximum distance of 15m
- **IEEE 802.3an working group**
 - 10GBASE-T
 - CAT6 UTP cabling with maximum distance of 55m to 100m
 - CAT7 cabling with maximum distance of 100m
 - Standard ratification expected in July 2006

Note



- **GE and 10GE use synchronous physical sublayer !!!**
- **Recommendation: Don't use GE over copper wires**
 - ◆ **Radiation/EMI**
 - ◆ **Grounding problems**
 - ◆ **High BER**
 - ◆ **Thick cable bundles (especially Cat-7)**

Both GE and 10GE are synchronous physical technologies on fiber. It is not recommended to use GE over copper wires anymore although 802.3ab would specify it. This is because the whole electrical hardware (cables and connectors) are re-used from older Ethernet technologies and have not been designed to support such high frequencies.

For example the RJ45 connector is not HF proof. Furthermore, shielded twisted pair cables require a very good grounding, seldom found in reality. The Bit Error Rate (BER) is typically so high that the effective data rate is much lower than GE, for example 30% only.

Summary



- **Ethernet evolved in the opposite direction:**
 - ♦ Collision free
 - ♦ WAN qualified
 - ♦ Switched
- **Several coding styles → Complex PHY architecture**
- **Plug & play through autonegotiation**
- **Much simpler than ATM but no BISDN solution – might change!**

Quiz



- **Why tends high-speed Ethernet to synchronous PHY?**
- **Can I attach a 100 Mbit/s port to a 1000 Mbit/s port via fiber?**
- **What is the idea of Etherchannels?
(Maximum bit rate, difference to multiple parallel links)**

Q1: On fiber its difficult to deal with asynchronous transmission, photons cannot be buffered easily, store and forward problems

Q2: No, autonegotiation on fiber does not care for data rates

Q3: "normal" parallel links would be disabled by STP, Etherchannel supports up to 8 links