L08 - Ethernet Evolution (v6.0)

The Ethernet Evolution

From 10Mbit/s to 10Gigabit/s Ethernet Technology
From Bridging to L2 Ethernet Switching and VLANs
From LAN to WAN Transmission Technique

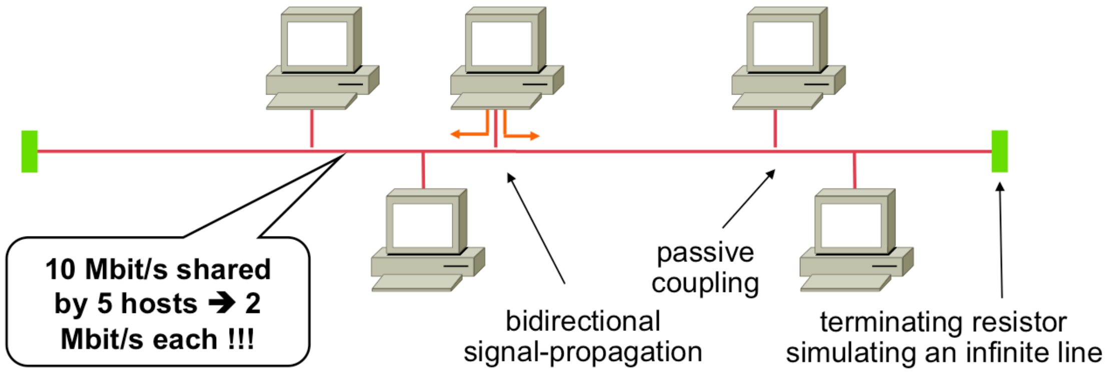
Agenda

- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

Ethernet Primer (v7.0)

History: Initial Idea

- Bus topology based on coax cable (two wires)
  - ➔ half duplex transmission, natural broadcast behavior
- No active network elements (no store and forward) ➔ low latency
- Shared media needs media access control (MAC)
  - ➔ CSMA/CD (Carrier Sense Multiple Access / Collision Detection)
  - ➔ MAC addresses (OSI L2 address)
- One single collision domain and one broadcast domain



 Ethernet Evolution, v6.0 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. A repeater is just a simple signal amplifier used to enlarge the network diameter according the repeater rules but there is not any kind of network segmentation -> it is still one collision domain!

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

L08 - Ethernet Evolution (v6.0)

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

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

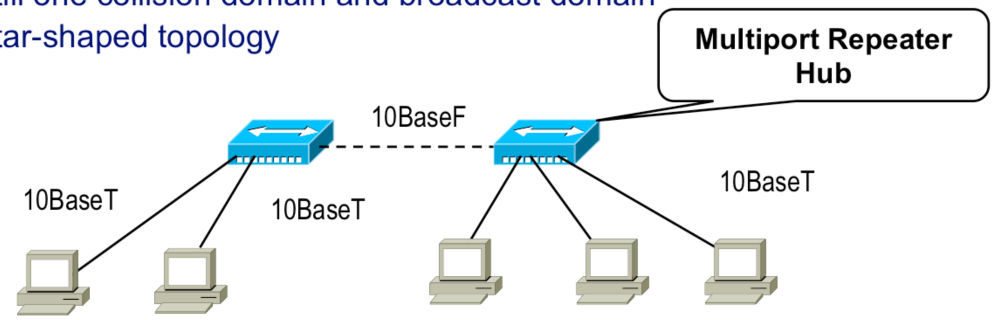
L08 - Ethernet Evolution (v6.0)

History: Multiport Repeaters

- Demand for structured cabling
  - 10BaseT (Cat3, Cat4, ...)
  - (voice-grade twisted-pair)
- Multiport repeater ("Hub") created
  - "CSMA/CD in a box“
  - The 180 Degree Turn (“Star instead Bus”)
- Still half duplex communication
- Still one collision domain and broadcast domain
- Star-shaped topology

——— represents four CU wires
2 for Tmt, 2 for Rcv
(e.g. 10BaseT)

- - - - represents two FO wires
e.g. 10BaseF



 Ethernet Evolution, v6.0 5

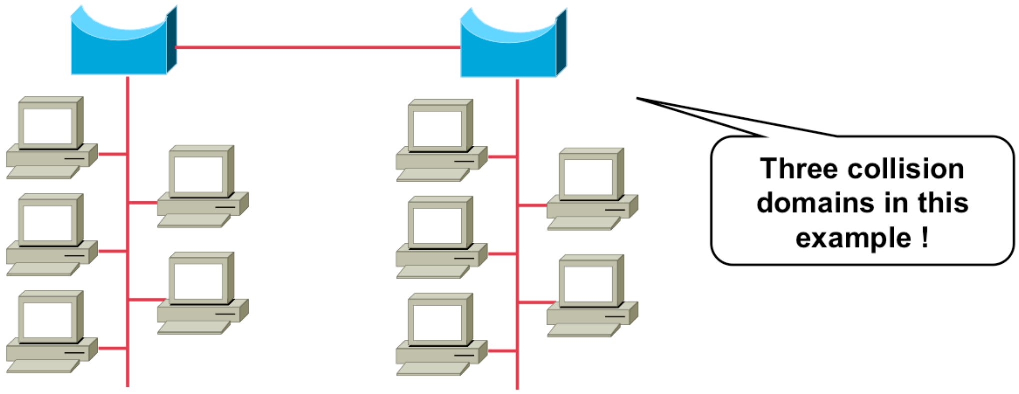
Later, Ethernet devices supporting structured cabling were created in order to reuse the voice-grade twisted-pair cables already installed in buildings. 10BaseT had been specified to support Cat3 cables (voice grade) or better, for example Cat4 (and today Cat5, Cat6, and Cat7).

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



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).

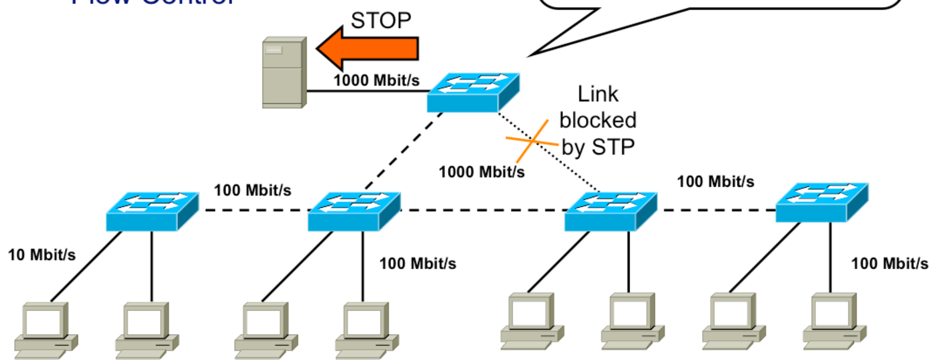
Bridges are store and forwarding devices (introducing significant delay) that can filter traffic based on the destination MAC addresses to avoid unnecessary flooding of frames to certain segments. Thus, bridges segment the LAN into several collision domains. Broadcasts are still forwarded to allow layer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

L08 - Ethernet Evolution (v6.0)

History: Switches

- Switch = Multiport Bridges with HW acceleration
- Full duplex ➔ Collision-free Ethernet ➔ No CSMA/CD necessary anymore
  - No collision domains anymore
- Different data rates at the same time supported
  - Autonegotiation
- Still one broadcast domain
- STP in case of redundancy
- Flow Control

Collision-free plug & play scalable Ethernet !



 Ethernet Evolution, v6.0 7

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. Cut-through means that forwarding a frame to the other port just happens when the Ethernet header of that frame is received on the incoming port without waiting for the frame to be complete and fully stored.

There is no need for collision detection (media access control) on a link which is shared by two devices only. All devices on such links use a separate physical wire for transmit and receive and inherently act as store and forward devices for each direction. Therefore CSMA/CD can be turned off and full duplex operation (receiving and transmitting at the same time) becomes possible. But now CSMA/CD is not able to slow down devices, if there is to much traffic in the LAN. We need a new element which is flow control between switch and end system. Now a switch can tell an end system to stop, if the switch recognizes congestion based on to much traffic is going to be stored in its buffers for transmission.

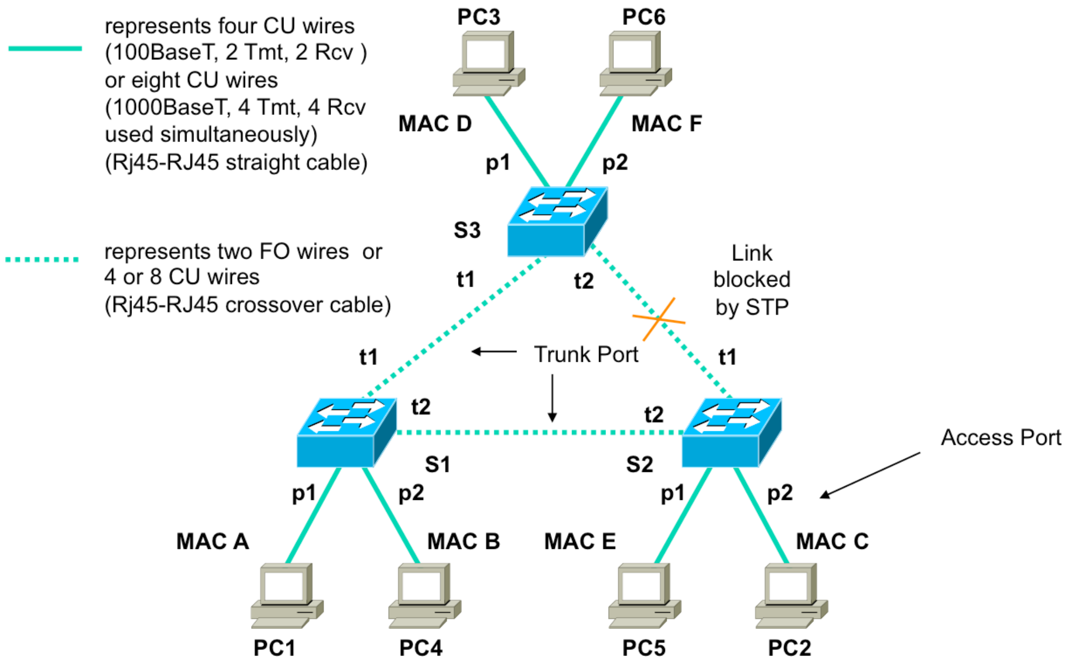
The next benefit of store and forward performed by L2 switches is that now different speeds on different links. Clients may use 10 Mbit/s, servers may use 100 Mbit/s or 1000Mbit/s and Interswitch links may go up to 10GBit/s speed nowadays. That was one reason for success of the Ethernet family. Even with change of speed the Ethernet frame looks just like in the old days. Of course cut-through switching is not possible if the speeds are different, because when forwarding a frame to the higher speed port before the frame is received on the lower speed port it can happen that the bits to be transmitted are not already there when needed.

No complicated translation technique has to be used when forwarding between links with different speeds. Recognize that a multiport repeater is not able to allow speed differences between links. All links must have the same speed.

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.

L08 - Ethernet Evolution (v6.0)

Ethernet Switching Topology
(MAC Address Table - Empty)

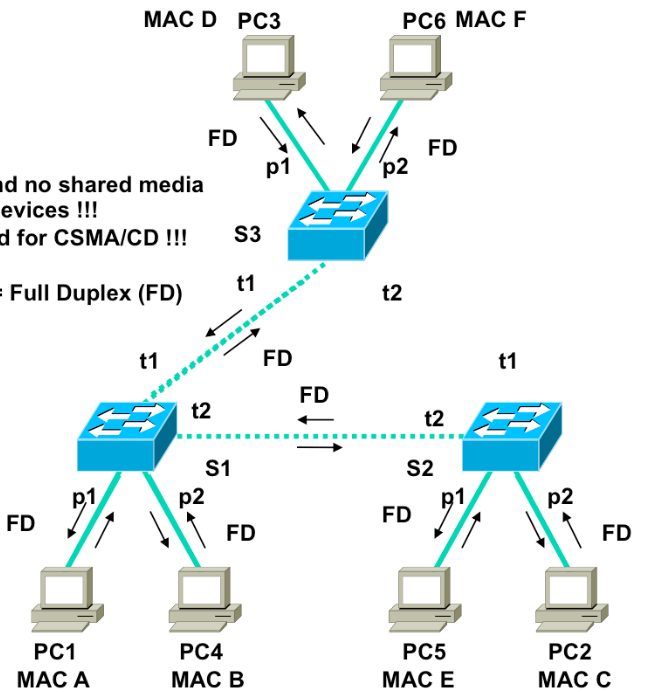


L08 - Ethernet Evolution (v6.0)

Ethernet Switching – Full Duplex (FD)
(Point-to-Point Links and FD Everywhere)

Only PTP links and no shared media for more than 2 Devices !!!
Therefore no need for CSMA/CD !!!

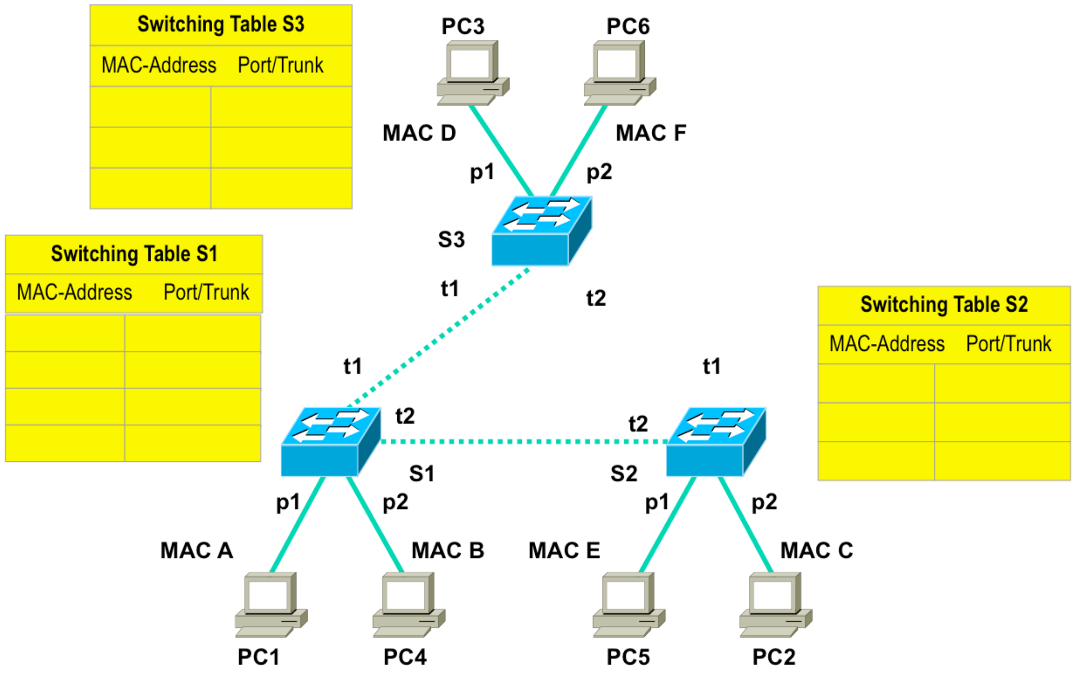
CSMA/CD OFF == Full Duplex (FD)



 Ethernet Evolution, v6.0 9

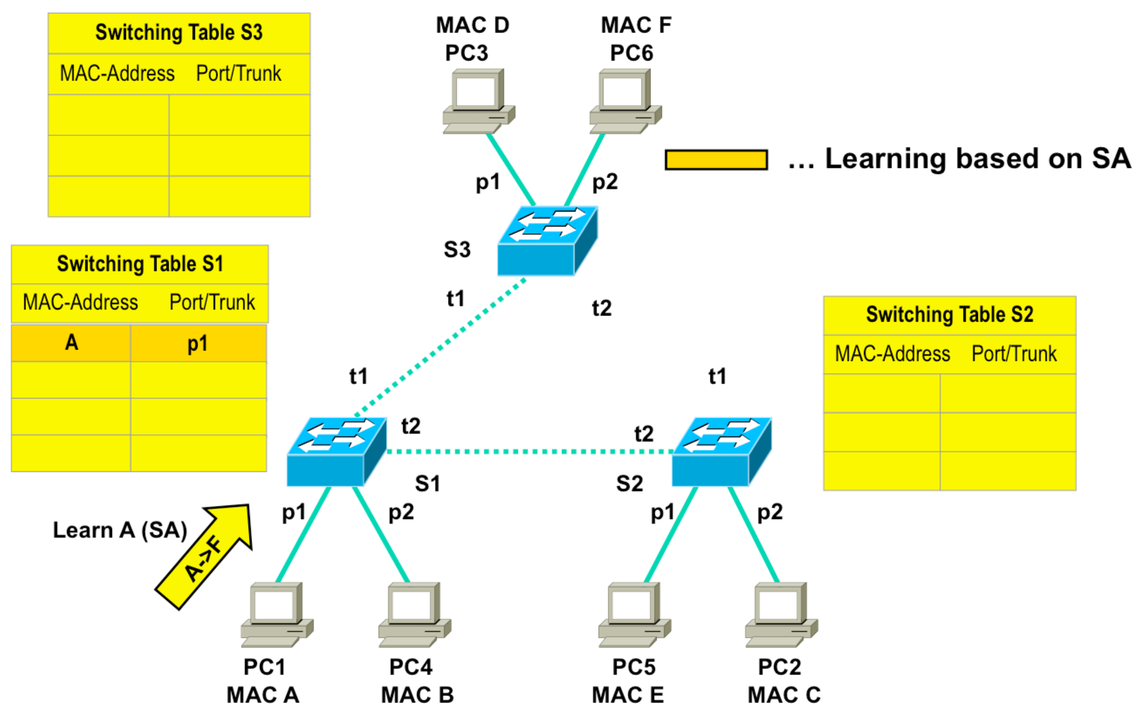
L08 - Ethernet Evolution (v6.0)

Ethernet Switch Table - Power On
(MAC Address Table - Empty)



 Ethernet Evolution, v6.0 10

Table Building for Ethernet Frame MAC-A to MAC-F 1

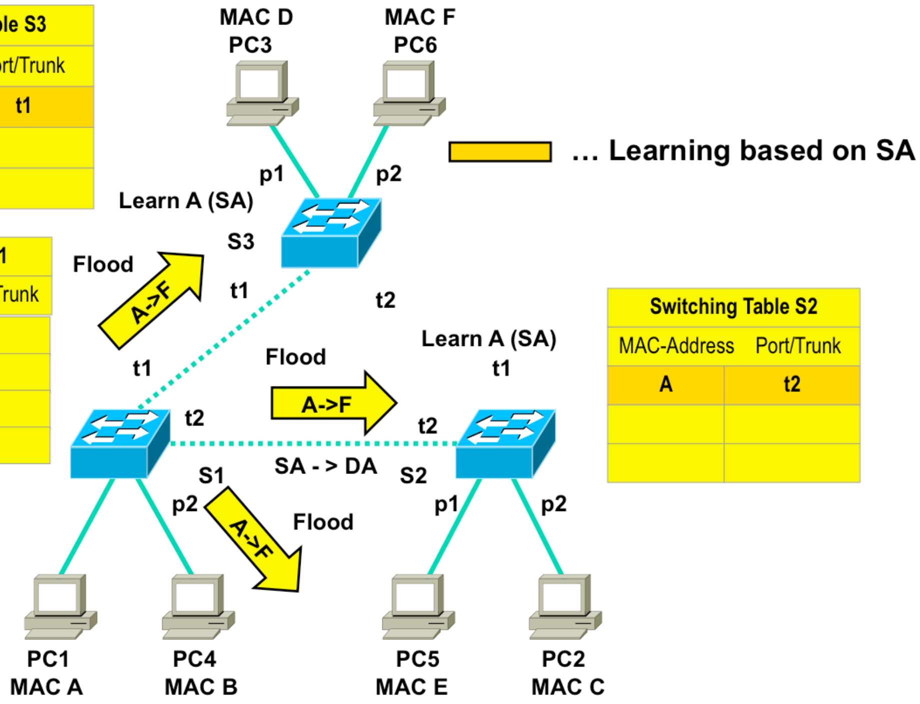


L08 - Ethernet Evolution (v6.0)

Table Building for Ethernet Frame MAC-A to MAC-F 2

Switching Table S3	
MAC-Address	Port/Trunk
A	t1

Switching Table S1	
MAC-Address	Port/Trunk
A	p1

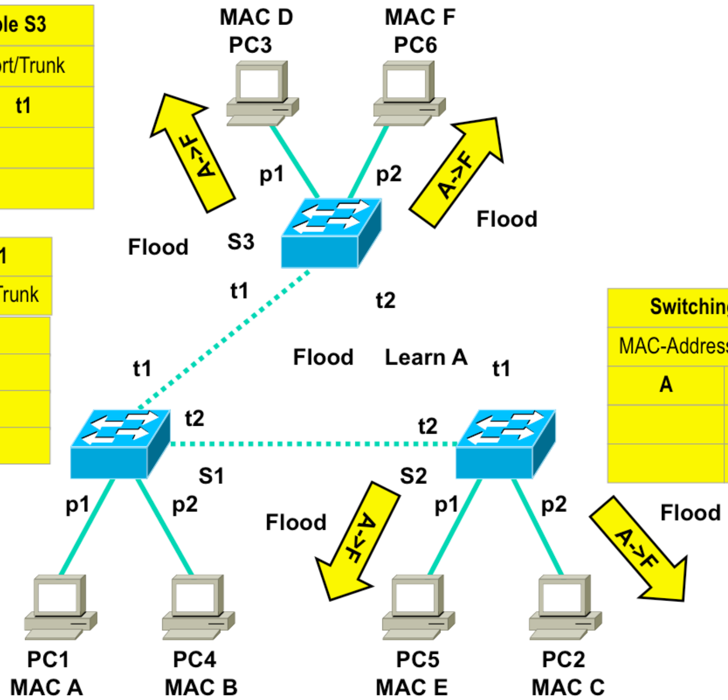


Switching Table S2	
MAC-Address	Port/Trunk
A	t2

Table Building for Ethernet Frame MAC-A to MAC-F 3

Switching Table S3	
MAC-Address	Port/Trunk
A	t1

Switching Table S1	
MAC-Address	Port/Trunk
A	p1



Switching Table S2	
MAC-Address	Port/Trunk
A	t2

 Ethernet Evolution, v6.0 13

L08 - Ethernet Evolution (v6.0)

Table Building / Table Usage for Ethernet Frame MAC-F to MAC-A 1

Switching Table S3	
MAC-Address	Port/Trunk
A	t1
F	p2

Switching Table S1	
MAC-Address	Port/Trunk
A	p1

Switching Table S2	
MAC-Address	Port/Trunk
A	t2

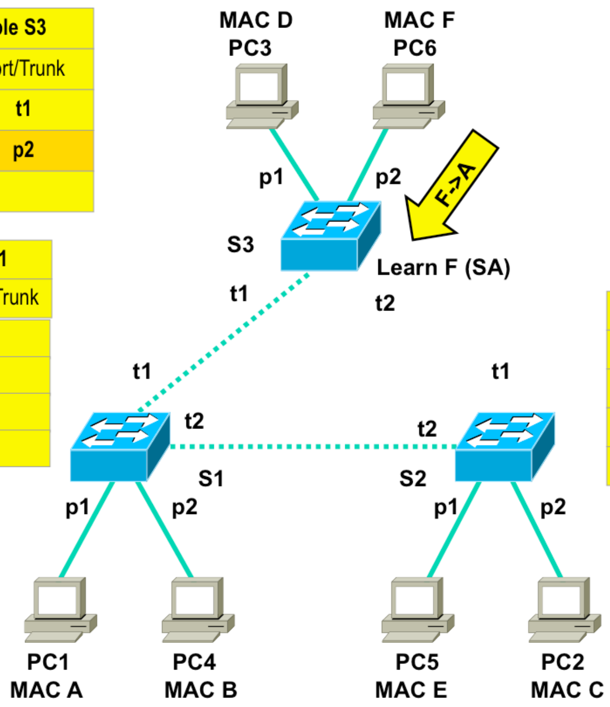


Table Building / Table Usage (Forwarding Decision) for Ethernet Frame MAC-F to MAC-A 2

Switching Table S3

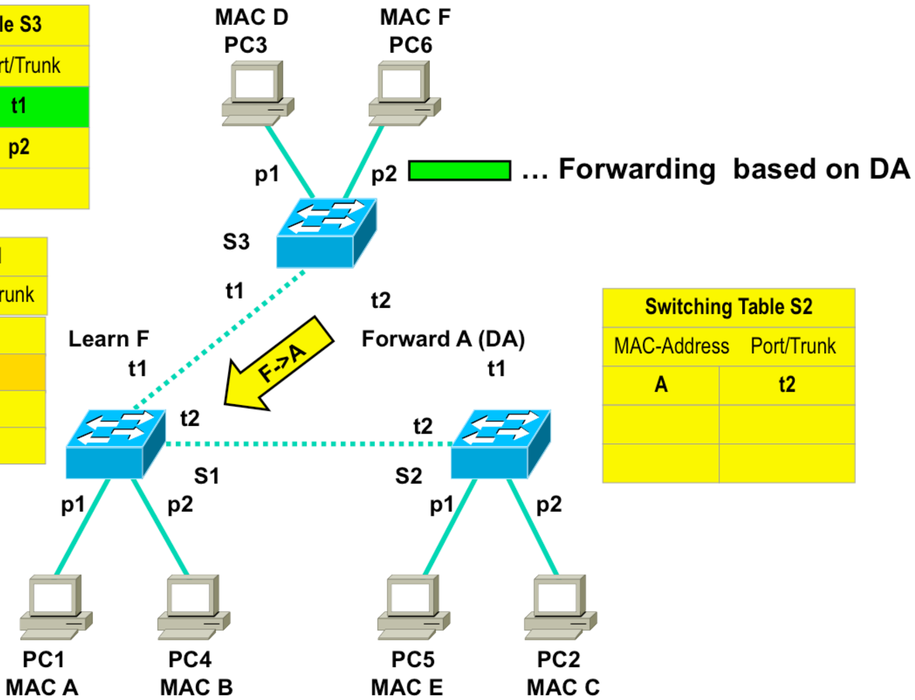
MAC-Address	Port/Trunk
A	t1
F	p2

Switching Table S1

MAC-Address	Port/Trunk
A	p1
F	t1

Switching Table S2

MAC-Address	Port/Trunk
A	t2



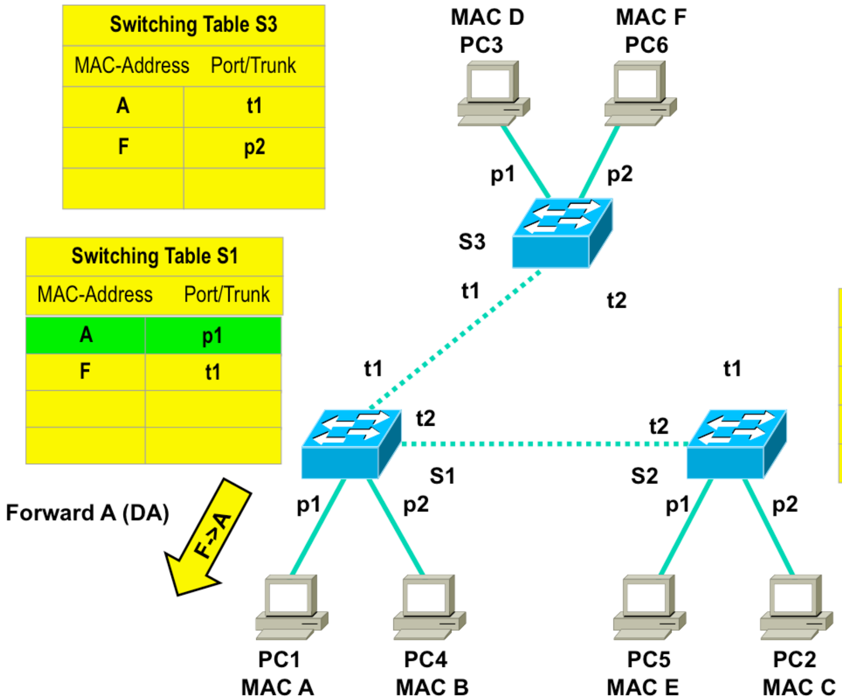
 Ethernet Evolution, v6.0 15

Table Building / Table Usage (Forwarding Decision) for Ethernet Frame MAC-F to MAC-A 3

Switching Table S3	
MAC-Address	Port/Trunk
A	t1
F	p2

Switching Table S1	
MAC-Address	Port/Trunk
A	p1
F	t1

Switching Table S2	
MAC-Address	Port/Trunk
A	t2

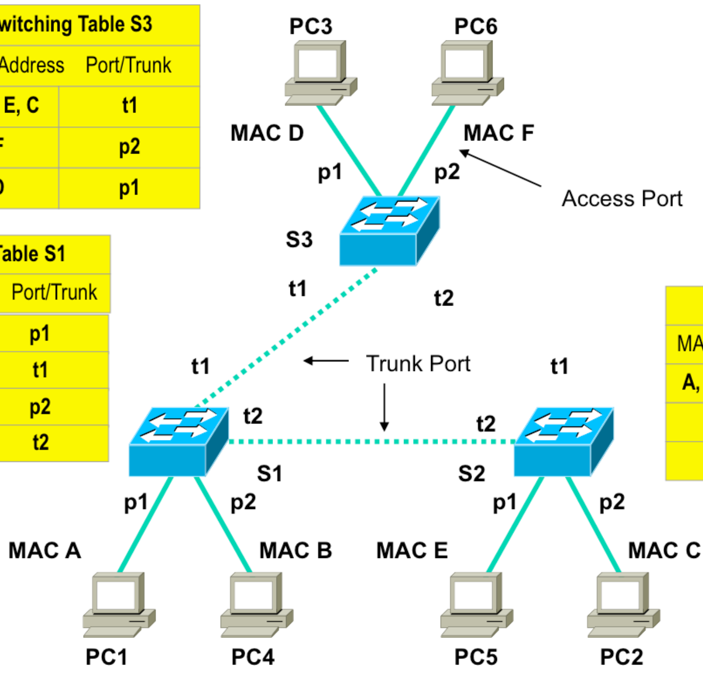


Ethernet Switch Table – Final State
(All MAC addresses learned)

Switching Table S3	
MAC-Address	Port/Trunk
A, B, E, C	t1
F	p2
D	p1

Switching Table S1	
MAC-Address	Port/Trunk
A	p1
F, D	t1
B	p2
E, C	t2

Switching Table S2	
MAC-Address	Port/Trunk
A, B, D, F	t2
E	p1
C	p2



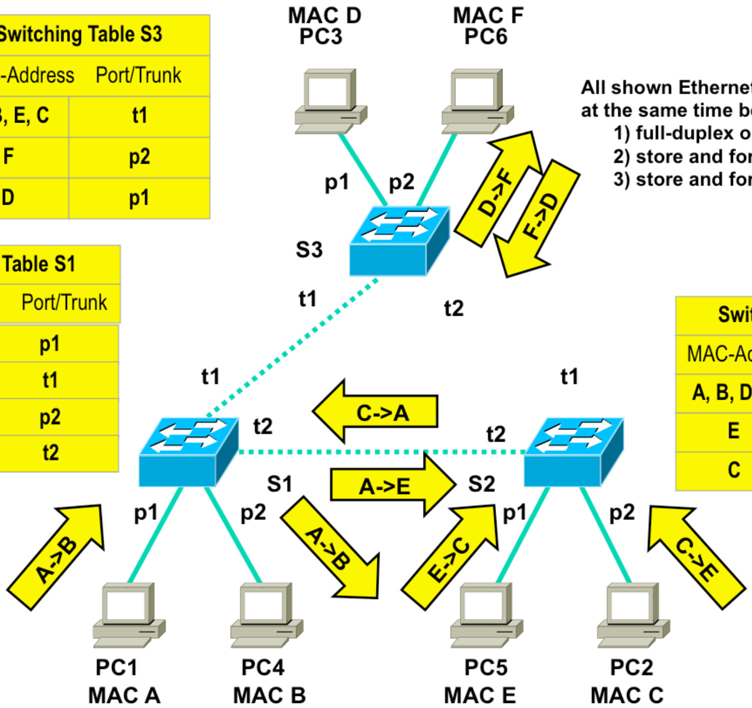
Ethernet Switching – Decoupling (Improved Performance <-> Collision Domains)

All shown Ethernet frames are forwarded at the same time because of:
1) full-duplex on every link
2) store and forward at switch
3) store and forward at NIC of PC

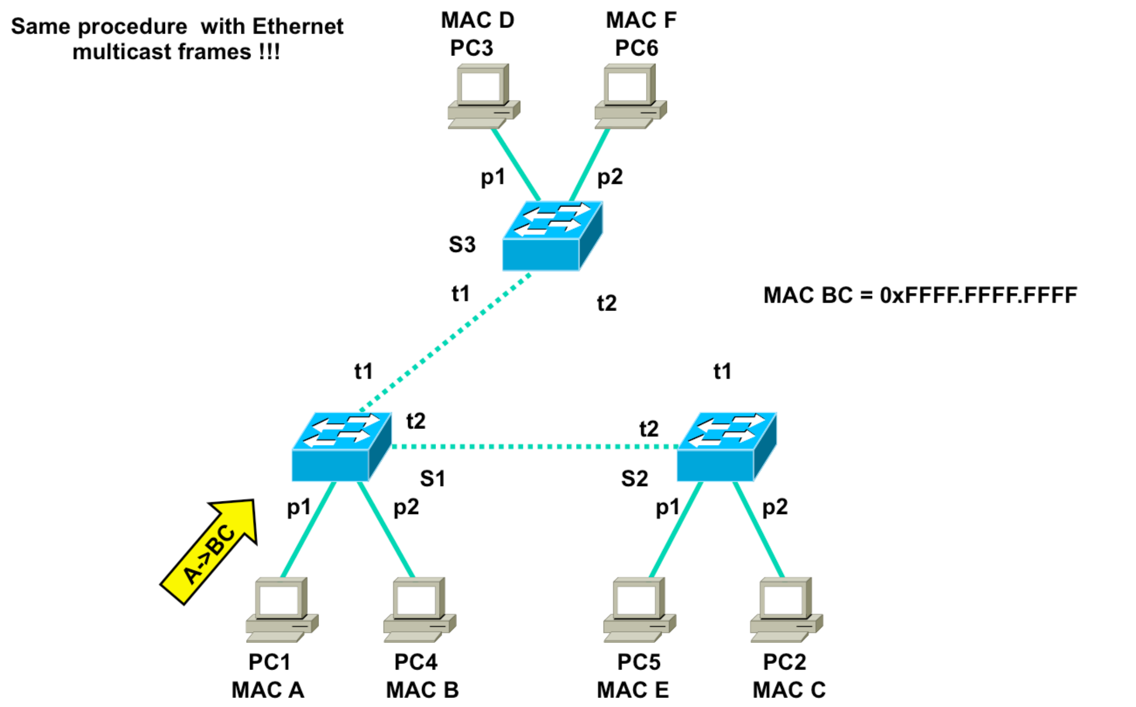
Switching Table S3	
MAC-Address	Port/Trunk
A, B, E, C	t1
F	p2
D	p1

Switching Table S1	
MAC-Address	Port/Trunk
A	p1
F, D	t1
B	p2
E, C	t2

Switching Table S2	
MAC-Address	Port/Trunk
A, B, D, F	t2
E	p1
C	p2



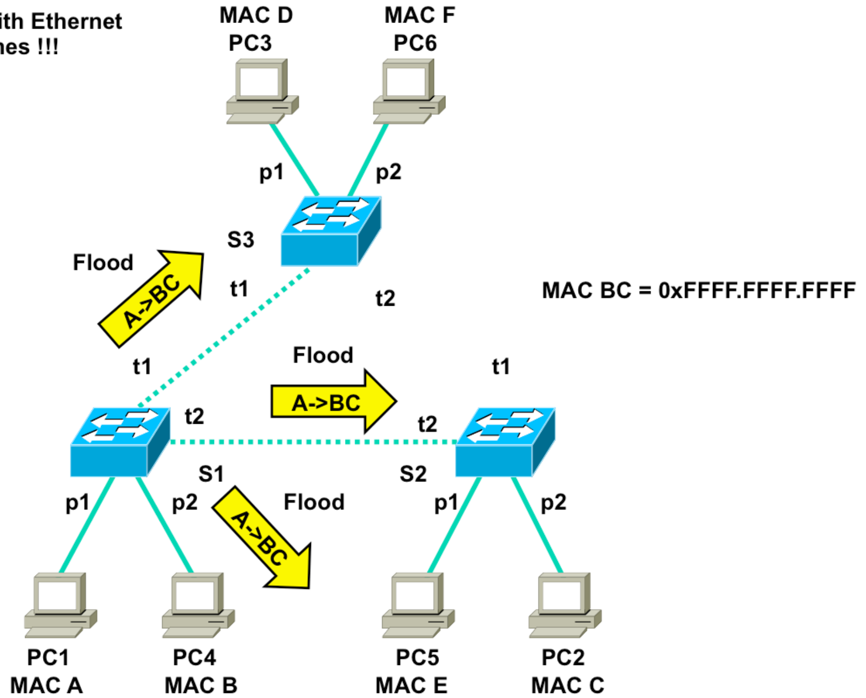
Ethernet Broadcast (BC) 1



L08 - Ethernet Evolution (v6.0)

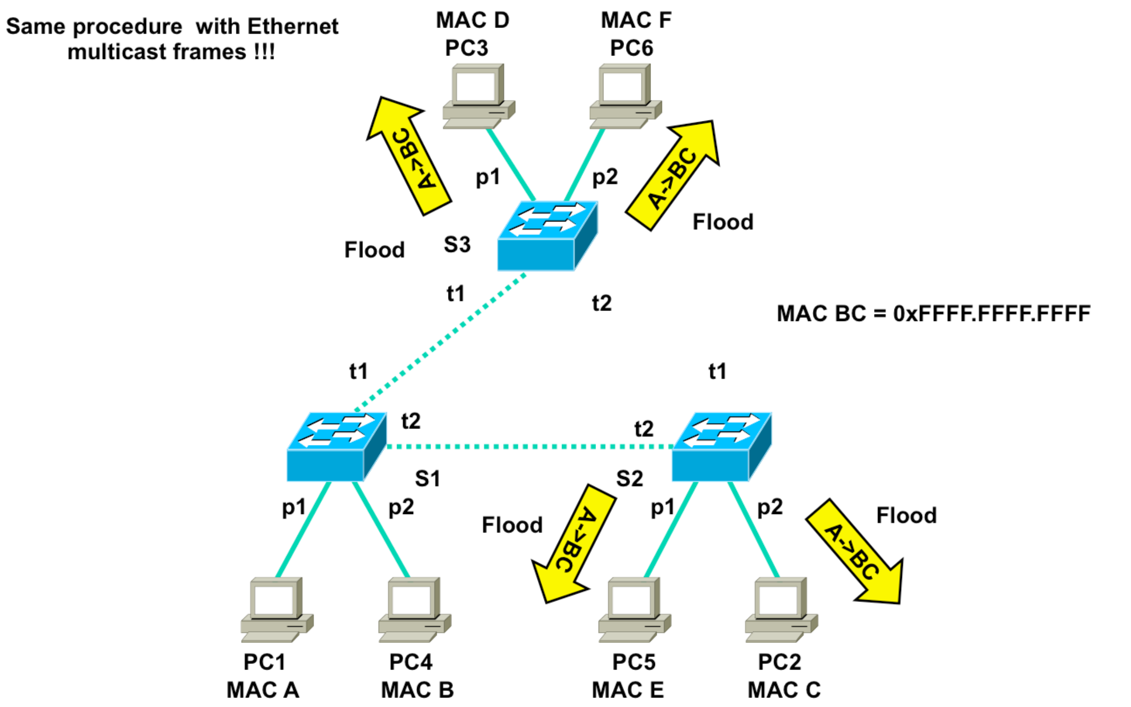
Ethernet Broadcast (BC) 2

Same procedure with Ethernet multicast frames !!!



L08 - Ethernet Evolution (v6.0)

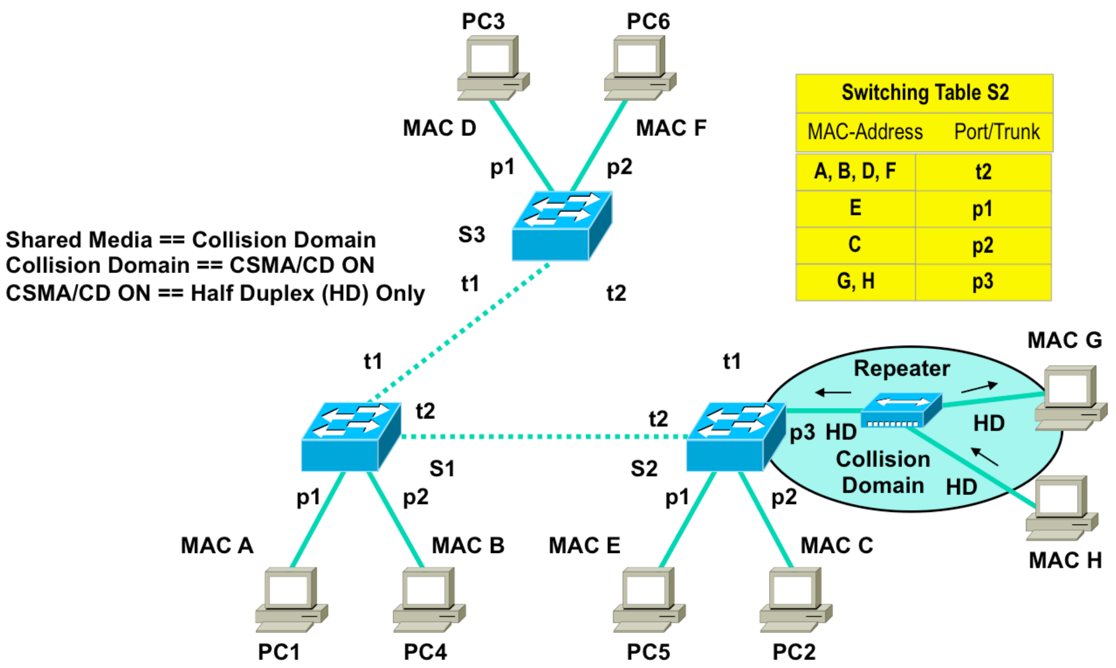
Ethernet Broadcast (BC) 3



 Ethernet Evolution, v6.0 21

L08 - Ethernet Evolution (v6.0)

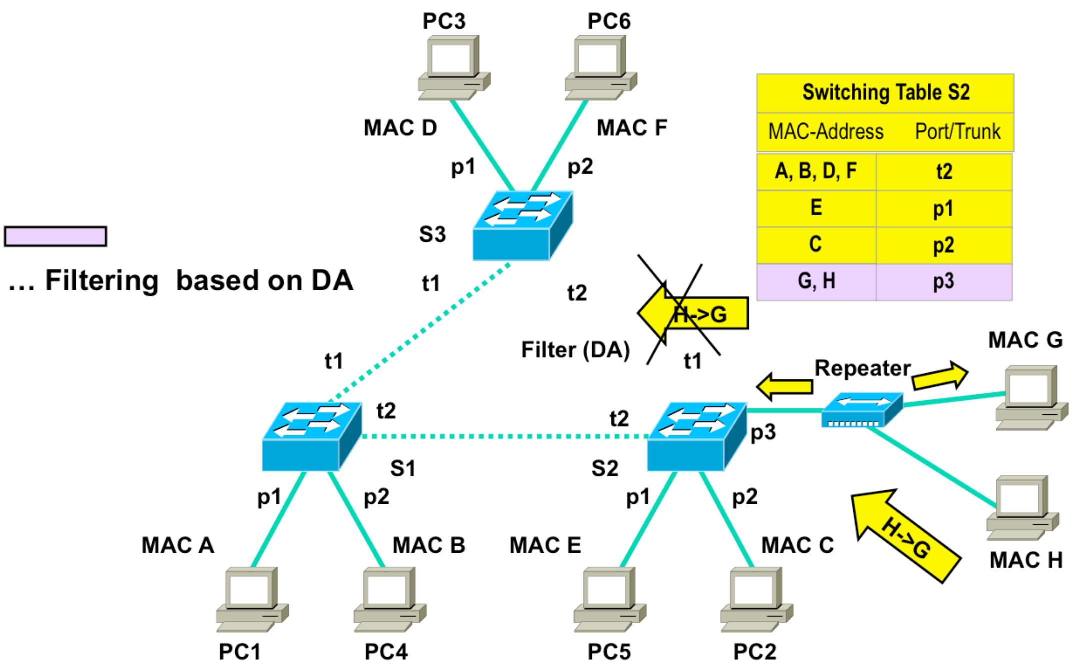
Ethernet Switching – Repeater (Hub)
(On One Link -> Shared Media – Half Duplex (HD))



Switching Table S2	
MAC-Address	Port/Trunk
A, B, D, F	t2
E	p1
C	p2
G, H	p3

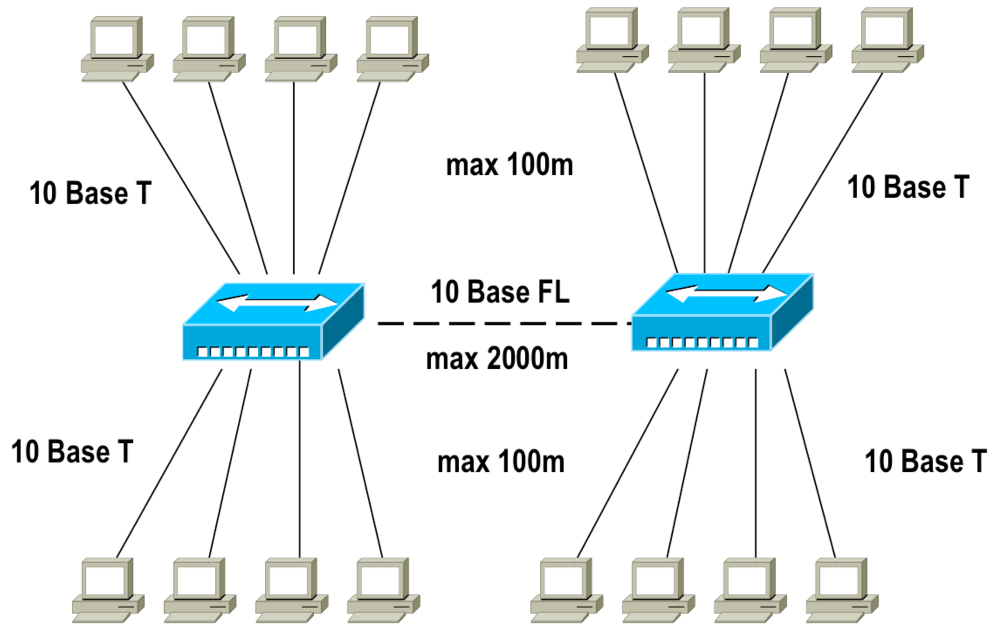
 Ethernet Evolution, v6.0 22

Table Usage (Filtering Decision) for Ethernet Frame MAC-H to MAC-G

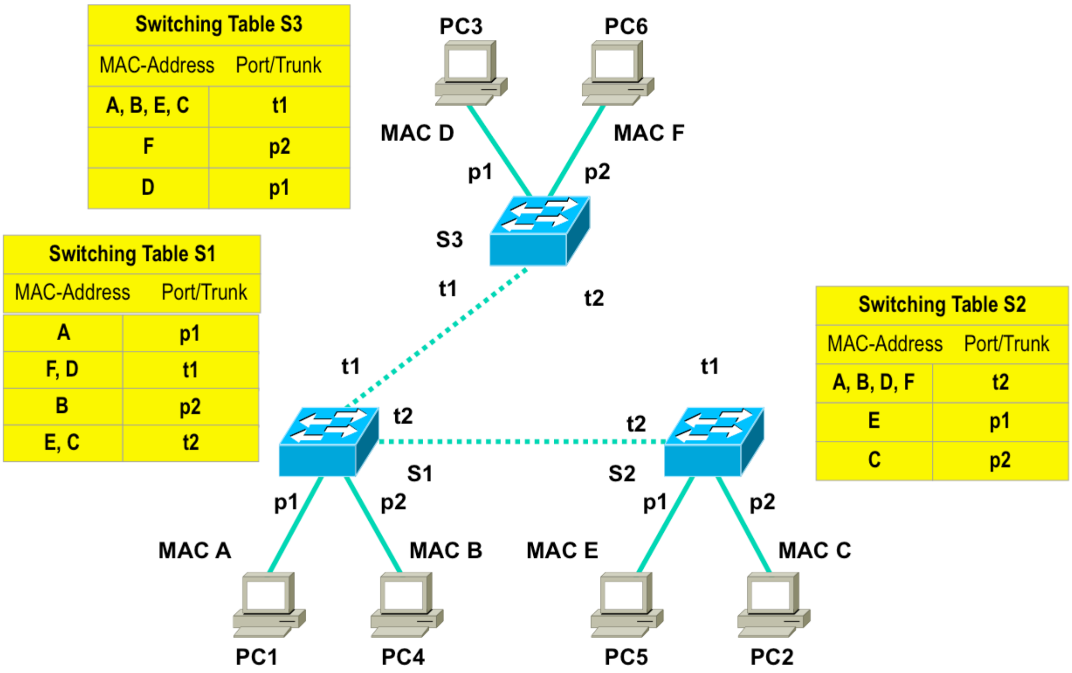


Switching Table S2	
MAC-Address	Port/Trunk
A, B, D, F	t2
E	p1
C	p2
G, H	p3

Ethernet with Repeater: Network Sniffing? Yes -> Ethernet Card -> Promiscuous Mode



Ethernet with Switches: Network Sniffing? Not So Easy -> Because of Inherent Filtering



Switching Table S3	
MAC-Address	Port/Trunk
A, B, E, C	t1
F	p2
D	p1

Switching Table S1	
MAC-Address	Port/Trunk
A	p1
F, D	t1
B	p2
E, C	t2

Switching Table S2	
MAC-Address	Port/Trunk
A, B, D, F	t2
E	p1
C	p2

Modern Switching Features

- **Different data rates supported simultaneously**
  - 10, 100, 1000, 10000 Mbit/s depending on switch
- **Full duplex operation**
- **QoS (802.1p)**
  - Queuing mechanisms
  - Flow control
- **Security features**
  - Restricted static mappings (DA associated with source port)
  - Port secure (Limited number of predefined users per port)
- **Different forwarding**
  - Store & Forward
  - Cut-through
  - Fragment-Free
- **VLAN support (tagging, trunking, 802.1Q)**
- **Spanning Tree (RSTP, MSTP, PVST+)**
- **SPAN (for monitoring traffic)**

 Ethernet Evolution, v6.0 26

Today most switches support different data rates at each interface or at selected interfaces. Also full duplex operation is standard today. QoS might be supported by using sophisticated queuing techniques, 802.1p priority tags, and flow control features, such as the pause MAC control frame.

Security is provided by statically entered switching tables and port locking (port secure), that is only a limited number or predefined users are allowed at some designated ports.

Forwarding of frames can be significantly enhanced using cut through switching: the processor immediately forwards the frame when the destination is determined. The switching latency is constant and very short for all length of packets but the CRC is not checked. In the Fragment-Free switching mode, the switch waits for the collision window (64 bytes) to pass before forwarding. If a packet has an error or better explained, a collision, it almost always occurs within the first 64 bytes. Fragment-Free mode provides better error checking than the Cut through mode with practically no increase in latency. The store and forward mode is the classical forwarding mode.

VLAN support allows to separate the whole LAN into multiple broadcast domains, hereby improving performance and security.

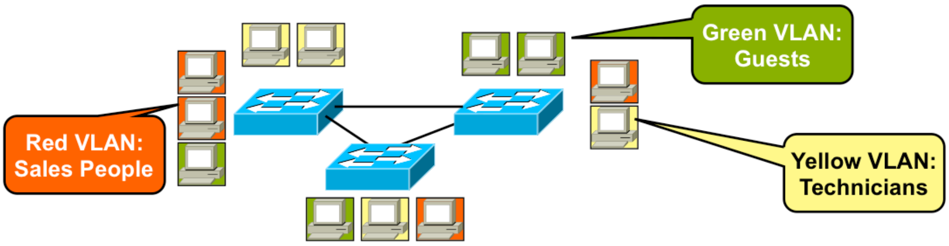
The spanning tree protocol (STP) avoids broadcast storms in a LAN. It was already described in the last chapter.

Agenda

- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

Virtual LANs

- **Separate LAN into multiple broadcast domains**
  - No global broadcasts anymore
  - For security reasons
- **Assign users to "VLANs“**
- **Base Idea:**
  - Multiplexing of several LANs over the same infrastructure (Ethernet switches and connection between switches)



 Ethernet Evolution, v6.0 28

Since most organizations consist of multiple "working groups" it is reasonable to confine their produced traffic somehow. Today's work-groups are expanding over the whole campus and users of one workgroup should be kept separated from other workgroups because of security reasons. They should see their necessary working environment only. End-systems of one workgroup should see broadcasts only from stations of same workgroup. But at all the network must be flexible to adapt to continuous location changes of the end-systems/users.

This is achieved using Virtual LANs (VLANs). Switches configured for VLANs consist logically of multiple virtual switches inside. Users/End systems are assigned to dedicated VLANs and there is no communication possible between different VLANs—even broadcasts are blocked! This significantly enhances security. On an Ethernet switch each VLAN is identified by a number and a name (optionally) but in our example we also use colors to differentiate them.

Ethernet switches supporting VLAN technique maintain separate bridging/switching tables per VLAN, handle separate broadcast domains per VLAN, but still have to deal with spanning-tree.

There are several solutions how to implement STP in case of VLANs:

1) original 802.1D standard specifies one single STP to be used for all VLANs together. That means the traffic of all VLANs travels along the same Spanning-Tree.

2) Cisco implements a per-VLAN STP. That means by differently tuning STP parameters per VLAN, different links are used by the VLAN traffic.

3) Later the MST (Multiple Instances Spanning Tree) standard allows something similar to the Cisco solution. The difference to Cisco is the better scalability if a large number of VLANs is used. MST allows to deal with a number of necessary Spanning-Trees given by the specific topology but avoids Spanning-Trees per VLAN.

Host to VLAN Assignment

- **Different solutions**
  - Port based assignment
  - Source address assignment
  - Protocol based
  - Complex rule based
  - 802.1X based on the credentials of a user / machine provided by EAP authentication
- **Bridges are interconnected via VLAN trunks**
  - IEEE 802.1Q (former 802.1s)
  - ISL (Cisco)
  - IEEE 802.10 (pre 802.1Q temporary solution, outdated)

 Ethernet Evolution, v6.0 29

There are different ways to assign hosts (users) to VLANs. The most common is the port-based assignment, meaning that each port has been configured to be member of a VLAN. Simply attach a host there and its user belongs to that VLAN specified.

Hosts can also be assigned to VLANs by their MAC address. Also special protocols can be assigned to dedicated VLANs, for example management traffic. Furthermore, some devices allow complex rules to be defined for VLAN assignment, for example a combination of address, protocol, etc.

Example how a station may be assigned to a VLAN:

Port-based: fixed assignment port 4 -> VLAN x, most common approach, a station is member of one specific VLAN only, administrator has to reconfigure a switch in order to support a location change of a user.

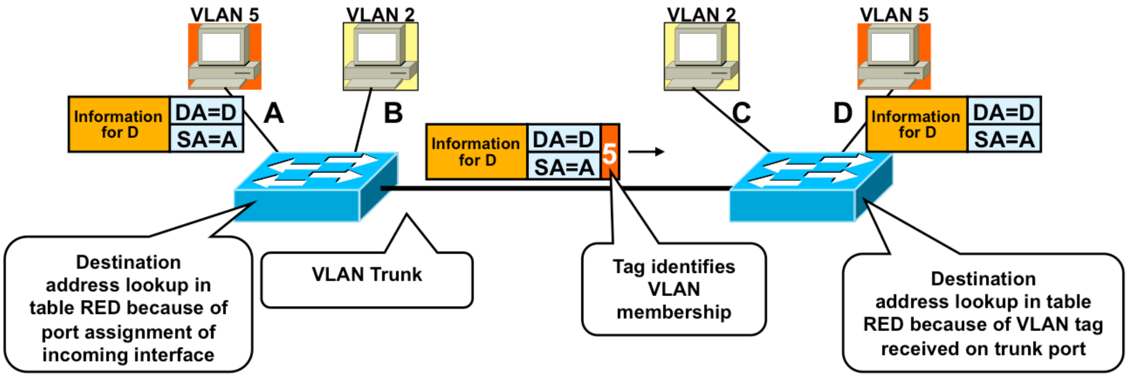
MAC-address based: MAC A -> VLAN x, allows integration of older shared-media components and automatic location change support, a station is member of one specific VLAN only.

Protocol-based: IP-traffic, port 1 -> VLAN x and NetBEUI-traffic, port 1 -> VLAN y, a station could be member of different VLANs

802.1X-based: User A -> VLAN engineering, User B -> VLAN finance, automatic location change support.

Of course VLANs should span over several bridges. This is supported by special VLAN trunking protocols, which are only used on the trunk between two switches. Two important protocols are commonly used: the IEEE 802.1q protocol and the Cisco Inter-Switch Link (ISL) protocol. Both protocols basically attach a "tag" at each frame which is sent over the trunk.

VLAN Trunking with Port Assignment

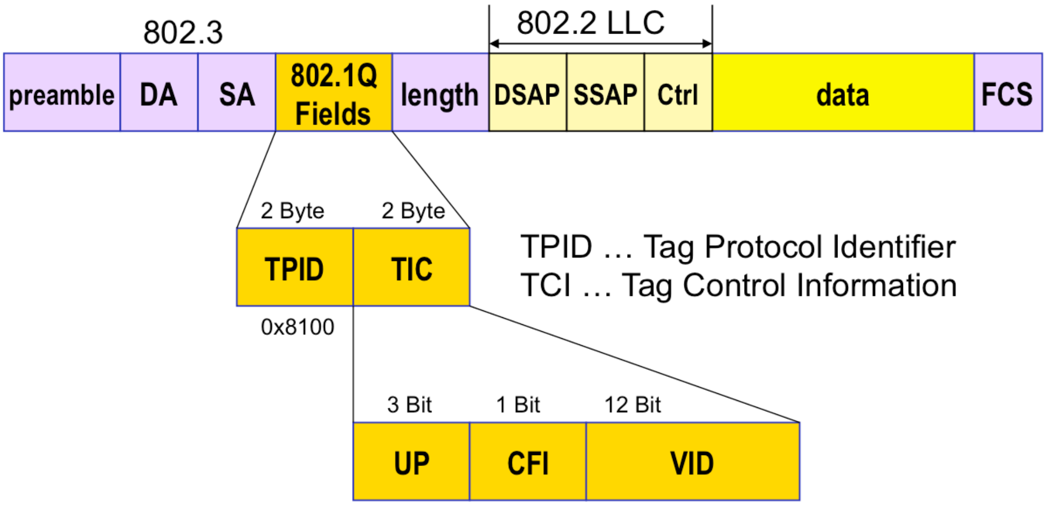


- **Packets across the VLAN trunk are tagged**
  - Either using 802.1Q or ISL tag
  - So next bridge is able to constrain frame to same VLAN as the source
- **Inter-VLAN communication not possible**
  - Only IP router can forward inter-VLAN traffic

 Ethernet Evolution, v6.0 30

By using VLAN tagging the "next" bridge knows whether the source address is also member of the same VLAN.

802.1Q VLAN Tagging – LLC (1)

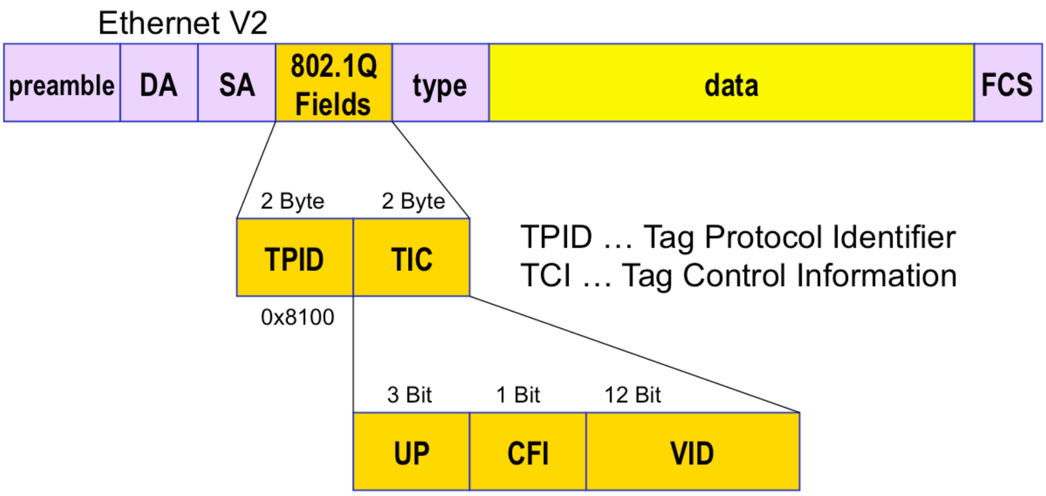


TPID … Tag Protocol Identifier
TCI … Tag Control Information

note: With tagging Ethernets maximal frame length = 1522, minimal frame length = 68

UP … User Priority for L2 QoS = COS
CFI … Canonical Format Identifier
VID … VLAN Identifier

802.1Q VLAN Tagging – Ev2 (2)

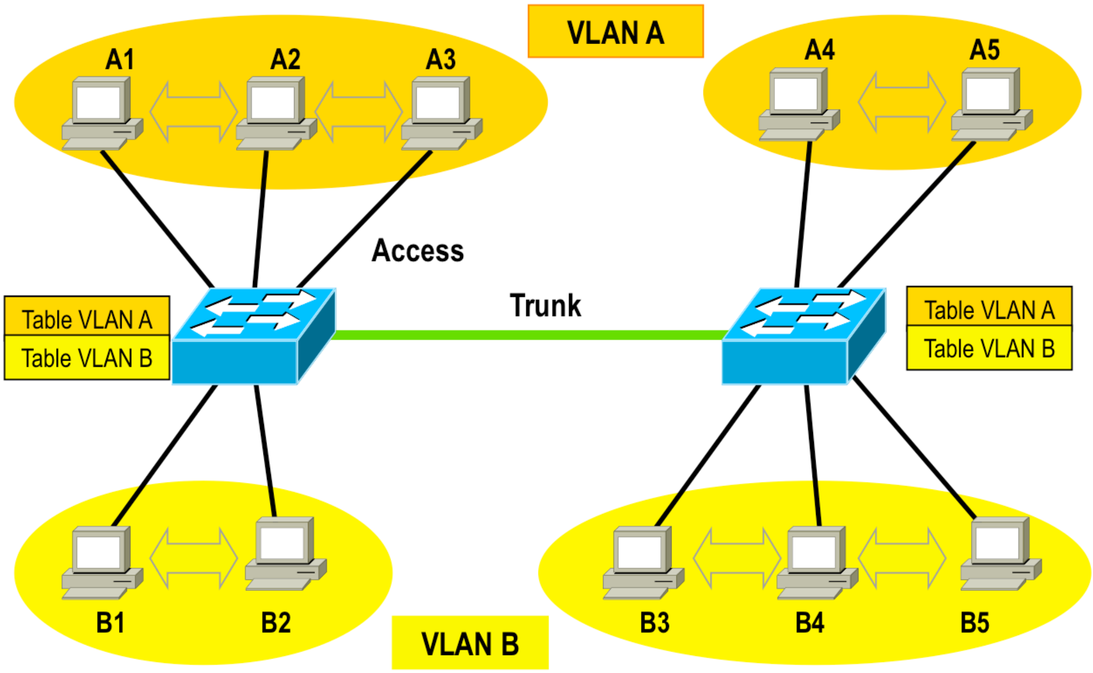


TPID … Tag Protocol Identifier
TCI … Tag Control Information

UP … User Priority for L2 QoS = COS
CFI … Canonical Format Identifier
VID … VLAN Identifier

note: With tagging Ethernets maximal frame length = 1522, minimal frame length = 68

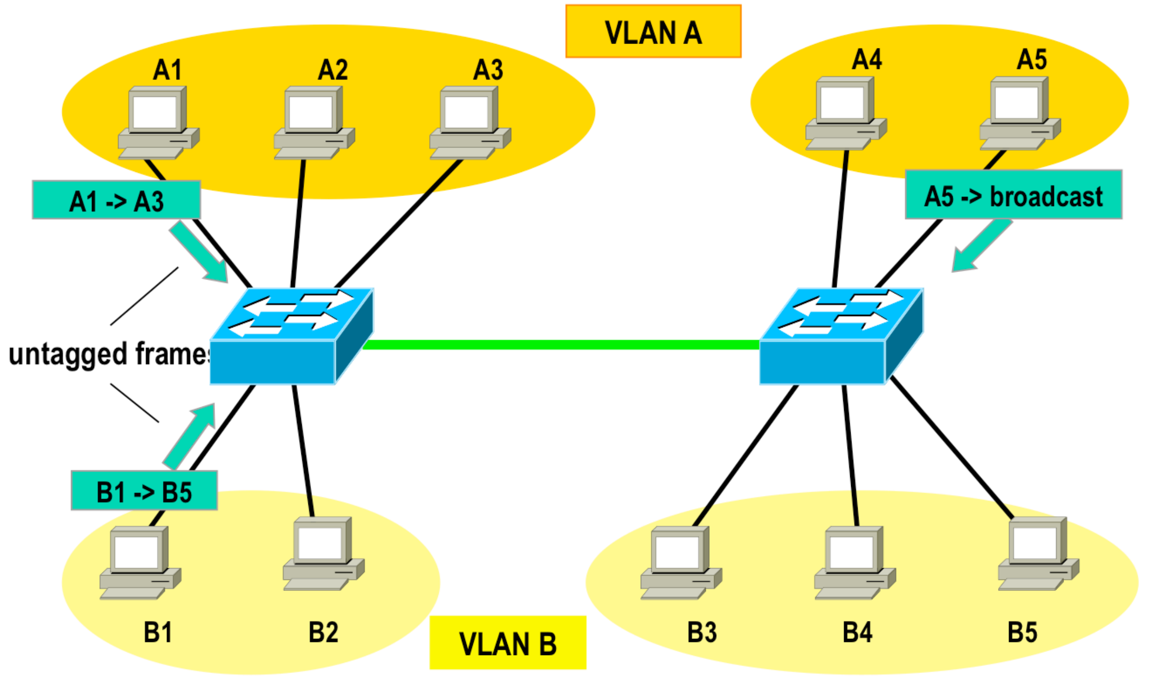
VLAN Trunking Example



 Ethernet Evolution, v6.0 33

L08 - Ethernet Evolution (v6.0)

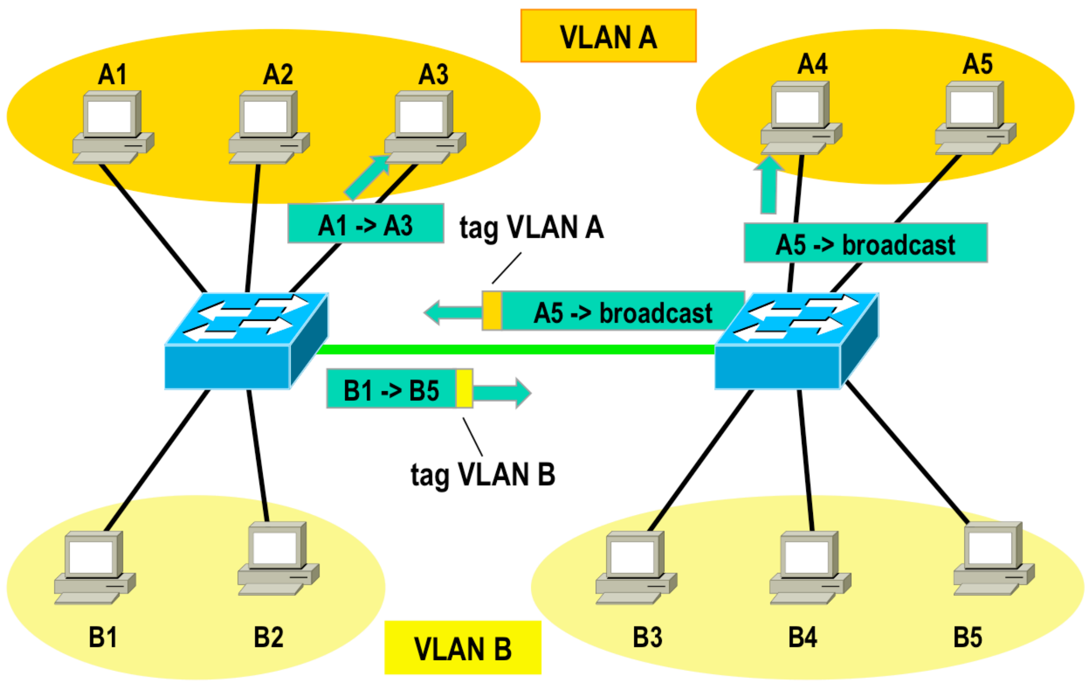
VLAN in Operation (1)



 Ethernet Evolution, v6.0 34

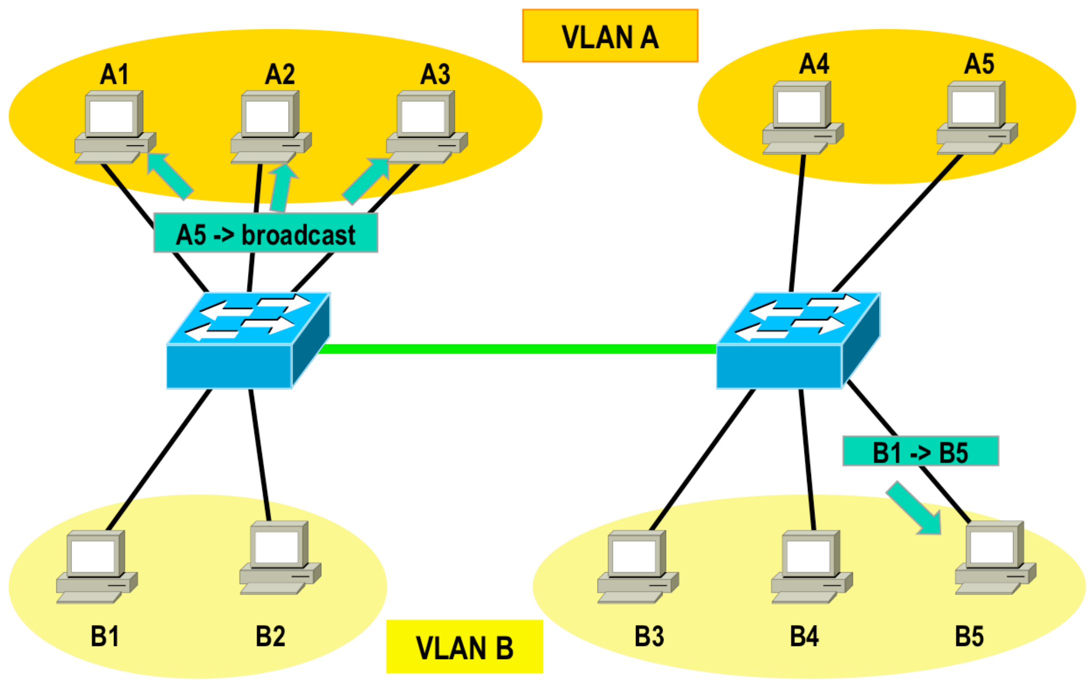
L08 - Ethernet Evolution (v6.0)

VLAN in Operation (2)

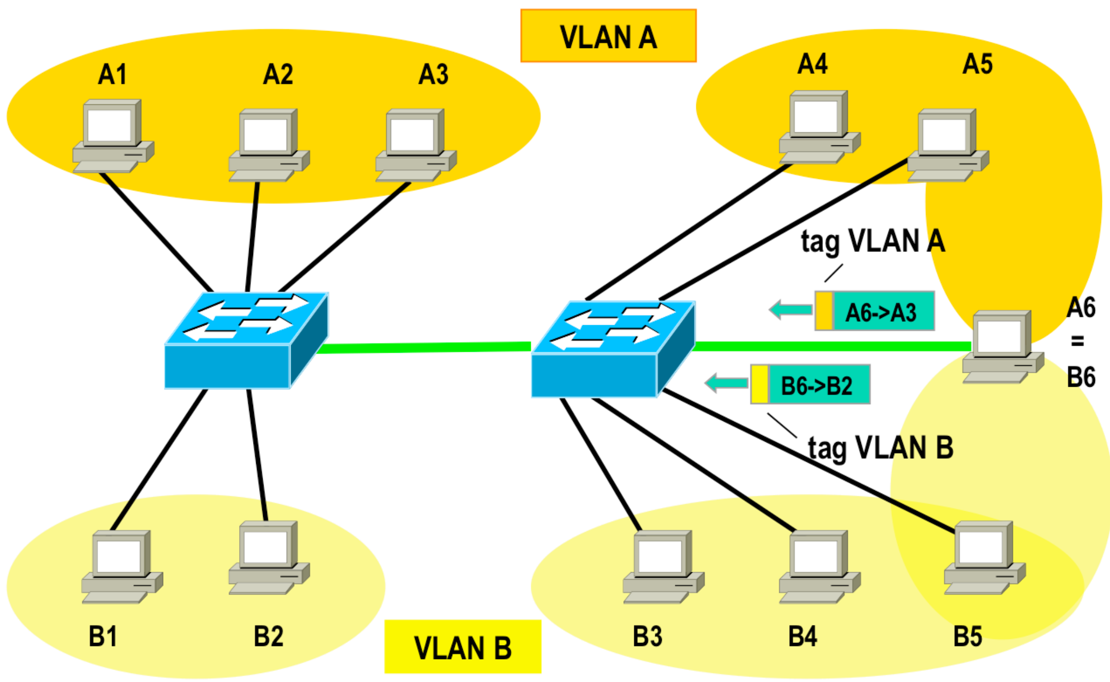


 Ethernet Evolution, v6.0 35

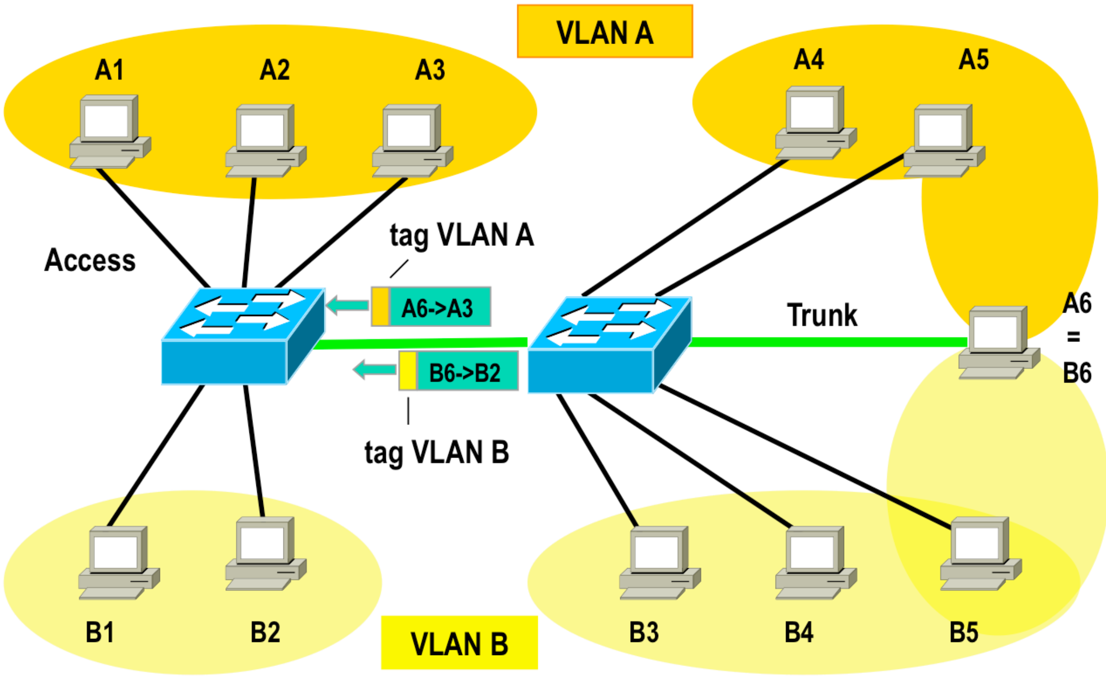
VLAN in Operation (3)



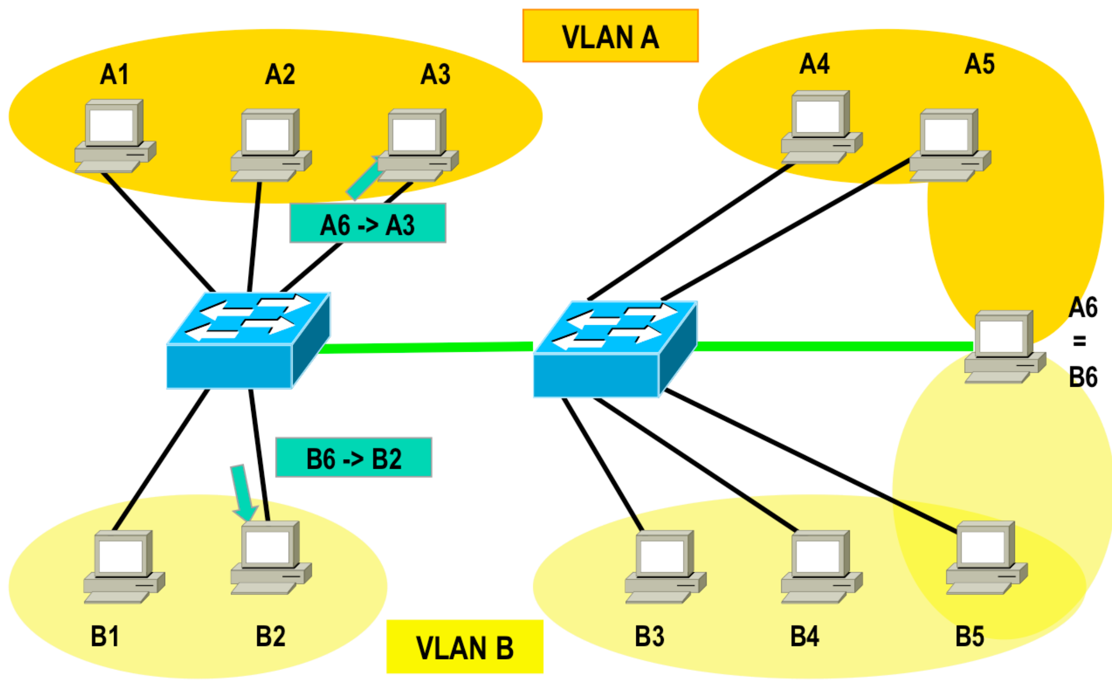
Multihomed VLAN (1)



Multihomed VLAN (2)

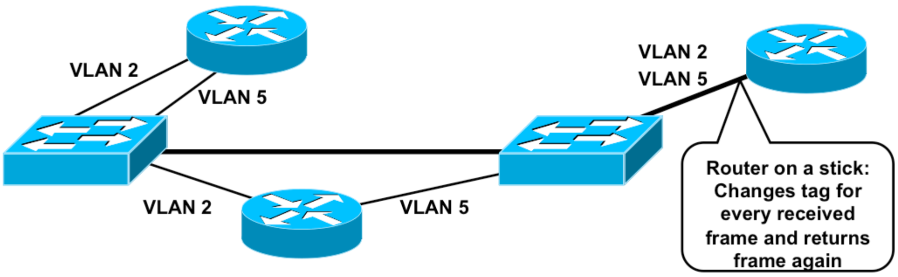


Multihomed VLAN (3)



Inter-VLAN Traffic

- **Router can forward inter-VLAN traffic**
  - Terminates Ethernet links
  - Requirement: Each VLAN in other IP subnet !
- **Two possibilities**
  - Router is member of every VLAN with one link each
  - Router attached on VLAN trunk port ("Router on a stick")



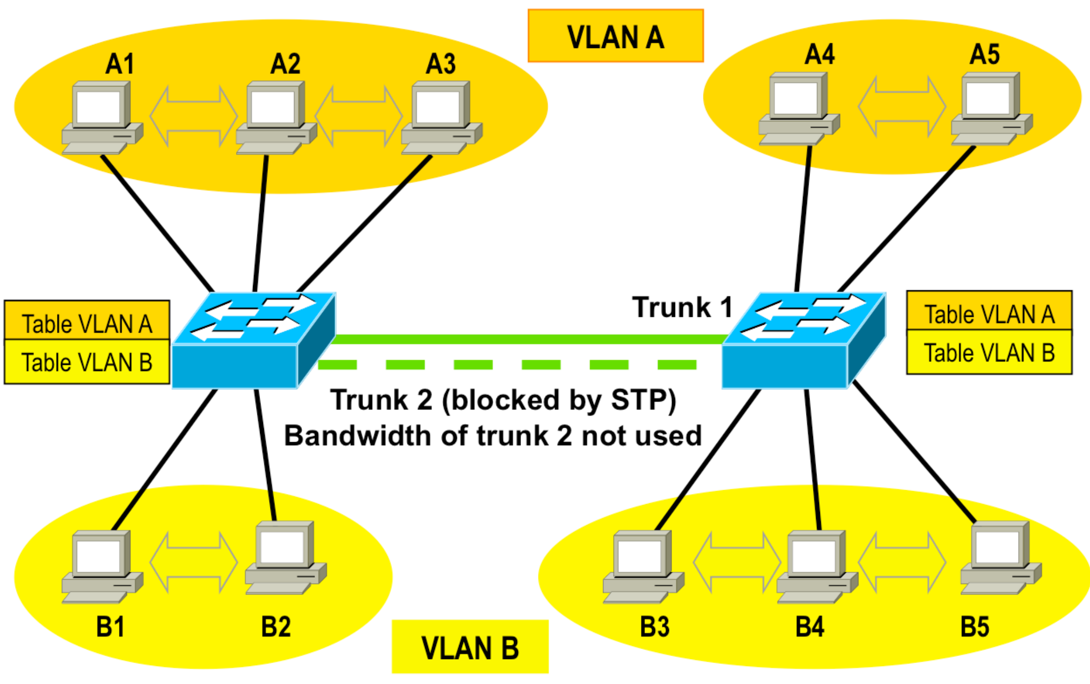
 Ethernet Evolution, v6.0 40

Now we admit the wholly truth: of course it is possible to communicate between different VLANs—using a router! A router terminates layer 2 and is not interested in VLAN constraints. Of course this requires that each VLAN uses another subnet IP address since the router needs to make a routing decision.

There are two possible configurations: The straightforward solutions is to attach a router to several ports on one or more switches, provided that each port is member of another VLAN.

Another method is the "Router on a stick" configuration, employing only a single attachment to a trunk port of a switch. This method saves ports (and cables) but requires trunking functionality on the router. Here the router simply changes the tag of each frame (after making a routing decision) and sends the frame back to the switch.

Trunking without LACP / FEC / GEC



 Ethernet Evolution, v6.0 41

On trunks between multiport switches full duplex operation is used of course. In case of parallel trunks the normal operation of STP will block one trunk link and hence bandwidth of this link can not be used.

Several techniques were developed by vendors and IEEE standardization to allow load balancing on a session-base in such a situation, meaning both trunks can be used for traffic forwarding.

Bundling (aggregation) of physical links to one logical link – which is seen by STP - can be done with:

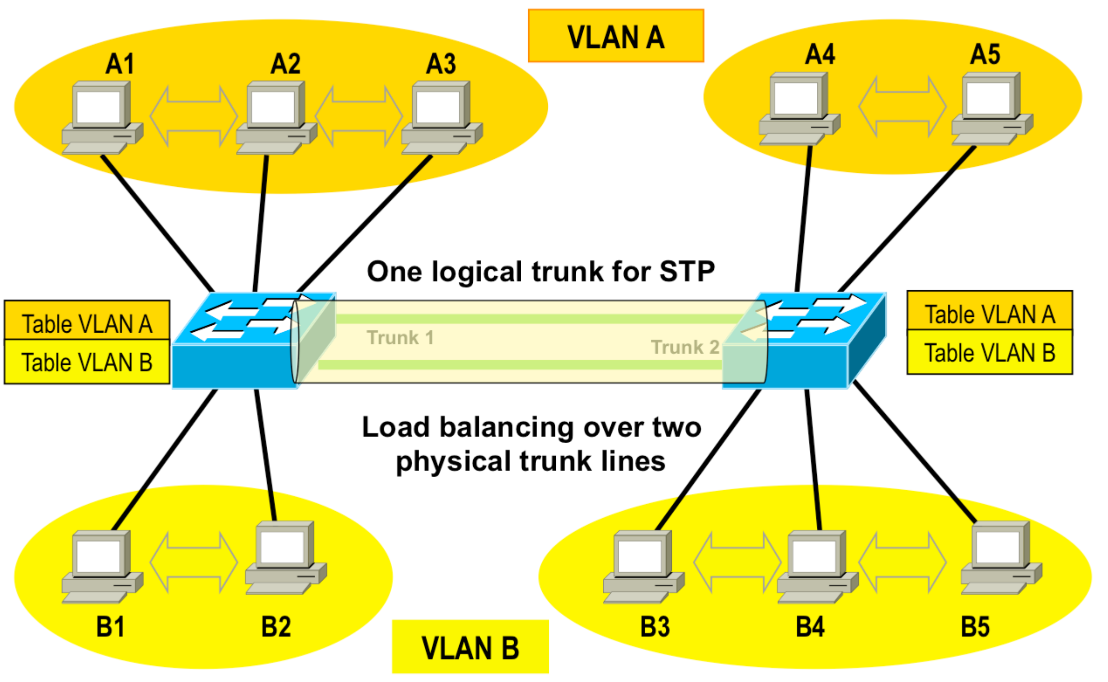
1. Fast Ethernet Channeling (FEC, Cisco), up to eight active ports can be bundled.
2. Gigabit Ethernet Channeling (GEC Cisco), up to eight active ports can be bundled.
3. Linux Bonding.
4. IEEE 802.1AX 2008 LACP (Link Aggregation Control Protocol), up to eight active ports can be bundled.

Note1: LACP appeared first in IEEE 802.3 – version 2002, nowadays handled in a separate standard IEEE 802.1AX-2008)

Note2: LACP is defined between switch and switch or end station and one switch but not between end-system and two switches. Although some vendor have proprietary solutions which allows two physical switches acting as one logical switch so that LACP can also be used between an end-system and two physical switches.

Of course, if a per-VLAN STP is used like in PVST+ or multiple instances of STP are possible like in MSTP, then by STP-tuning of VLAN orange to use trunk 1 and STP-tuning of VLAN yellow to use trunk 2 a alternate method exiists for solving that problem.

Trunking with LACP / FEC / GEC



 Ethernet Evolution, v6.0 42

Agenda

- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

© 2016, D.I. Lindner / D.I. Haas

Ethernet Switching

- **Ethernet switches can connect end systems with 10 Mbit/s, 100 Mbit/s or 1000 Mbit/s for example**
  - Clients may using 100 Mbit/s and server may use 1000 Mbit/s using a full duplex, point-to-point connection with switch.
  - Note: multiport repeater is not able to do this!
  - Ethernet frame has not changed!
- **It is still connectionless packet switching on L2 based**
  - Asynchronous TDM principle, buffers
  - Flow control would be great
    - Modern switches can avoid congestion can by supporting a new MAC control frame (so called pause command)

Ethernet MAC frame format was preserved up nowadays. Bridging from 10 Mbit/s Ethernet to 100 Mbit/s Ethernet does not require a bridge to change the frame format. (Remark: bridging from 10 Mbit/s Ethernet to FDDI (100 Mbit/s Token ring) requires frame format changing which makes it slower). Therefore Ethernet L2 switches can connect Ethernets with 10 Mbit/s, 100 Mbit/s or 1000 Mbit/s easily and fast.

IEEE 802.3 (2008)

- **The latest version specifies**
  - Operation for 10 Mbit/s, 100 Mbit/s, Gigabit/s and 10Gigabit/sEthernet over copper and fiber
  - Full duplex Ethernet
  - Auto-negotiation
  - Flow control
- **It is still backward compatible to the old times of Ethernet**
  - CSMA/CD (half-duplex) operation in 100 and 1000 Mbit/s Ethernets with multiport repeater possible
  - Frame bursting or carrier extension for ensuring slot-time demands in 1000 Mbit/s Ethernet
  - 10Gigabit/s Ethernet is full duplex only
    - CSMA/CD has died!!!
  - Ethernet frame is identical across all speeds

 Ethernet Evolution, v6.0 45

Note: Full-duplex mode is possible on point-to-point links between two elements in the network (end-system to switch, switch to switch, end-system to end-system) if there are two physical communication paths available (2 fiber optic links or 4 copper wires used for symmetrical transmission). Now CSMA/CD is not in necessary and can be switched off. A station can send frames immediately (without CS) using the transmit-line of the cable and simultaneously receive data on the other line. At both end of the link we have store and forward behavior hence collision detection (CD) is not necessary anymore.

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 (such as frame bursting and carrier extension) had been specified in the standard to support length up to 200m, no vendor developed an GE hub as for today. Thus: Forget GE Hubs!

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.

Even if 1Gbit/s Ethernet is backward compatible (CSMA/CD capable) with initial 10 Mbit/s shared media idea, a multiport repeater with Gigabit Ethernet seems absurd because bandwidth sharing decreases performance; every collision produces an additional delay for a crossing packet. Therefore nobody uses it as shared media!

Flow Control

- **Speed-requirements for switches are very high**
  - Especially in full duplex operation also powerful switches can not avoid buffer overflow
  - Earlier, high traffic caused collisions and CSMA/CD interrupted the transmission in these situations, now high traffic is normal
- **L4 flow control (e.g. TCP) between end-systems is not efficient enough for a LAN**
  - switches should be involved to avoid buffer overflow
- **Therefore a MAC based (L2) flow control is specified**
  - MAC control frame with the Pause command

MAC Control Frame

- Identified among other frames by setting length field = 8808 hex

always 64 octets

8 octets	6	6	2	2	44	4
preamble	DA	SA	8808h	MAC-ctrl opcode	MAC-ctrl parameters	FCS
			(Length)			

MAC ctrl opcode .... defines function of control frame

MAC- trl parameters .... control parameter data; always filled up to 44 bytes, by using zero bytes if necessary

- Currently only the "pause" function is available (opcode 0x0001)

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.

The Pause Command 1

- **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)

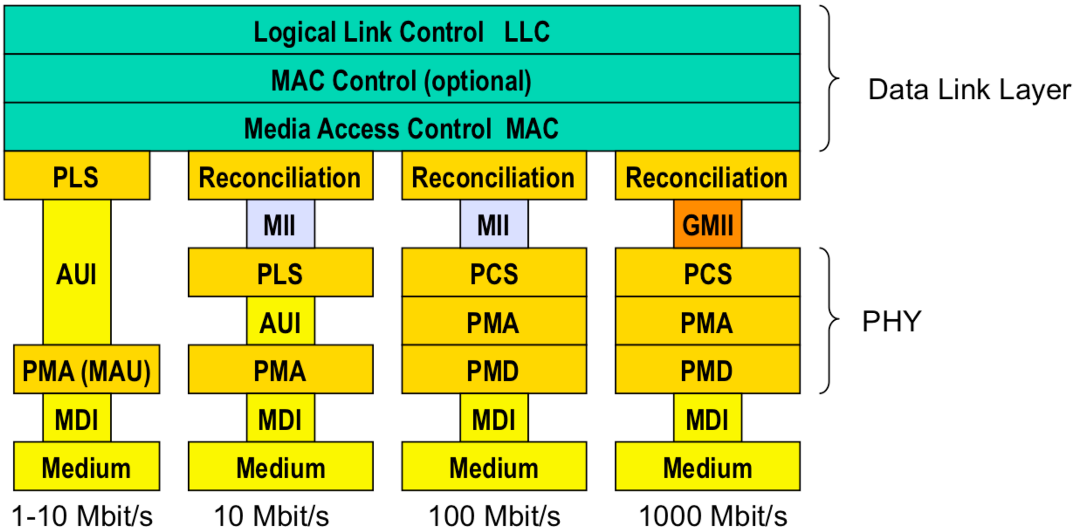
The Pause Command 2

- **Destination address is either**
  - Address of destination station or
  - Broadcast address or
  - Special multicast address 01-80-C2-00-00-01
- **The 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

Demand for Higher Speed

- **Higher data rates need more sophisticated coding**
  - 10 Mbit/s Ethernet: Manchester coding
  - Fast Ethernet (100 Mbit/s): 4B/5B block code
  - Gigabit Ethernet 1000 Mbit/s): 8B/10B block code
- **New implementations should be backwards-compatible**
  - Old physical layer signaling interface (PLS), represented by AUI, was not suitable for new coding technologies
- **AUI has been replaced**
  - MII (Media Independent Interface) for Fast Ethernet
  - GMII for Gigabit Ethernet

New Physical Sublayers



AUI...Attachment Unit Interface, PLS...Physical Layer Signaling, MDI...Medium Dependent Interface, PCS...Physical Coding Sublayer, MII...Media Independent Interface, GMII...Gigabit Media Independent Interface, PMA...Physical Medium Attachment, MAU...Medium Attachment Unit, PMD...Physical Medium Dependent

PHY Sublayers (1)

- **Physical Layer Signaling (PLS) serves as abstraction layer between MAC and PHY**
- **PLS provides**
  - Data encoding/decoding (Manchester)
  - Translation between MAC and PHY
  - Attachment Unit Interface (AUI) to connect with PMA
- **Several new coding techniques demands for a Media Independent Interface (MII)**
- **Today coding is done through an media-dependent Physical Coding Sublayer (PCS) below the MII**

PHY Sublayers (2)

- **PLS has been replaced with the Reconciliation sublayer**
  - Reconciliation layer transforms old MAC PLS-primitives into MII control signals
- **MII / GMII serves as an interface between MAC and PHY**
  - Hides coding issues from the MAC layer
  - MII: often a mechanical connector for a wire; GMII is an interface specification between MAC-chip and PHY-chip upon a circuit board
  - One independent specification for all physical media
  - Supports several data rates (10/100/1000 Mbits/s)
  - MII:4 bit / GMII: 8 bit parallel transmission channels to the physical layer

PHY Sublayers (3)

- **Physical Coding Sublayer (PCS)**
  - Encapsulates MAC-frame between special PCS delimiters
  - E.g. 4B/5B or 8B/10B encoding
  - Appends idle symbols
- **Physical Medium Attachment (PMA)**
  - Interface between PCS and PMD
  - (de) serializes data for PMD (PCS)
- **Physical Medium Dependent (PMD)**
  - Serial transmission of the code groups
  - Specification of the various connectors (MDI)

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), FEC, GEC
- **Trend moves towards layer 3 switching**
  - High amount of today's traffic goes beyond the border of the LAN
  - Routing stop broadcast domains, enable load balancing and decrease network traffic
  - "**Route if you can bridge if you must**"

Note: Spanning Tree regards this as one logical link! => Load balancing!

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".

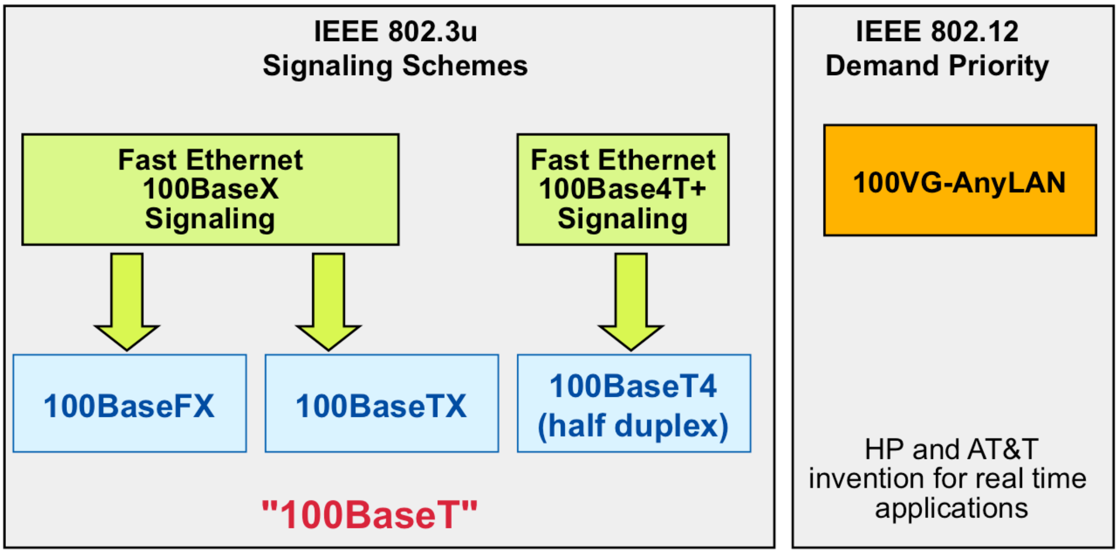
Agenda

- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

100 Mbit/s Ethernet

- **Access method disagreement split 100 Mbit/s LAN development into two branches:**
  - Fast Ethernet - IEEE-802.3u (today 802.3-2008)
  - 100VG-AnyLAN - IEEE-802.12 (disappeared)
- **Fast Ethernet was designed as 100 Mbit/s and backwards-compatible 10Mbit/s Ethernet**
  - CSMA/CD but also
  - Full-duplex connections (collision free)
- **Network diameter based on collision window requirement (512 bit times)**
  - Reduced by factor 10
  - e.g. 250m compared with 2500m at 10 Mbit/s

100 Mbit Ethernet Overview



 Ethernet Evolution, v6.0 59

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-µm multimode fiber-optic cables.

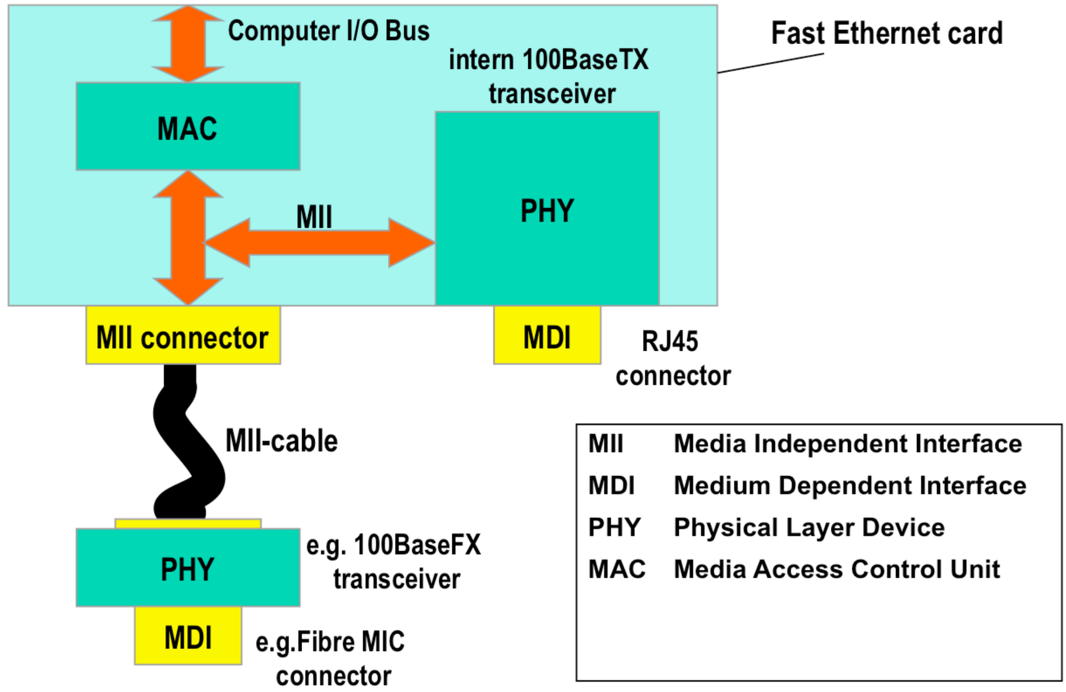
100Base4T+ signaling has been specified to support 100 Mbit/s over Cat3 cables. This mode allows half duplex operation only and uses a 8B6T code over 4 pairs of wires; one pair for collision detection, three pairs for data transmission. One unidirectional pair is used for sending only and two bi-directional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

Fast Ethernet

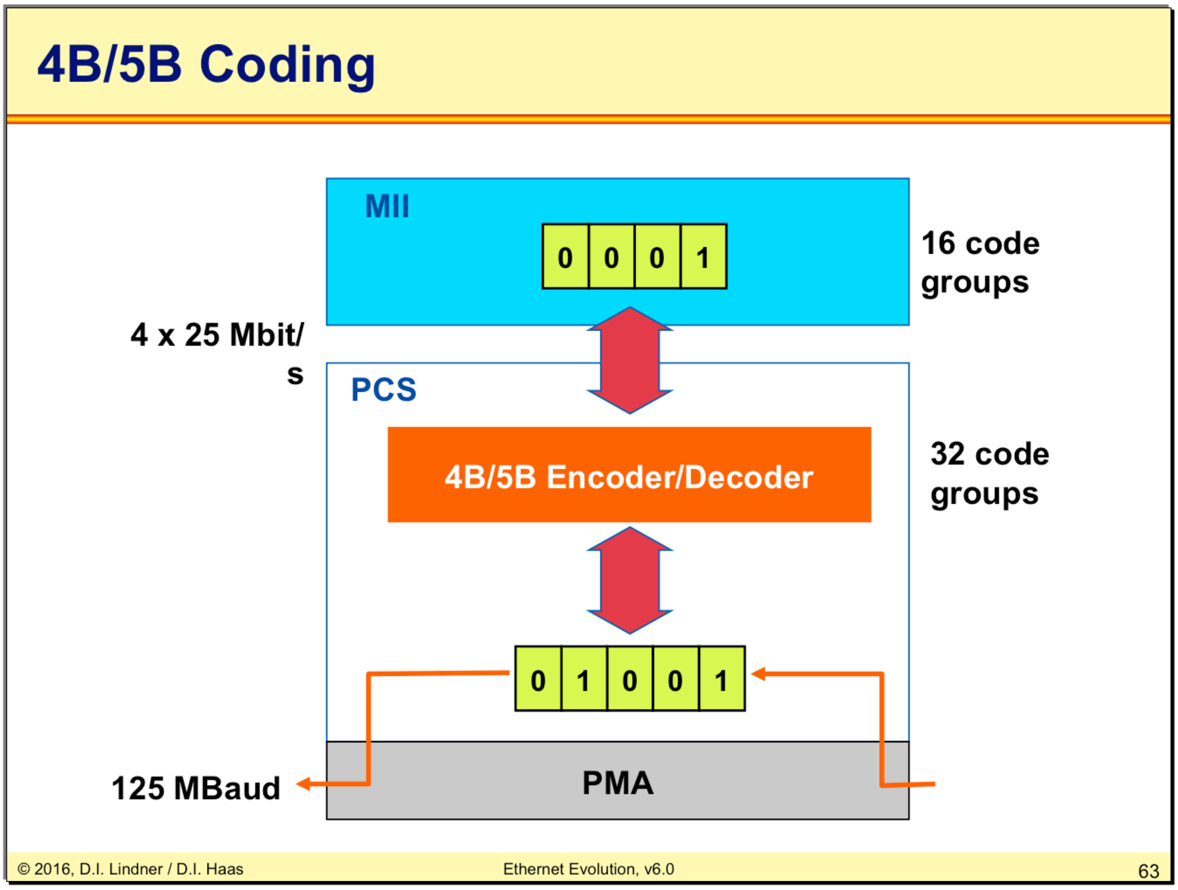
- **AUI has been replaced with the Media Independent Interface (MII)**
  - New coding (4B/5B, 8B/6T, PAM 5x5) and bandwidth constrains demand for a redesigned abstraction layer
- **MII defines a generic 100BaseT interface**
  - Allows utilization of a 100BaseTX, 100BaseFX, 100BaseT4 or a 100BaseT2 transceiver
    - On-board or cable-connector with
    - 20 shielded, symmetrically twisted wire pairs -> 40 poles
    - One additional main-shield
    - 68 Ohm impedance; 2.5 ns maximal delay
    - 50 cm maximal length

Typical Fashion



100BaseX Signaling – 4B/5BCoding

- **Each 4-bit group is encoded by a 5 bit run-length limited "code-group"**
  - Code groups lean upon FDDI-4B/5B codes
  - Some additional code groups are used for signaling purposes; remaining code groups are violation symbols
    -> easy error detection
  - Groups determinate maximal number of transmitted zeros or ones in a row -> easy clock synchronization
    - Keeps DC component below 10%
- **Code groups are transmitted using NRZI-encoding**
  - Code efficiency: 4/5 = 100/125 = 80% (Manchestercode only 50 %)



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 → 12.5 %.

L08 - Ethernet Evolution (v6.0)

Code Group Table

	PCS code-group	name	MII group	
DATA	11110	0	0000	
	01001	1	0001	
	10100	2	0010	
	10101	3	0011	
	01010	4	0100	
	01011	5	0101	
	01110	6	0110	
	01111	7	0111	
	10010	8	1000	
	10011	9	1001	
	10110	A	1010	
	10111	B	1011	
	11010	C	1100	
	11011	D	1101	
	11100	E	1110	
	11101	F	1111	
Control	11111	I	undefined	Idle pattern between streams
	11000	J	0101	Start of Stream Delimiter (1st part)
	10001	K	0101	Start of Stream Delimiter (2nd part)
	01101	T	undefined	End of Stream Delimiter (1st part)
	00111	R	undefined	End of Stream Delimiter (2nd part)
	00100	H	undefined	signals receiving errors

Remaining code groups are not valid (triggers error detection)

Recognize: Start Delimiter (SD) and End Delimiter (ED) for frame synchronization of an Ethernet frame are coded in control code-points -> that are code-violations in this context. Real data is coded DATA code-points.

Autonegotiation (1)

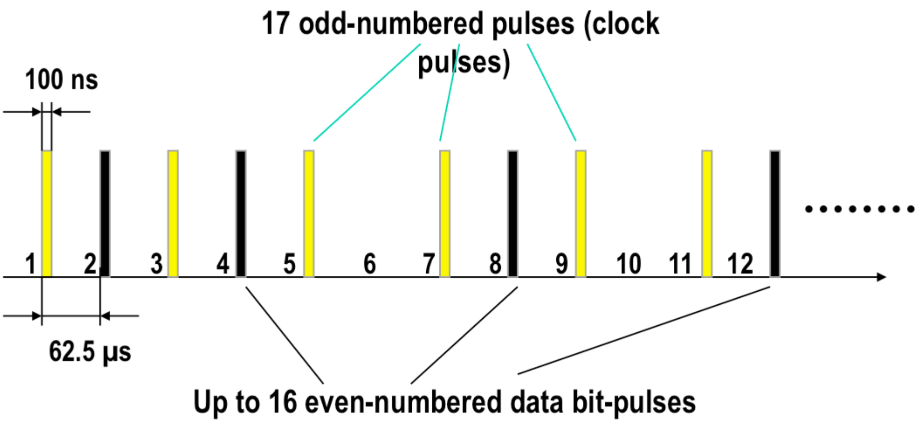
- **Enables each two Ethernet devices to exchange information about their capabilities**
  - Signal rate, CSMA/CD, half- or full-duplex
- **Modern Ethernet NICs send bursts of so called**
  - Fast-Link-Pulses (FLP) for autonegotiation signaling
  - Each FLP burst represents a 16 bit word
- **FLP**
  - Consists of 17-33 so called Normal-Link-Pulses (NLPs)
  - NLP are used for testing link-integrity
    - 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.

FLP Burst Coding



= 1 1 0 1 0 1 ....

 Ethernet Evolution, v6.0 66

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".

Autonegotiation (2)

- **FLP-bursts are only sent on connection-establishments**
- **100BaseT stations recognizes 10 Mbit/s stations by receiving a single NLP only**
- **Two 100BaseT stations analyze their FLP-bursts and investigate their largest common set of features**
- **Last frames are sent 3 times -> other station responds with acknowledge-bit set**
- **Negotiated messages are sent 6-8 times**
  - FLP- session stops here

Autonegotiation (3)

- **The first FLP-burst contains the base-link codeword**
- **By setting the NP bit a sender can transmit several "next-pages"**
  - Next-pages contain additional information about the vendor, device-type and other technical data
- **Two kinds of next-pages**
  - Message-pages (predefined codewords)
  - Unformatted-pages (vendor-defined codewords)
- **After reaching the last acknowledgement of this FLP-session, the negotiated link-codeword is sent 6-8 times**

Base Page

S0	S1	S2	S3	S4	A0	A1	A2	A3	A4	A5	A6	A7	RF	Ack	NP

Selector field (S0–S4) — Technology ability field (A0–A7)

Selector 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

Remote Fault (RF)

Signals that the remote station has recognized an error

Next Page (NP)

Signals following next-page(s) after the base-page

Acknowledge (Ack)

Signals the receiving of the data (not the feasibility)

If the base-page has been received 3 times with the NP set to zero, the receiver station responds with the Ack bit set to 1

If next-pages are following, the receiver responds with Ack=1 after receiving 3 FLP-bursts

Next-Pages Codeword

M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	T	Ack 2	MP	Ack	NP

Message code field (M0–M10)

Examples:

10000000000 ....null message, station has no further information to send

01000000000 ....next page contains technology ability information

10100000000 ....next 4 pages contain Organizationally Unique Identifier (OUI) information

U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	T	Ack 2	MP	Ack	NP

Unformatted code field (U0–U10)

Acknowledge 2 (Ack2)

Ack2 is set to 1 if station can perform the declared capabilities

Message Page (MP)

Differentiates between message-pages (MP=1) and

Unformatted-pages (MP=0)

Toggle (T)

Provides synchronization during exchange of next-pages information

T-bit is always set to the inverted value of the 11th bit of the last received link-codeword

Signaling Types (1)

- **Three signaling types :**
  - 100BaseX:
    - refers to either the 100BaseTX or 100BaseFX specification
  - 100BaseT4+
- **100BaseX**
  - combines the CSMA/CD MAC with the FDDI Physical Medium Dependent layer (PMD)
  - 4B/5B code
  - allows full duplex operation on link

Signaling Types (2)

- **100BaseT4+**
  - allows half duplex operation only
  - 8B6T code
  - Uses 4 pairs of wires; one pair for collision detection, three pair for data transmission
  - One unidirectional pair is used for sending only and two bi-directional pairs for both sending and receiving
  - Same pinout as 10BaseT specification
  - Transmit on pin 1 and 2, receive on 3 and 6; bi-directional on 4 and 5; bi-directional on 7 and 8

100BaseTX and 100BaseFX

- **100BaseTX:**
  - 125 MBaud symbol rate, full duplex, binary encoding
  - 2 pair Cat 5 unshielded twisted pair (UTP) or 2 pair STP or type 1 STP
  - RJ45 connector; same pinout as in 10BaseT (transmit on 1 and 2, receive on 3 and 6)
- **100BaseFX:**
  - 125 MBaud symbol rate, full duplex, binary encoding
  - Two-strand (transmit and receive) 50/125 or 62.5/125-µm multimode fiber-optic cable
  - SC connector, straight-tip (ST) connector, or media independent connector (MIC)

100BaseT4

- **100BaseT4:**
  - 25 MBaud, half duplex, ternary encoding
  - Cat3 or better, needs all 4 pairs installed
  - 200 m maximal network diameter
  - maximal 2 hubs

Agenda

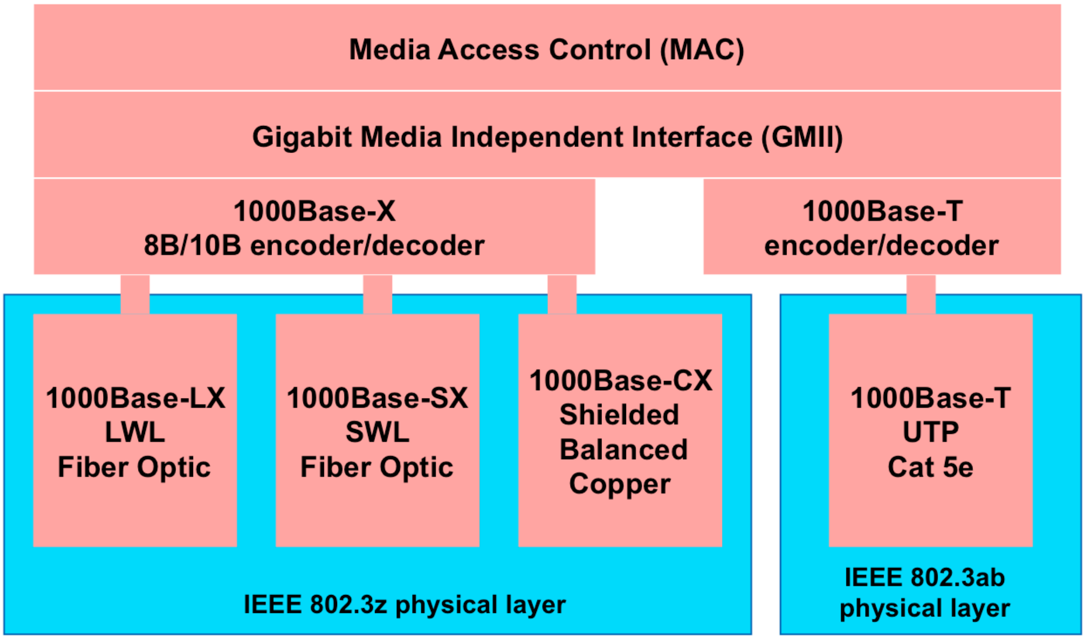
- Ethernet Evolution
- VLAN
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

 Ethernet Evolution, v6.0 75

Gigabit-Ethernet: IEEE-802.3z / IEEE802.3ab

- **Easy integration into existing 802.3 LAN configurations**
  - Because of backward compatibility
  - Access methods: CSMA/CD or full duplex
  - Autonegotiation and flow control
  - IEEE-802.3z/802.3ab are now part of IEEE 802.3-2008
- **Backbone technology**
  - GE link as WAN transmission technique
  - Reaches 70 km length using fibre optics

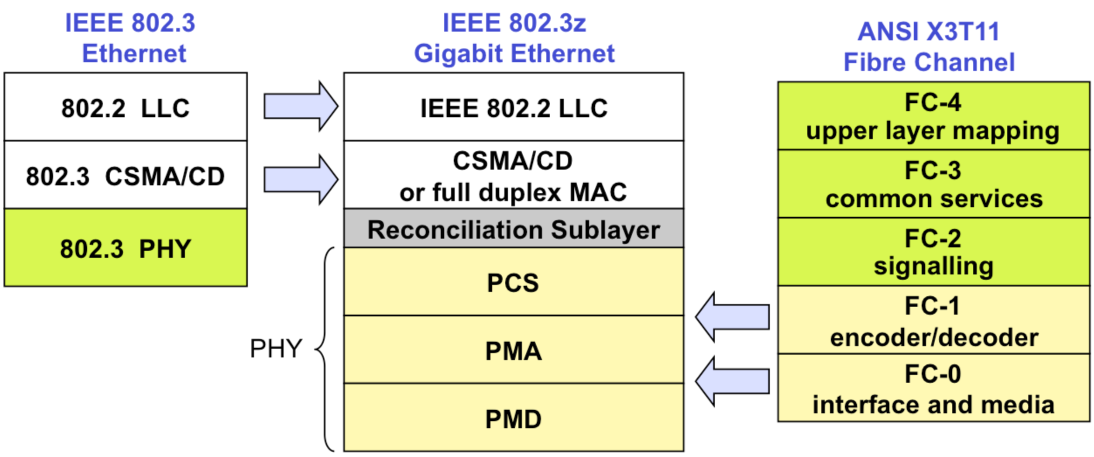
Gigabit Ethernet



 Ethernet Evolution, v6.0 77

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.

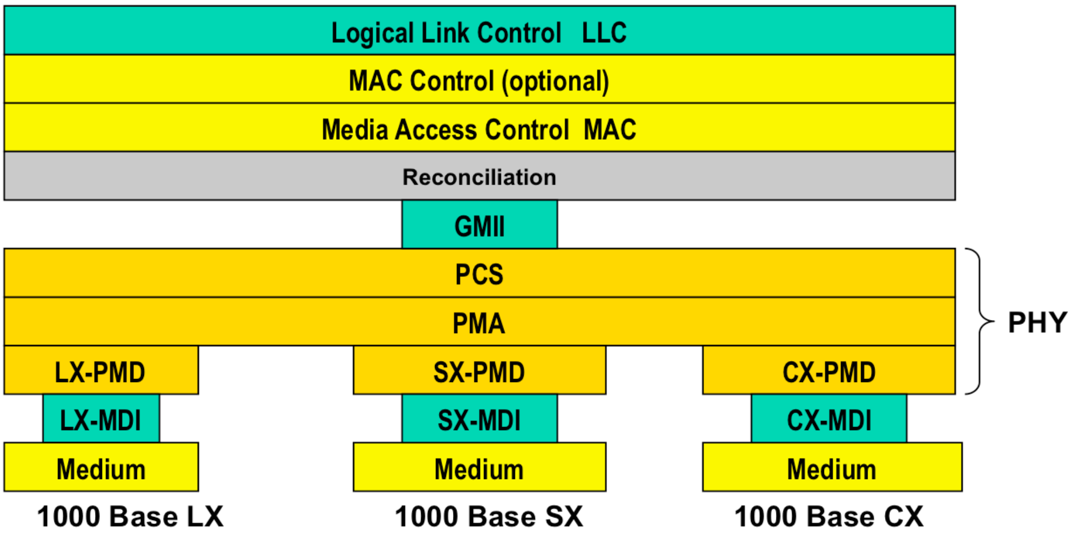
GE Signaling



 Ethernet Evolution, v6.0 78

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.

Physical Sublayers



CSMA/CD Restrictions (Half Duplex Mode)

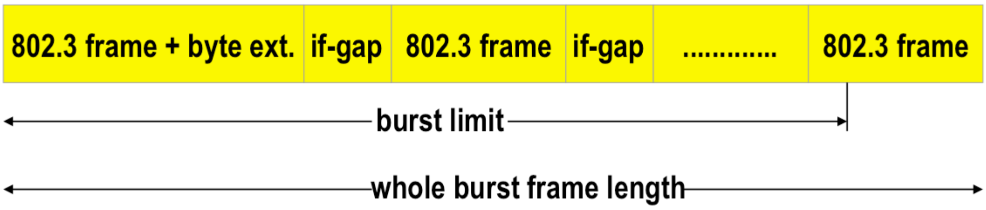
- **The conventional collision detection mechanism CSMA/CD**
  - Limits the net diameter to 20m!
- **Solutions to increase the maximal net expansion:**
  - Carrier Extension:
    - Extension bytes appended to (and removed from) the Ethernet frame by the physical layer
    - Frame exists a longer period 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

 Ethernet Evolution, v6.0 80

Remember: CSMA/CD requires that stations have to listen (CS) twice the signal propagation time to detect collisions. A collision window of 512 bit times at a rate of 1Gbit/s limits the maximal net expansion to 20m!

Frame Bursting

- **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 µs = 12 bytes

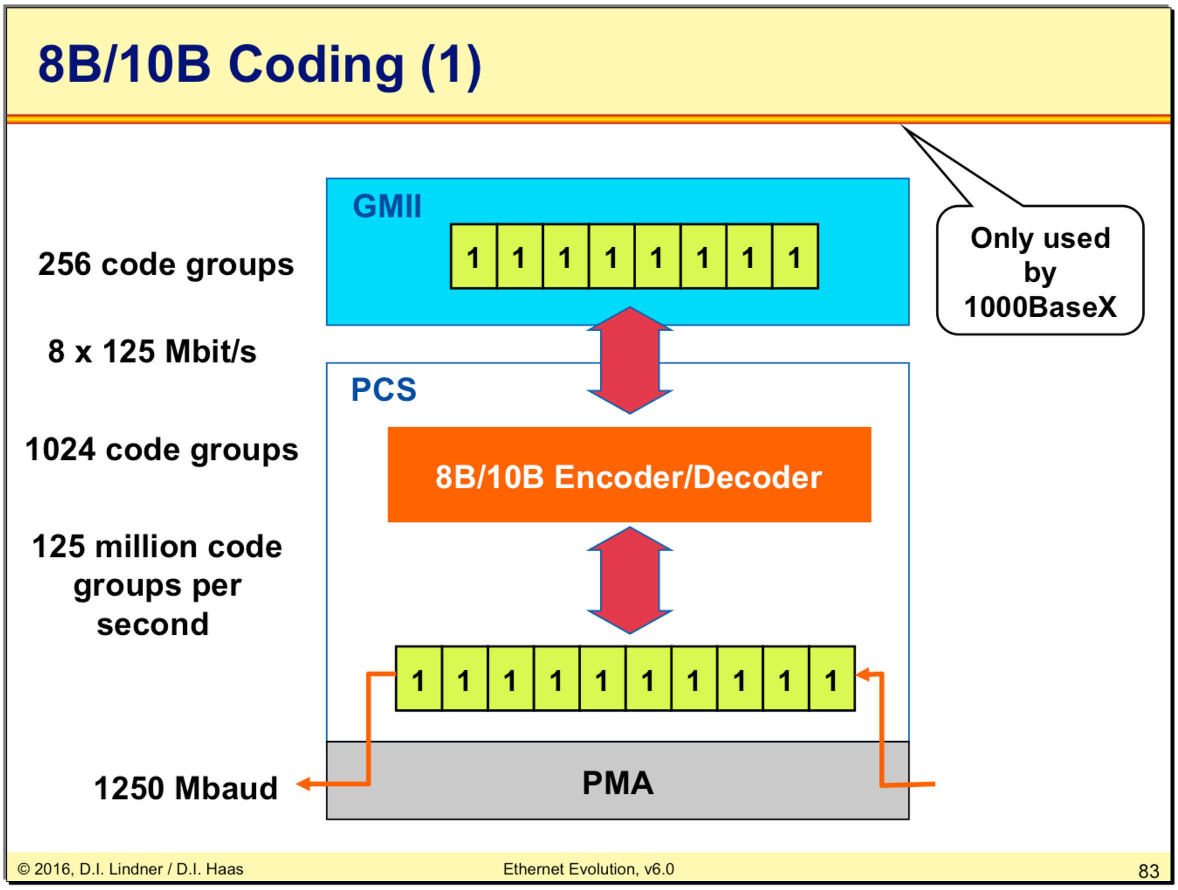


 Ethernet Evolution, v6.0 81

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)

1000BaseX Coding

- **8B/10B block encoding: each 8-bit group encoded by a 10 bit "code-group" (symbol)**
  - Half of the code-group space is used for data transfer
  - Some code groups are used for signaling purposes
  - Remaining code groups are violation symbols
    - -> Easy error detection
  - Groups determine the maximal number of transmitted zeros or ones in a 10 bit symbol
    - -> Easy clock signal detection (bit synchronization)
  - No baselinewander (DC balanced)
    - lacking DC balance would result in data-dependent heating of lasers which increases the error rate
  - Code efficiency: 8/10 = 1000/1250 = 80%



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.

8B/10B Coding (2)

- **Each GMII 8 bit group (data) can be represented by an associated pair of 10 bit code groups**
  - Each pair has exactly 10 ones and 10 zeros in sum
- **Sender toggles Running Disparity flag (RD) to remember which code group to be sent for the next data-octet**
- **Hence, only non-symmetric code groups need a compensating code group**
  - symmetric code groups already have equal number of ones and zeros

8B/10B Coding (3)

- **Code groups which are not registered in the code-table are considered as code-violation**
  - These code groups are selected to enable detection of line errors with high probability
- **256 data and 12 control code-group-pairs are defined**
- **Control-code-groups are used independently or in combination with data-code-groups**
  - Autonegotiation is performed by using such control-code-groups
  - Therefore no kind of NLP or FLP bursts for that purpose

8B/10B Coding (4)

- **Control-code-groups are classified by "ordered sets" after their usage:**
  - Configuration C for autonegotiation
  - Idle I used between packets
  - Encapsulation:

    R for separating burst frames

    S as start of packet delimiter

    T as end of packet delimiter

    V for error propagation

1000BaseX

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

 Ethernet Evolution, v6.0 87

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

- **1000Base-T defined by 802.3ab task force**
  - UTP uses all 4 line pairs simultaneously for duplex transmission!
    - Using echo-cancelling: receiver subtracts own signal
  - 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
    - note: collision domains should be avoided

 Ethernet Evolution, v6.0 88

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.

Autonegotiation

- **Both 1000Base-X and 1000Base-T provide autonegotiation functions to determinate the**
  - Access mode (full duplex - half duplex)
  - Flow control mode
- **Additionally 1000Base-T can resolve the data rate**
  - Backward-compatibility with 10 Mbit/s and 100 Mbit/s
  - Also using FLP-burst sessions

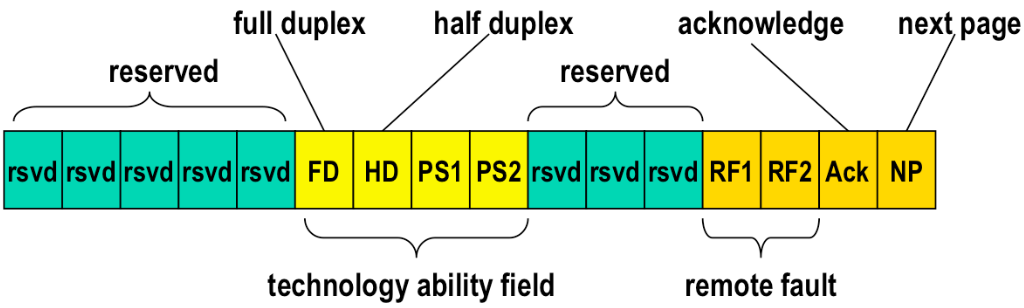
 Ethernet Evolution, v6.0 89

1000BaseX Autonegotiation

- **1000Base-X autonegotiation uses normal (1000Base-X) signaling !**
  - "Ordered sets" of the 8B/10B code groups
  - No fast link pulses !
    - Autonegotiation had never been specified for traditional fiber-based Ethernet
    - So there is no need for backwards-compatibility
- **1000Base-X does not negotiate the data rate !**
    - Only gigabit speeds possible
- **1000Base-X autonegotiation resolves**
  - Half-duplex versus full-duplex operation
  - Flow control

Autonegotiation is part of the Physical Coding sublayer (PCS).
Content of base-page register is transmitted via ordered set /C/.
On receiving the same packet three times in a row the stations replies with the Ack -bit set.
Next-pages can be announced via the next-page bit NP.

Base-Page



PS1	PS2	description
0	0	no pause
0	1	asymmetrical pause
1	0	symmetrical pause
1	1	symmetrical and asymmetrical pause

RF1	RF2	description
0	0	no error
0	1	offline
1	0	connection error
1	1	autonegotiation error (no common capabilities)

Next-Pages

Normal message-page (predefined codes)

M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	T	Ack 2	MP	Ack	NP

Message code field (M0–M10)

Vendor specific page (non predefined codes)

U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	T	Ack 2	MP	Ack	NP

Unformatted code field (U0–U10)

 Ethernet Evolution, v6.0 92

Acknowledge 2 (Ack2)
- Ack2 is set to 1 if station can perform the declared capabilities

Message Page (MP)
- Differentiates between message-pages (MP=1) and
- Unformatted-pages (MP=0)

Toggle (T)
- Provides synchronization during exchange of next-pages information
- T-bit is always set to the inverted value of the 11th bit of the last received link-codeword

1000BaseT Autonegotiation

- Autonegotiation is only triggered when the station is powered on
- At first the stations expects Gigabit-Ethernet negotiation packets (replies)
- If none of them can be received, the 100Base-T fast link pulse technique is tried
- At last the station tries to detect 10Base-T stations using normal link pulses

Agenda

- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

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

 Ethernet Evolution, v6.0 95

Originally the 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. On fiber its difficult to deal with asynchronous transmission, photons cannot be buffered easily, store and forward problems

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

As a WAN technology 10GE is much simpler than ATM (hopefully cheaper) but of course it can not be compared with cell switching based on store and forward and sophisticated QoS support.

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 defined in 2006**
  - 10GBASE-T
  - CAT6 UTP cabling with maximum distance of 55m to 100m
  - CAT7 cabling with maximum distance of 100m
- **Nowadays 802.3ab, 802.3ak, 802.3an are covered by the IEEE 802.3-2008 document**

 Ethernet Evolution, v6.0 99

GE and 10GE over copper is a challenge because of radiation/EMI, grounding problems, high BER, thick cable bundles (especially Cat-7).

Often 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.

Think about that before you use GE or 10GE over copper!