L23 - The Ethernet Evolution

The Ethernet Evolution

From 10Mbit/s to 10Gigabit/s Ethernet Technology
From Bridging to L2 Ethernet Switching and VLANs

Agenda

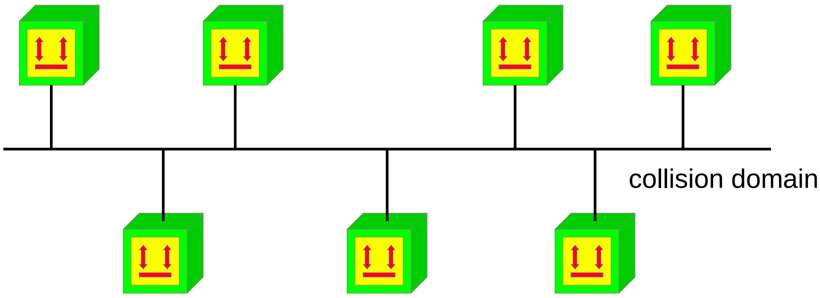
- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 2

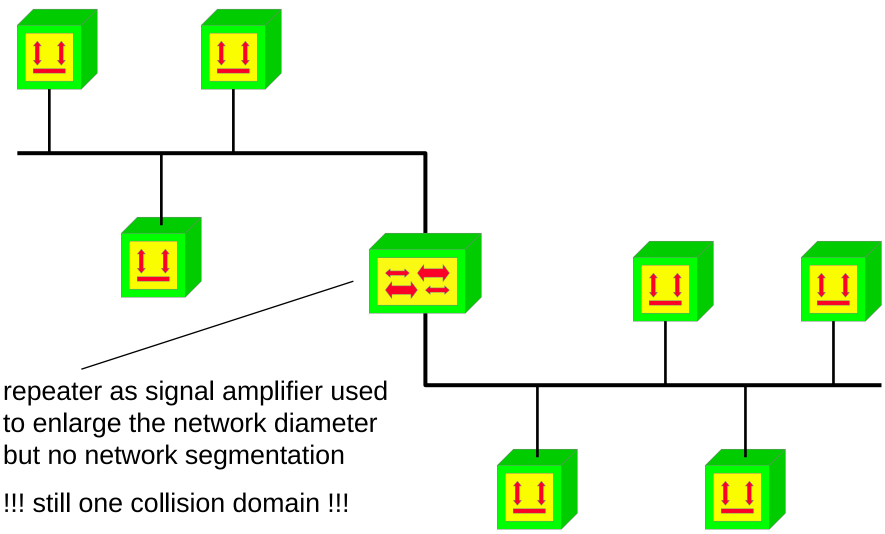
© 2011, D.I. Manfred Lindner

The Beginning

- **initial idea: shared media LAN**
  - bus structure, CSMA/CD was access method
  - coax cable, transmission rate up to 10 Mbit/s
  - half-duplex transmission (two physical wires e.g. coax)



Enlarging the Network

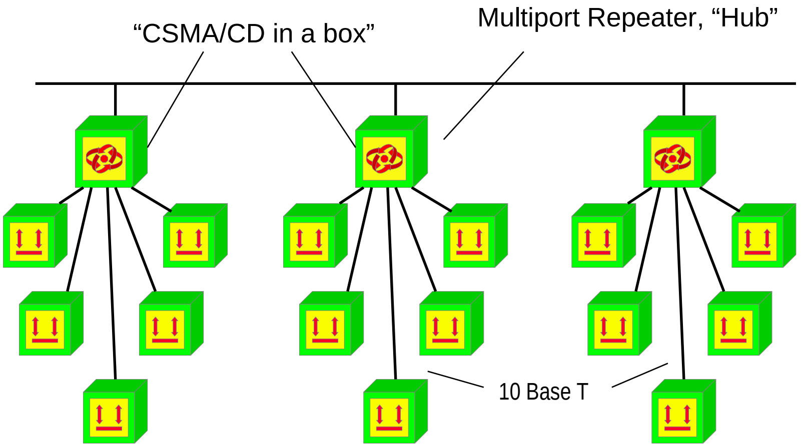


© 2011, D.I. Manfred Lindner

Multiport Repeater

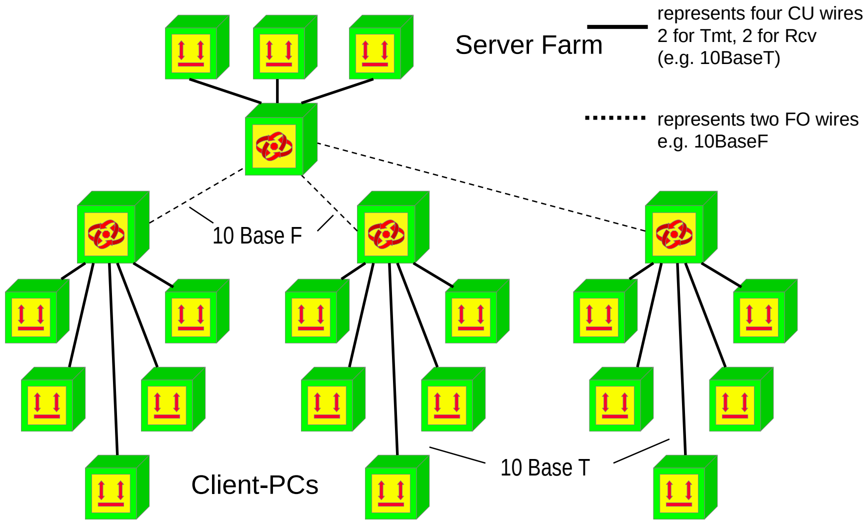
- **demand for telephony-like point-to-point cabling using Twisted Pair wires**
  - based on structured cabling standard
  - 10BaseT as new Ethernet type to support this demand
  - four physical wires (2 for tmt, 2 for rcv)
- **network stations are connected star-like to a multiport repeater**
  - multiport repeater is called “hub”
- **hub simulates the bus: "CSMA/CD in a box"**
- **only half-duplex**
  - only one network station can use the network at a given time, all others have to wait

Structured Cabling (1)



© 2011, D.I. Manfred Lindner

Structured Cabling (2)



© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 7

Bridging

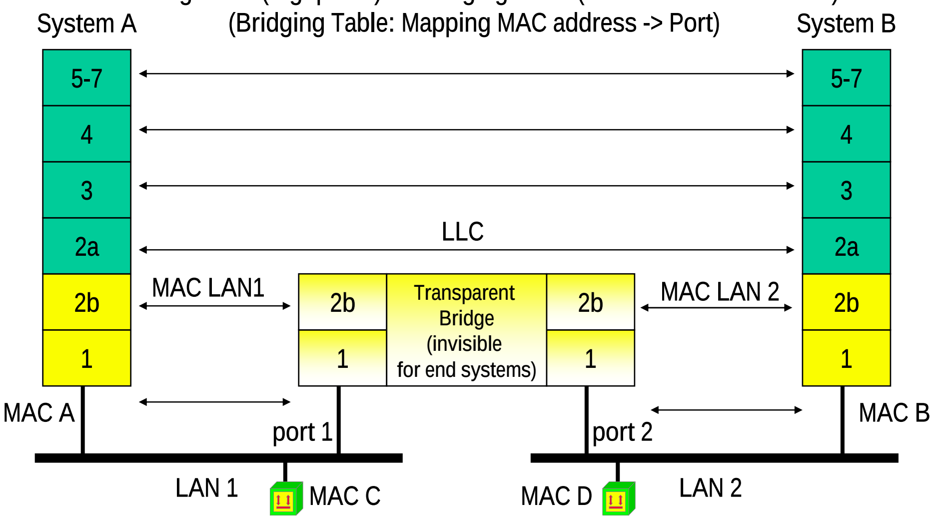
- **simple physical amplification with repeaters became insufficient**
  - with repeaters all nodes share the given bandwidth
  - the whole network is still one collision domain
  - -> technology moved toward layer 2
- **bridges segment a network into smaller collision domains**
  - store and forward technology (packet switching)
  - the whole network is still a broadcast domain
  - Spanning Tree provides a unique path between each two devices and avoids broadcast storms

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 8

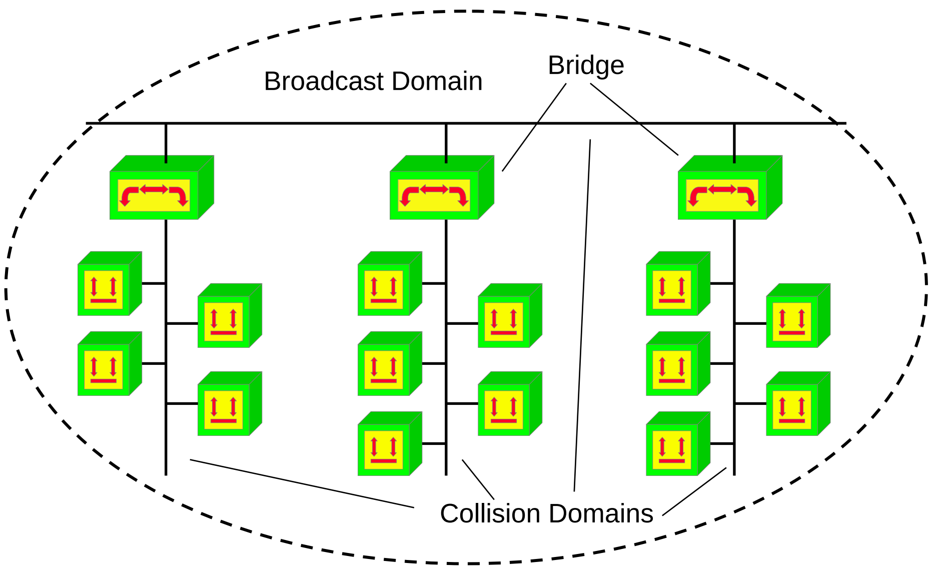
© 2011, D.I. Manfred Lindner

Transparent Bridge = Ethernet Switch

Packet Switching (PS) in Connectionless Service Mode on OSI Layer 2
Routing Table (Signposts) –> Bridging Table (= Ethernet Switch Table)
(Bridging Table: Mapping MAC address -> Port)



Network Segmentation with Bridges



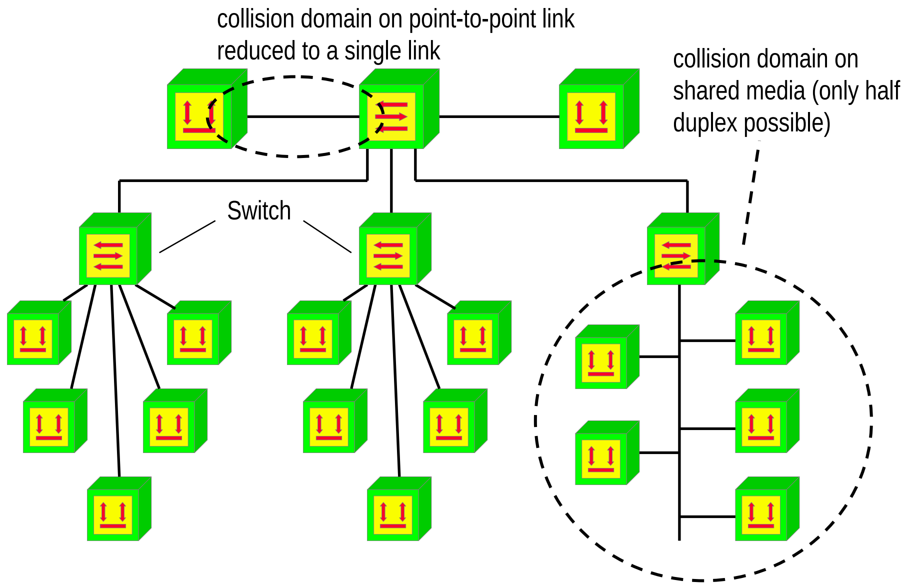
© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

Ethernet Switching (1)

- **"switching" means fast transparent bridging**
  - implemented in hardware
  - also called Layer 2 (L2) switching or Ethernet switching
- **multiport switches allow full duplex operation on point-to-point links**
  - no need for collision detection (media access control) on a link which is shared by two devices only
    - network station <-> switch port
    - switch <-> switch
- **multiport switches replaces multiport repeaters**
  - a collision free Ethernet can be built, if network consists of point-to-point links only

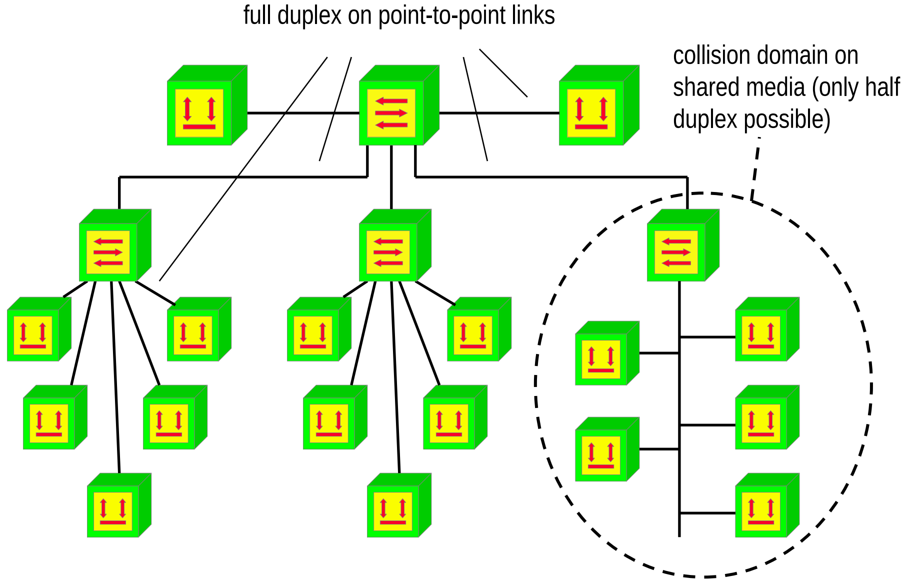
Ethernet Switching (2)



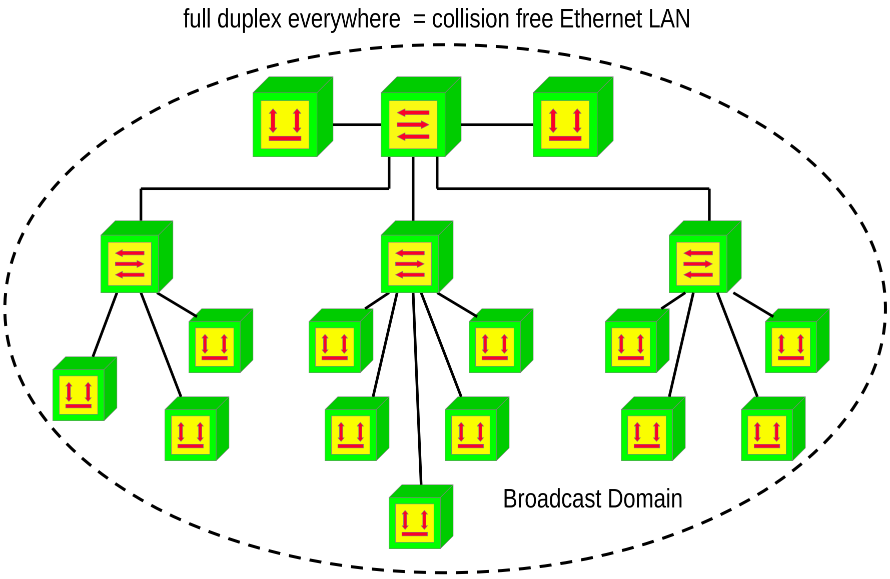
© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

Ethernet Switching (3)



Ethernet Switching (4)



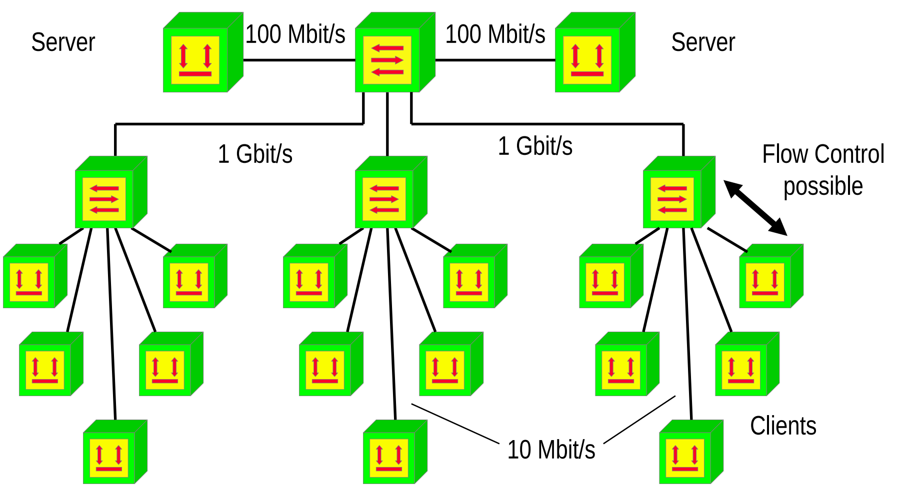
© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

Ethernet Switching (5)

- **L2 switches can connect Ethernets with 10 Mbit/s, 100 Mbit/s or 1000 Mbit/s for example**
  - clients using 10 Mbit/s either half duplex on shared media or full duplex on point-to-point connection with switch
  - server uses 100 Mbit/s, full duplex, point-to-point connection with switch
  - note: multiport repeater is not able to do this !
- **L2 switch as packet switch operates with asynchronous TDM**
  - congestion can be avoided by using a new MAC based flow control (pause command)

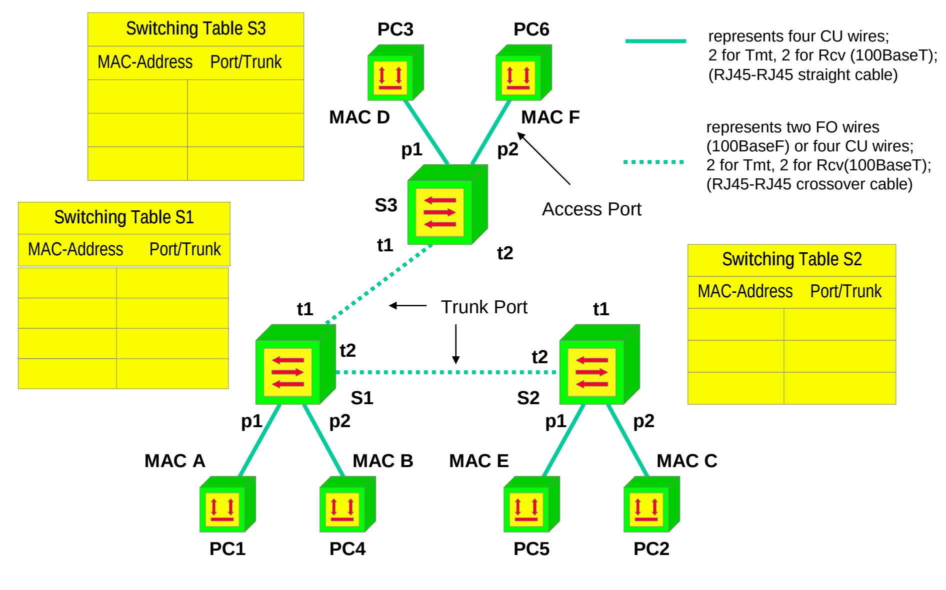
Ethernet Switching (6)



© 2011, D.I. Manfred Lindner

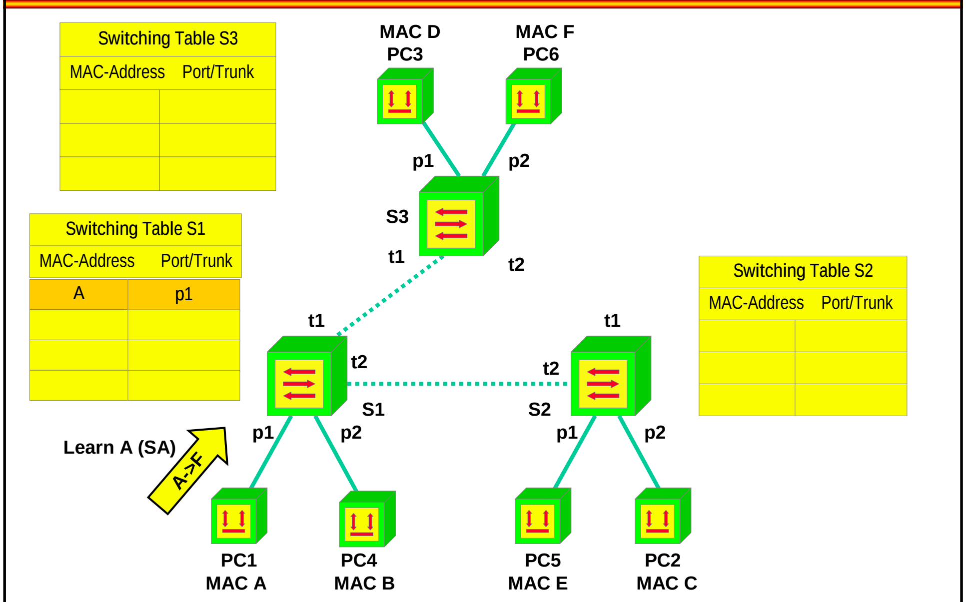
L23 - The Ethernet Evolution

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



© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 17

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



© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 18

© 2011, D.I. Manfred Lindner

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

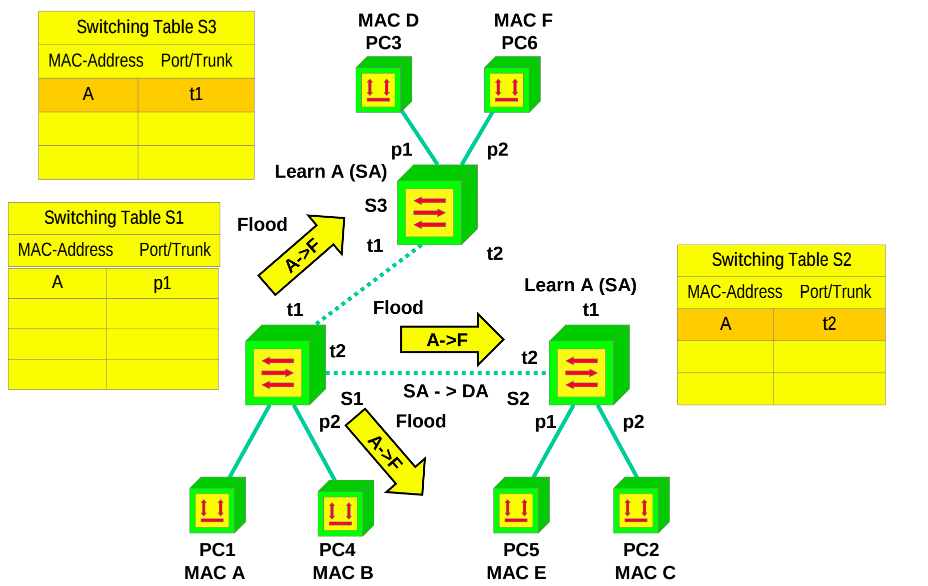
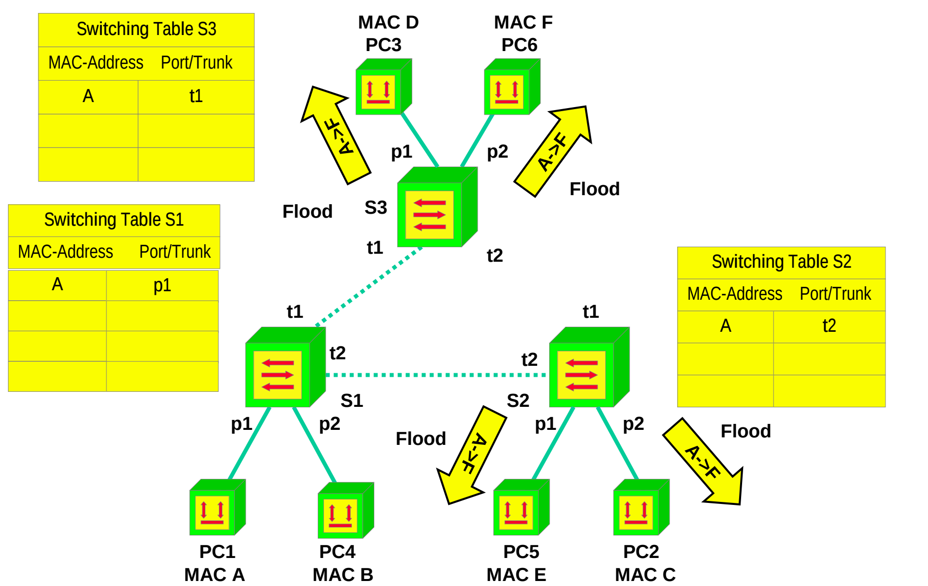


Table Building for Ethernet Frame MAC-A to MAC-F 3



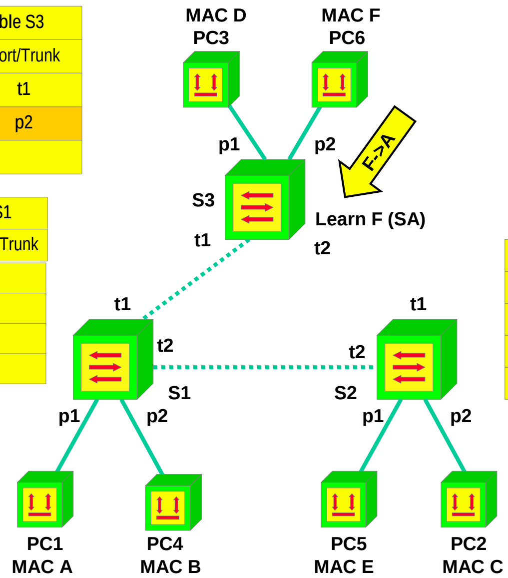
© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

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



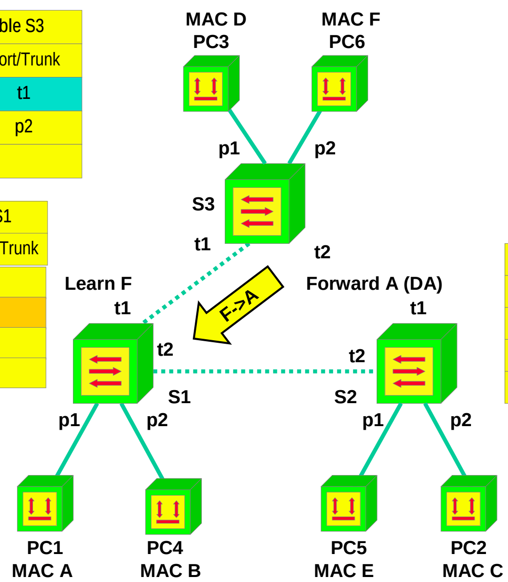
Switching Table S2	
MAC-Address	Port/Trunk
A	t2

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 21

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

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 22

© 2011, D.I. Manfred Lindner

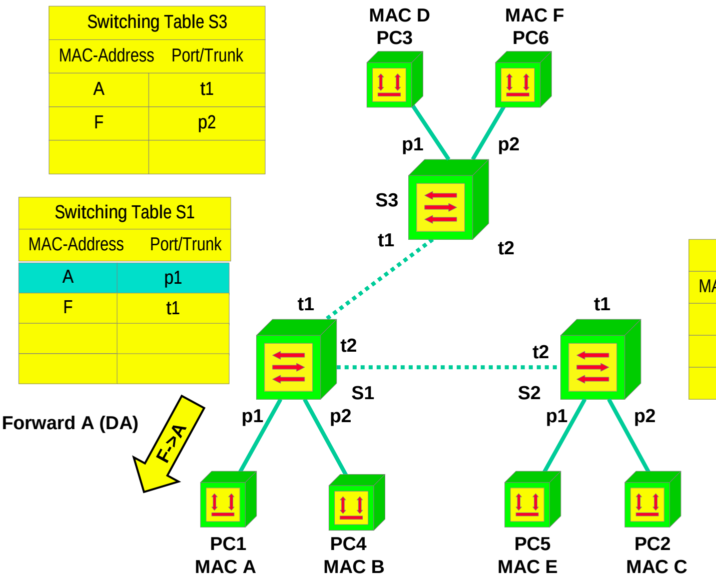
L23 - The Ethernet Evolution

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

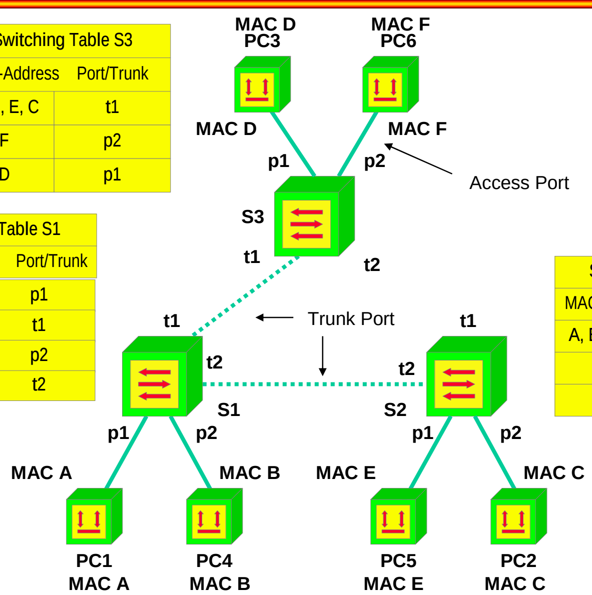


Ethernet Switch Table – Final State
(All MAC addresses learned)

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

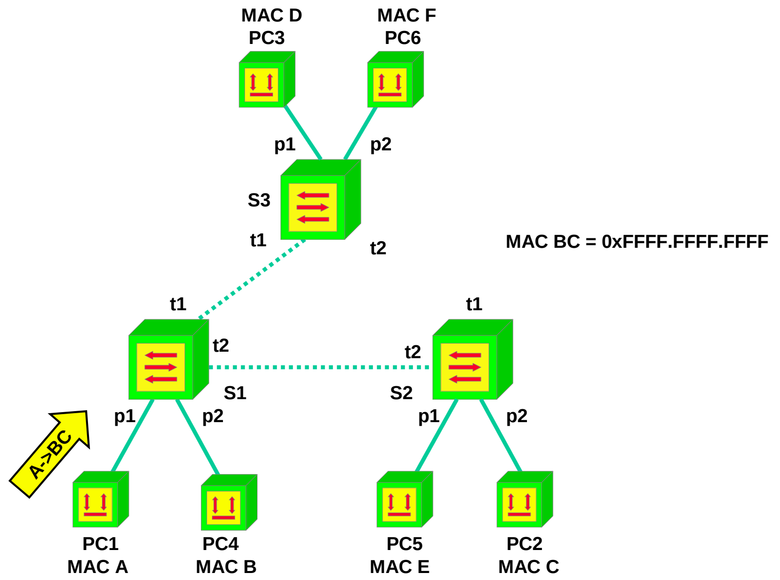
Switching Table S1	
MAC-Address	Port/Trunk
A	p1
F, D	t1
B	p2
E, C	t2

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

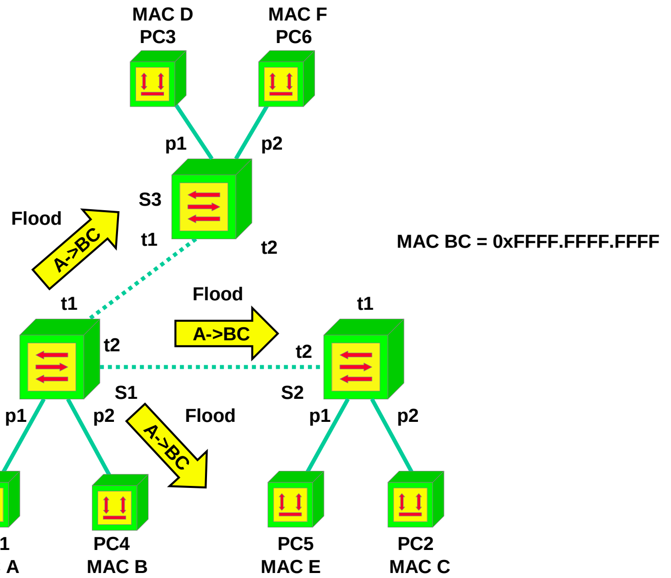


© 2011, D.I. Manfred Lindner

Ethernet Broadcast (BC) 1



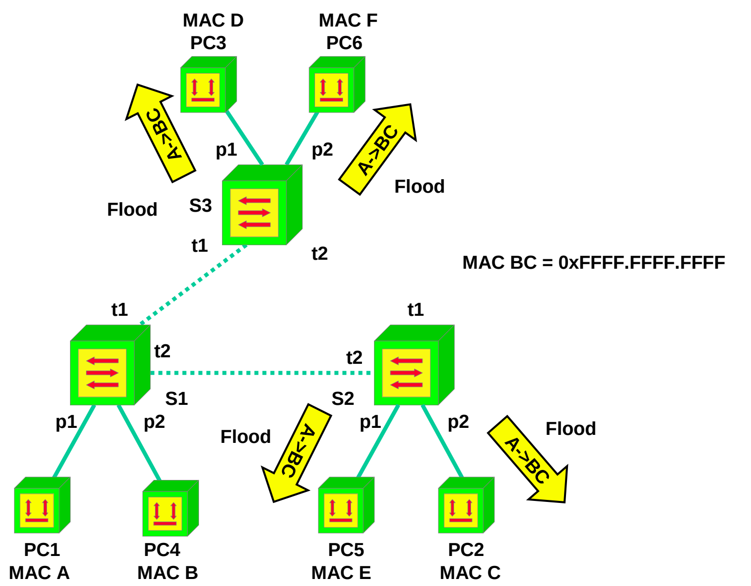
Ethernet Broadcast (BC) 2



© 2011, D.I. Manfred Lindner

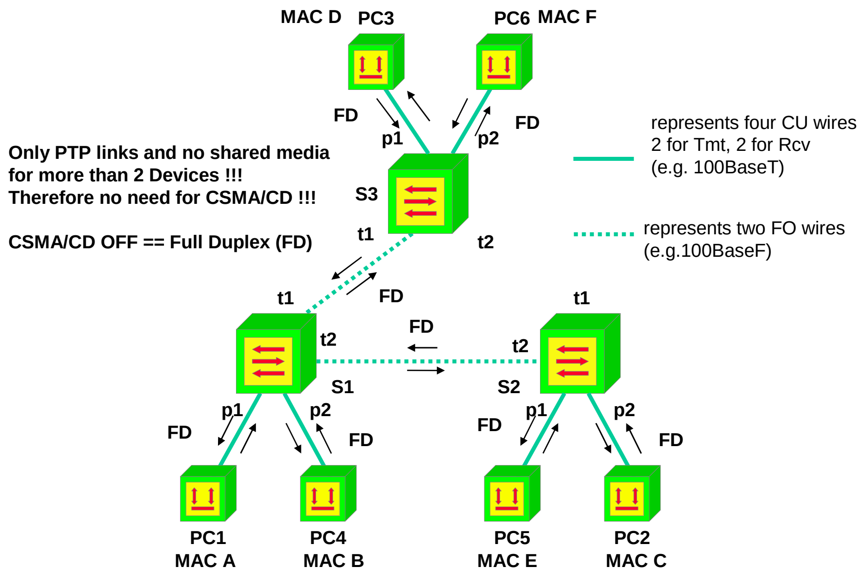
L23 - The Ethernet Evolution

Ethernet Broadcast (BC) 3



© 2011, D.I. Manfred Lindner | Ethernet Evolution, v4.11 | 27

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



© 2011, D.I. Manfred Lindner | Ethernet Evolution, v4.11 | 28

© 2011, D.I. Manfred Lindner

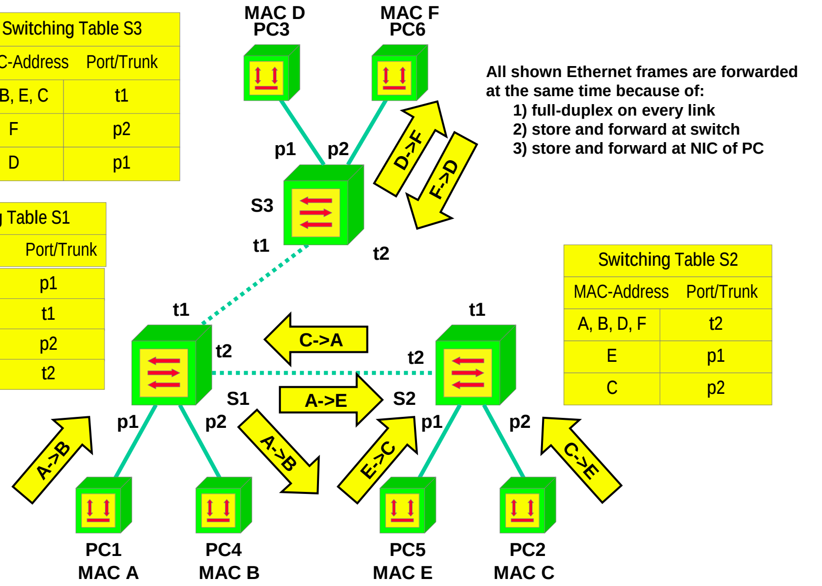
L23 - The Ethernet Evolution

Ethernet Switch Table – Decoupling (Improving Performance <-> Collision Domains)

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

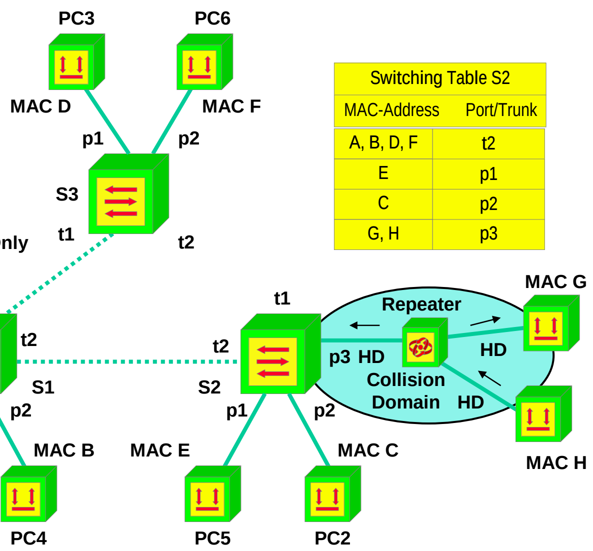


All shown Ethernet frames are forwarded at the same time because of:
1) full-duplex on every link
2) store and forward at switch
3) store and forward at NIC of PC

Ethernet Switching – Repeater (Hub)
(Point-to-Point Links Everywhere but on Shared Media – Half Duplex)

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

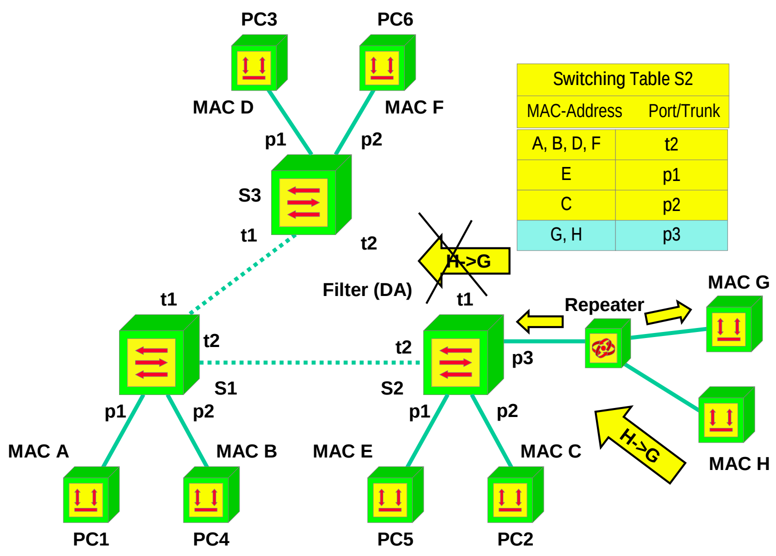
Shared Media == Collision Domain
Collision Domain == CSMA/CD ON
CSMA/CD ON == Half Duplex (HD) Only



© 2011, D.I. Manfred Lindner

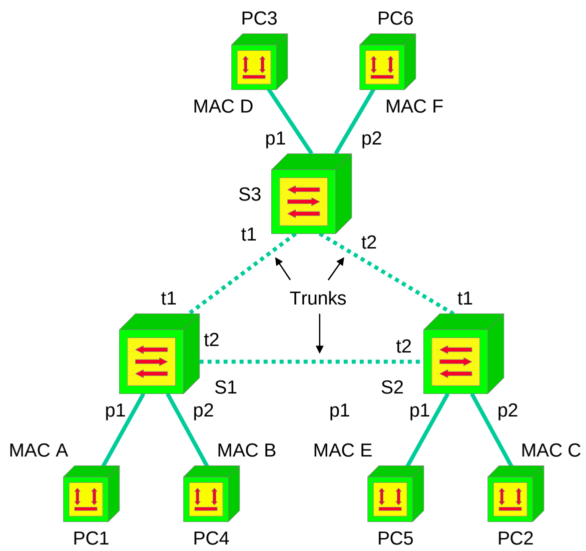
L23 - The Ethernet Evolution

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

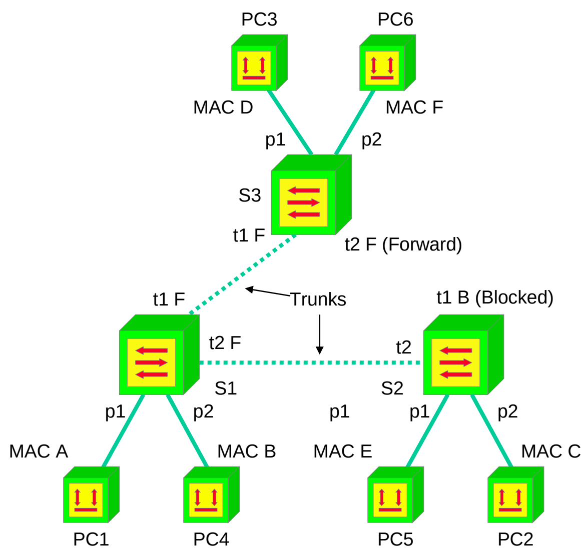
Redundant Topology L2 Switching



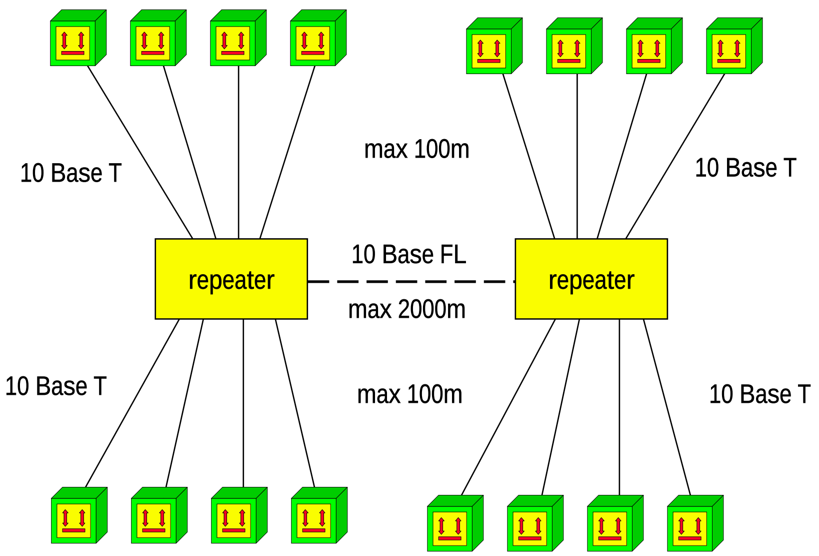
© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

Spanning Tree Applied



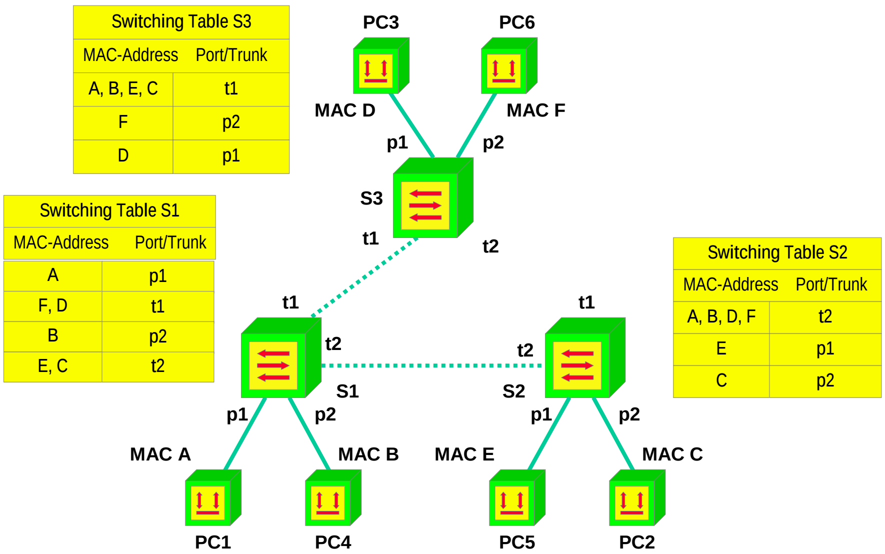
Ethernet with Repeater: Network Sniffing? Yes -> Ethernet Card -> Promiscuous Mode



© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

Ethernet with Switches: Network Sniffing? Not so easy -> Because of Inherent Filtering



Agenda

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

© 2011, D.I. Manfred Lindner

Virtual LANs (1)

- **today's work-groups are expanding over the whole campus in case of local environment**
- **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**
- **the network must be flexible**
  - to adapt continuous location changes of the end-systems/users

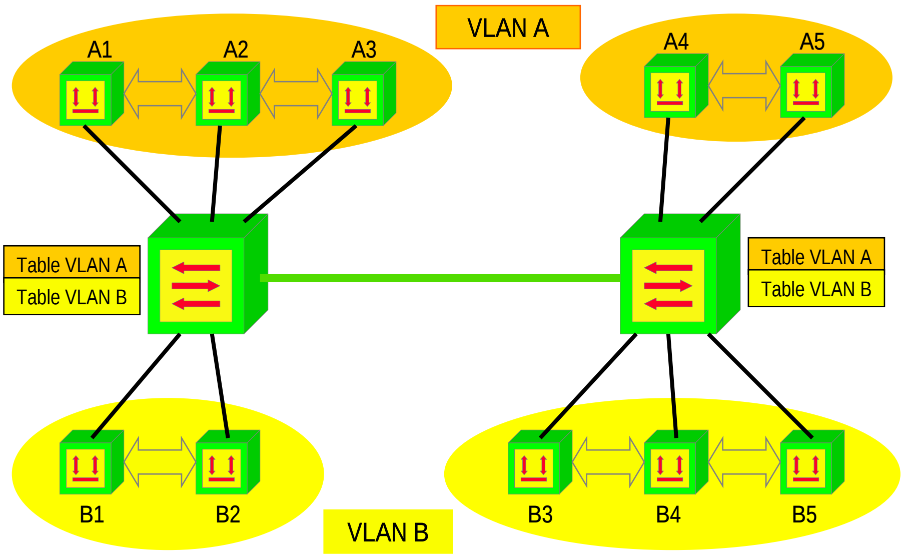
Virtual LANs (2)

- **base idea of VLAN:**
  - multiplexing of several LANs via same infrastructure (switches and connection between switches)
- **today's switches got the ability to combine several network-stations to so-called "Virtual LANs“**
  - separate bridging/switching table maintained for every single VLAN
  - separate broadcast handling for every single VLAN
    - each Virtual LAN is its own broadcast domain
  - separate Spanning Tree for every single VLAN
    - note: IEEE 802.1w specifies a method to share one Rapid Spanning Tree among all VLANs

© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

VLAN Example



VLAN Assignment

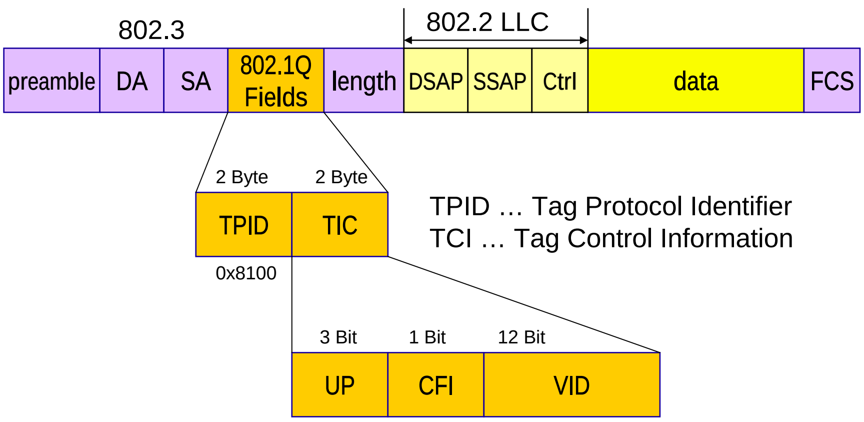
- **a station may be assigned to a VLAN**
  - port-based
    - fixed assignment port 4 -> VLAN x
    - most common approach
    - a station is member of one specific VLAN only
  - MAC-based
    - MAC A -> VLAN x
    - allows integration of older shared-media components and automatic location change support
    - a station is member of one specific VLAN only
  - protocol-based
    - IP-traffic, port 1 -> VLAN x
    - NetBEUI-traffic, port 1 -> VLAN y
    - a station could be member of different VLANs

© 2011, D.I. Manfred Lindner

Virtual Trunks - VLAN tagging

- **switches must be connected via VLAN-trunks on which each particular VLAN-frame is "tagged" (marked) with an identifier**
  - examples for tagging standards:
    - IEEE 802.10 (pre 802.1Q temporary solution)
    - ISL (Cisco)
    - IEEE 802.1Q
- **so switches can distinguish between several VLANs and manage their respective traffic**

802.1Q VLAN Tagging 1



TPID … Tag Protocol Identifier
TCI … Tag Control Information

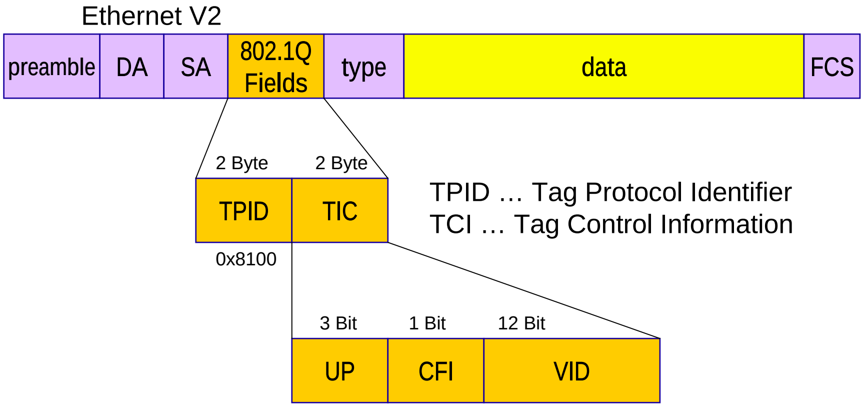
UP … User Priority
CFI … Canonical Format Identifier
VID … VLAN Identifier

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

© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

802.1Q VLAN Tagging 2

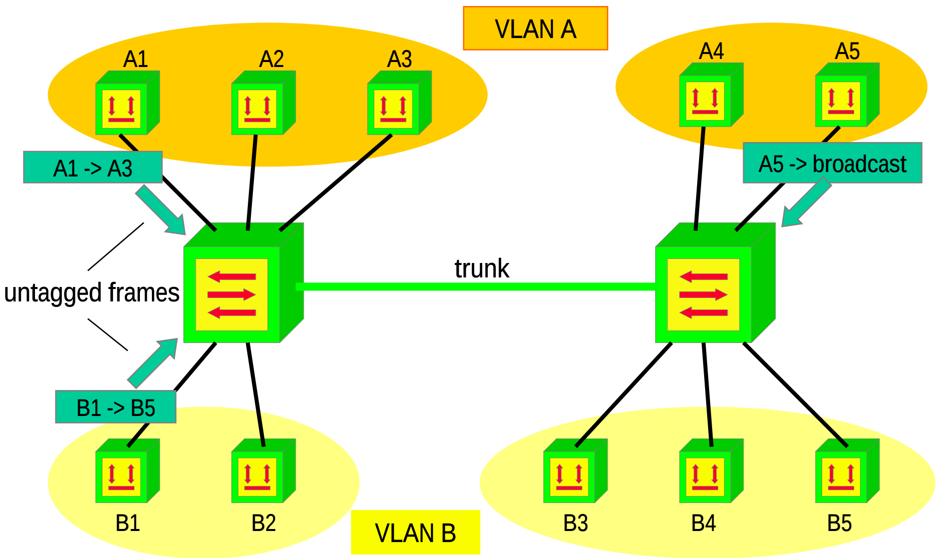


TPID … Tag Protocol Identifier
TCI … Tag Control Information

UP … User Priority
CFI … Canonical Format Identifier
VID … VLAN Identifier

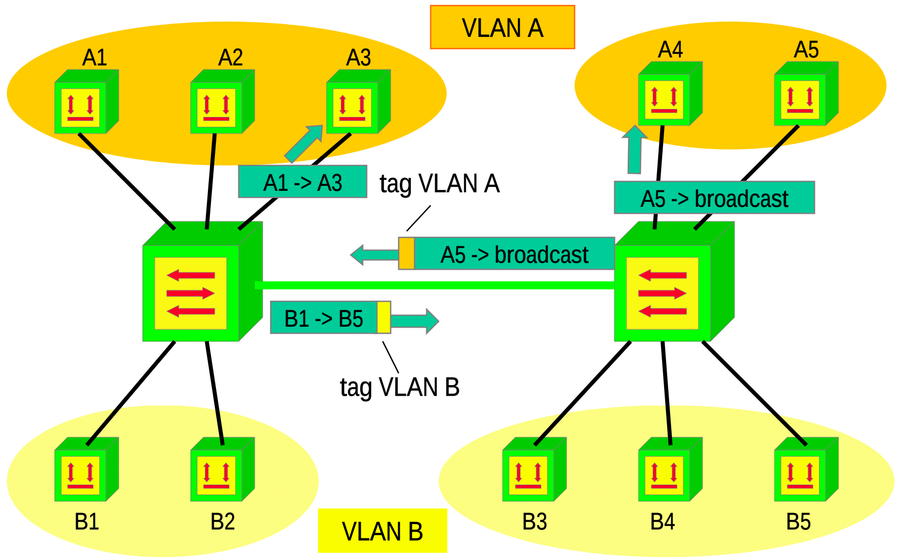
note: With tagging Ethernets maximal frame length = 1522, minimal frame length = 68

VLAN Operation (1)

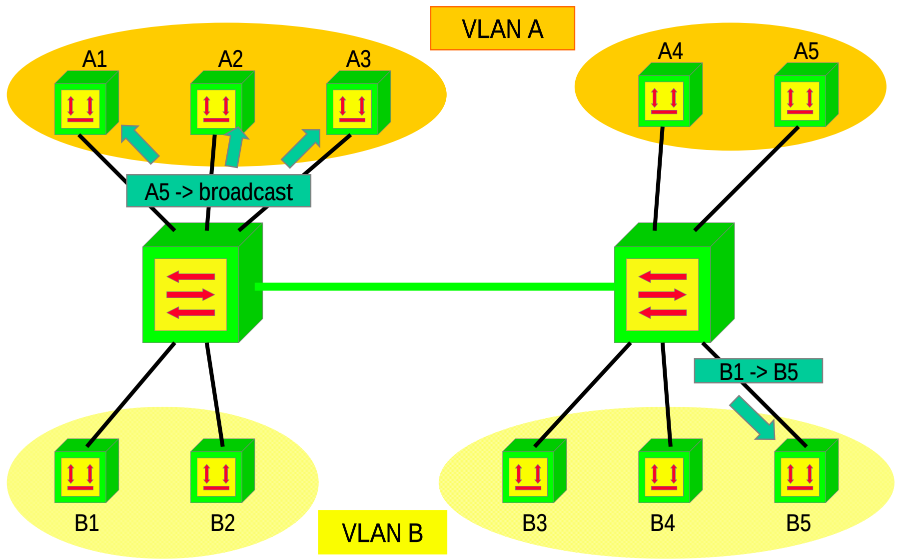


© 2011, D.I. Manfred Lindner

VLAN Operation (2)



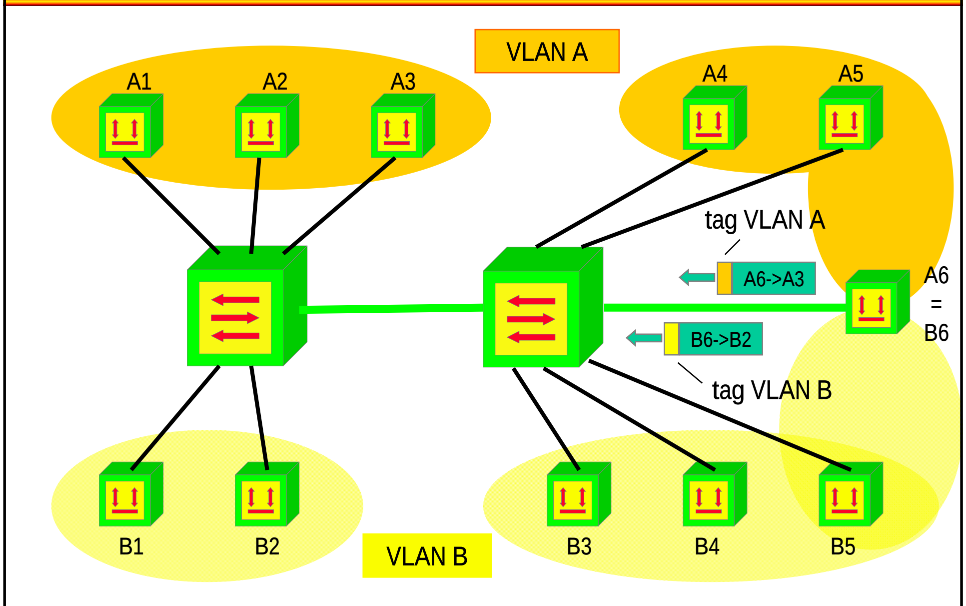
VLAN Operation (3)



© 2011, D.I. Manfred Lindner

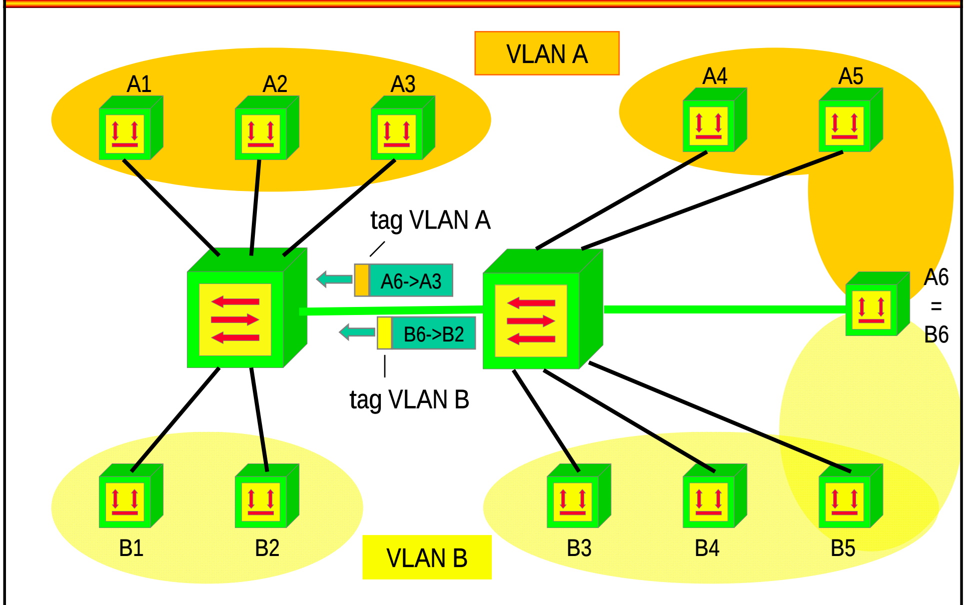
L23 - The Ethernet Evolution

Multihomed VLAN 1



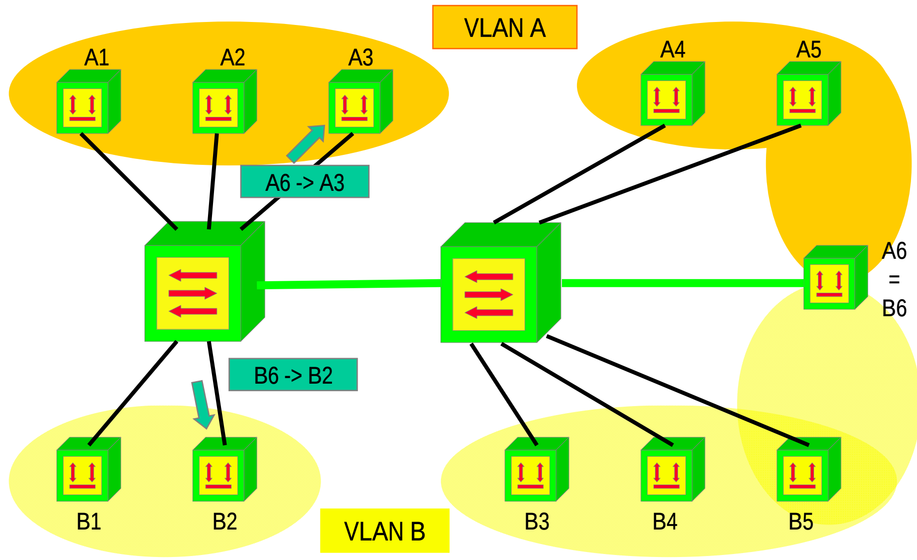
© 2011, D.I. Manfred Lindner | Ethernet Evolution, v4.11 | 47

Multihomed VLAN 2



© 2011, D.I. Manfred Lindner | Ethernet Evolution, v4.11 | 48

Multihomed VLAN 3



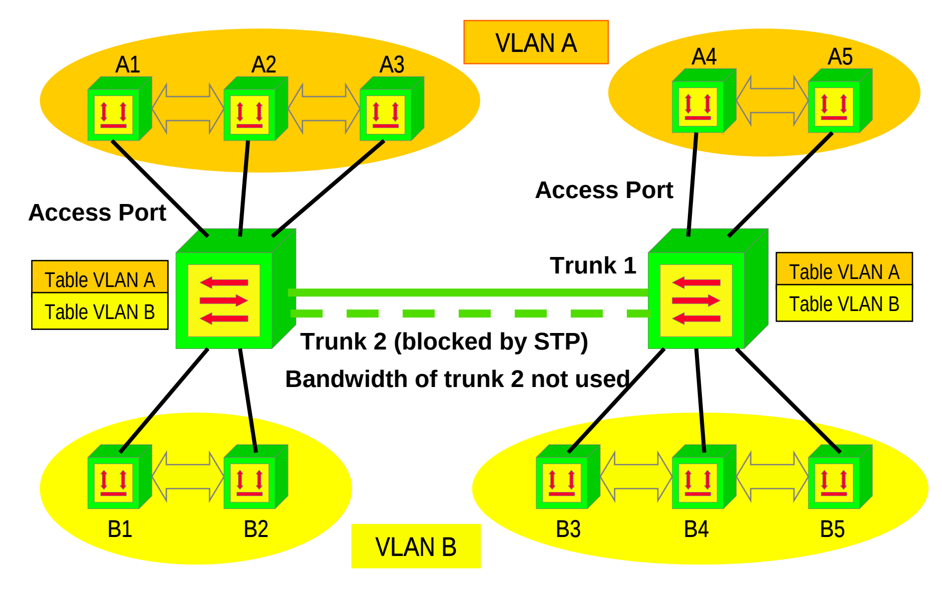
Trunking between L2 Switches

- **on trunks between multiport switches full duplex operation is possible**
  - hence "200 Mbit/s" with Fast Ethernet
  - hence "2 Gbit/s" with Gigabit Ethernet
- **on trunks <u>bundling</u> (aggregation) <u>of physical links</u> to <u>one logical link</u> is possible**
  - Fast Ethernet Channeling (Cisco)
    - 400 / 800 Mbit/s
  - Gigabit Ethernet Channeling (Cisco)
    - 4 / 8 Gbit/s
  - IEEE 802.3 (2002) LACP (Link Aggregation Control Protocol

© 2011, D.I. Manfred Lindner

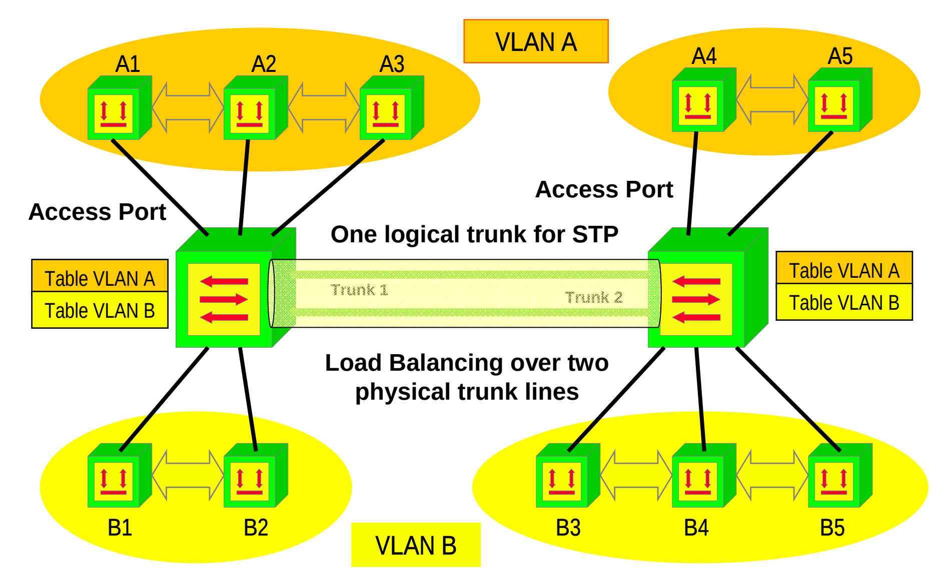
L23 - The Ethernet Evolution

Trunking without LACP / FEC / GEC



© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 51

Trunking with LACP / FEC / GEC



© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 52

© 2011, D.I. Manfred Lindner

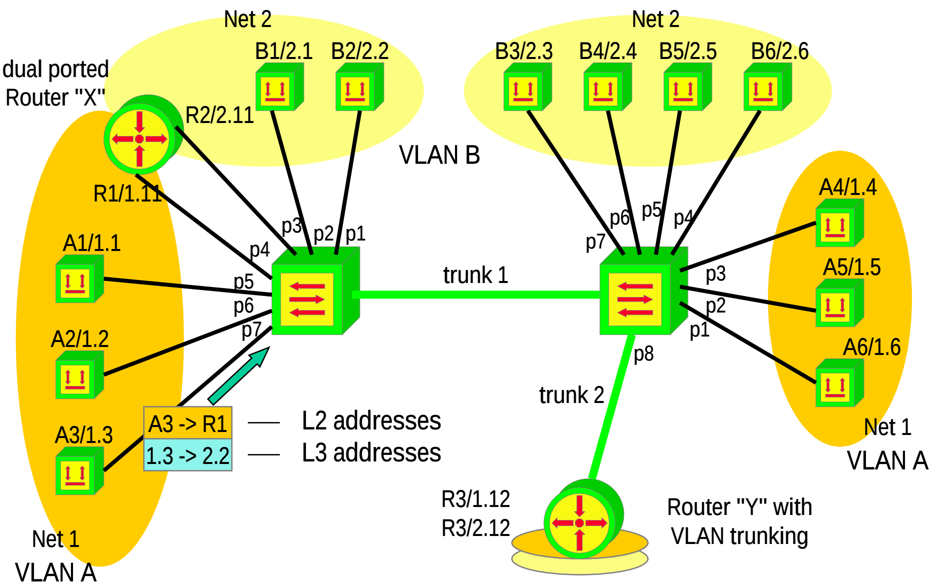
L23 - The Ethernet Evolution

Communication between VLANs

- switches do not allow traffic between (different) VLANs
- end-systems have to make use of routers
- routers can be either part of several VLANs (via multiple physical ports), or
- routers provide VLAN-trunk capabilities -> router must be able to recognize and change tags

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 53

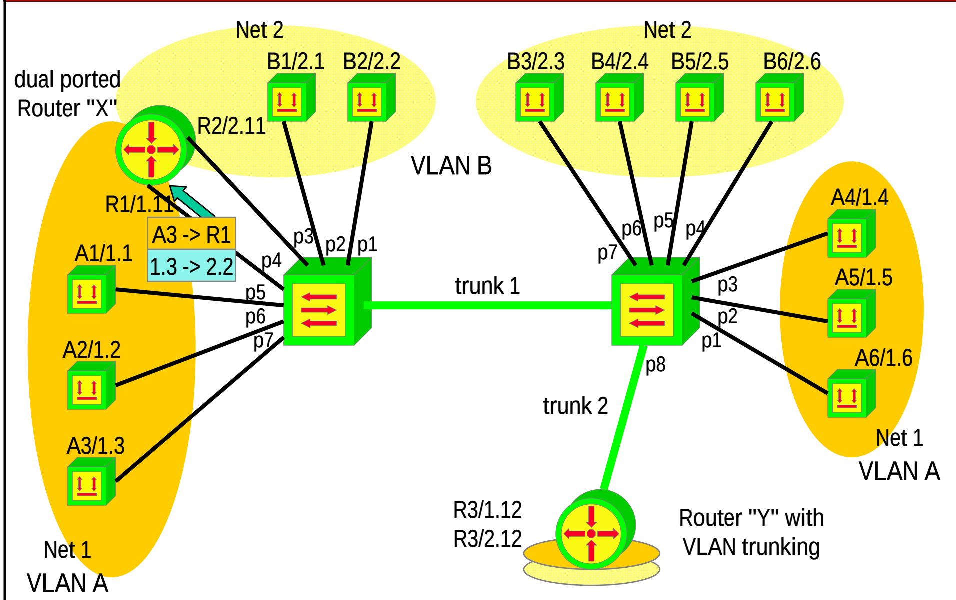
Inter-VLAN-Traffic (1)



© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 54

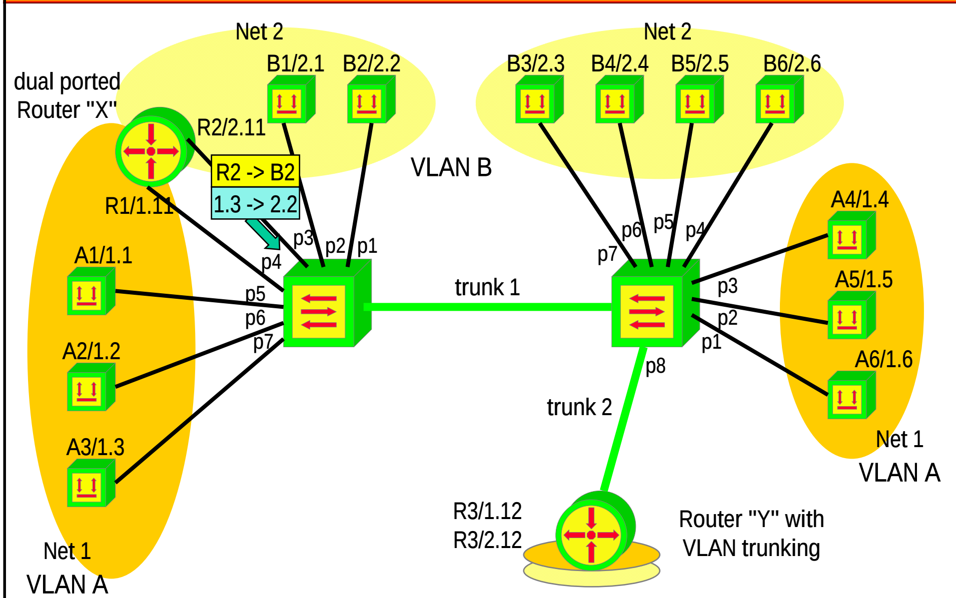
© 2011, D.I. Manfred Lindner

Inter-VLAN-Traffic (2)



© 2011, D.I. Manfred Lindner — Ethernet Evolution, v4.11 — 55

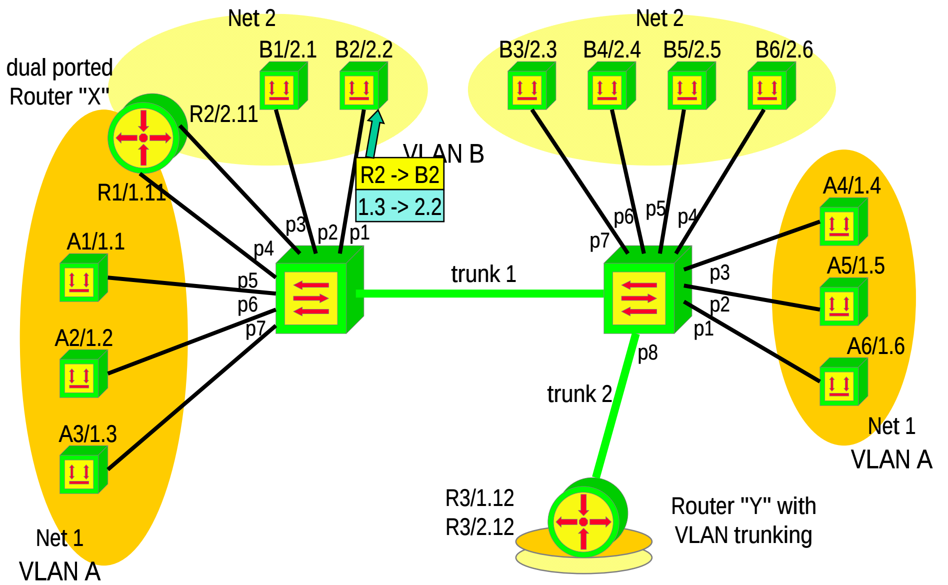
Inter-VLAN-Traffic (3)



© 2011, D.I. Manfred Lindner — Ethernet Evolution, v4.11 — 56

© 2011, D.I. Manfred Lindner

Inter-VLAN-Traffic (4)



Switching and Routing Tables

Switching Table 1

VLAN A	
A1	p5
A2	p6
A3	p7
A4	trunk 1
A5	trunk 1
A6	trunk 1
R1	p4

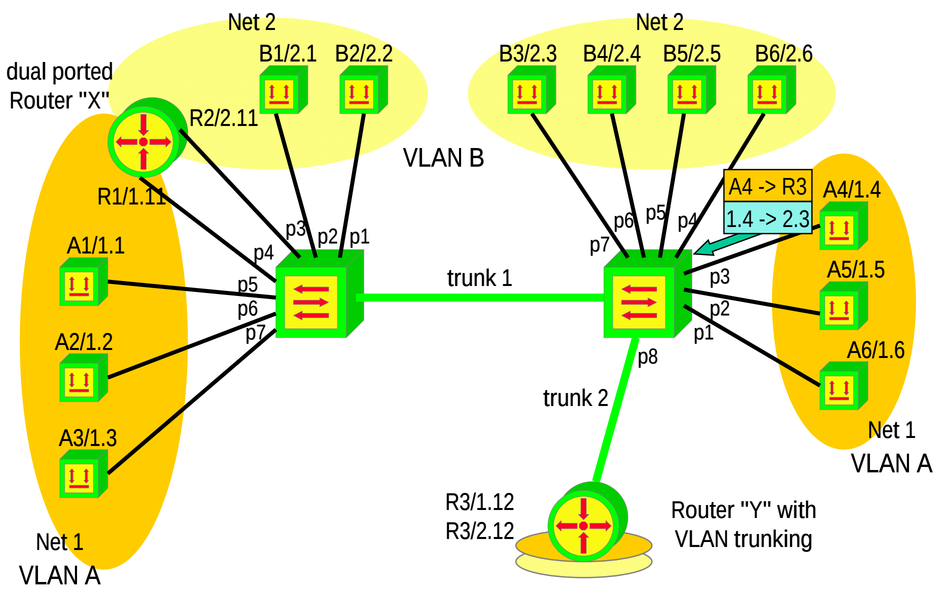
VLAN B	
B1	p2
B2	p1
B3	trunk 1
B4	trunk 1
B5	trunk 1
B6	trunk 1
R2	p3

Routing Table X

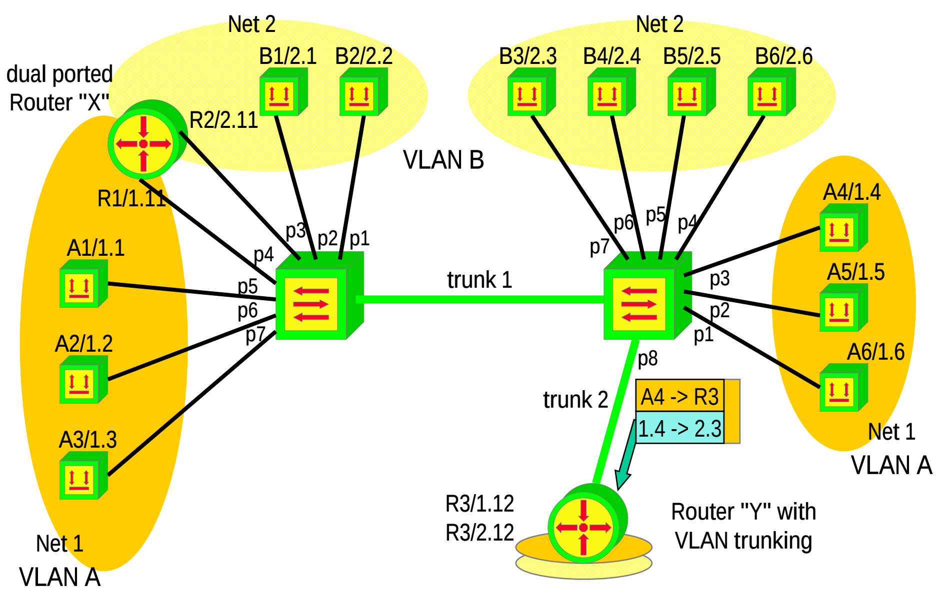
net	next hop	port (MAC)
1	local	R1
2	local	R2

© 2011, D.I. Manfred Lindner

Inter-VLAN-Traffic (5)



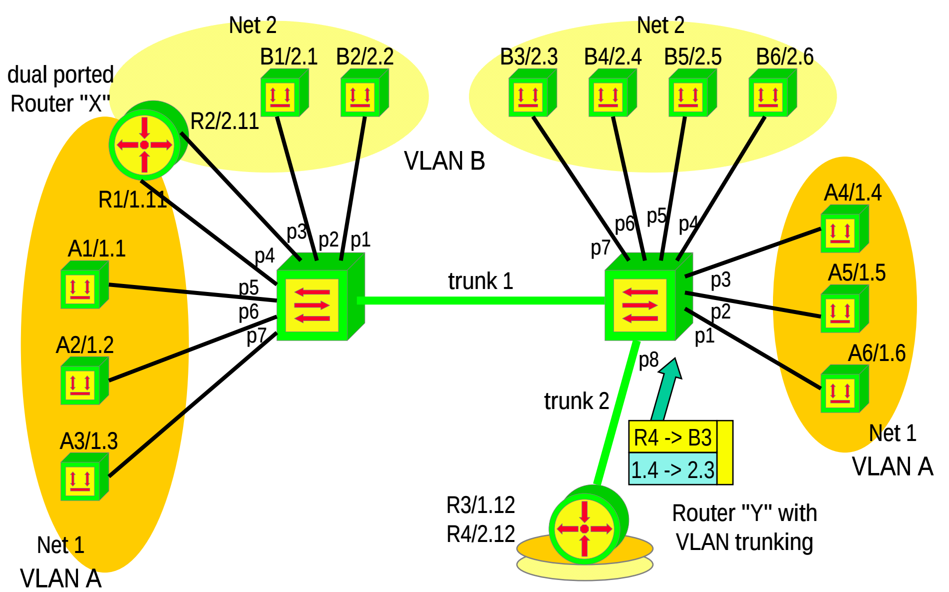
Inter-VLAN-Traffic (6)



© 2011, D.I. Manfred Lindner

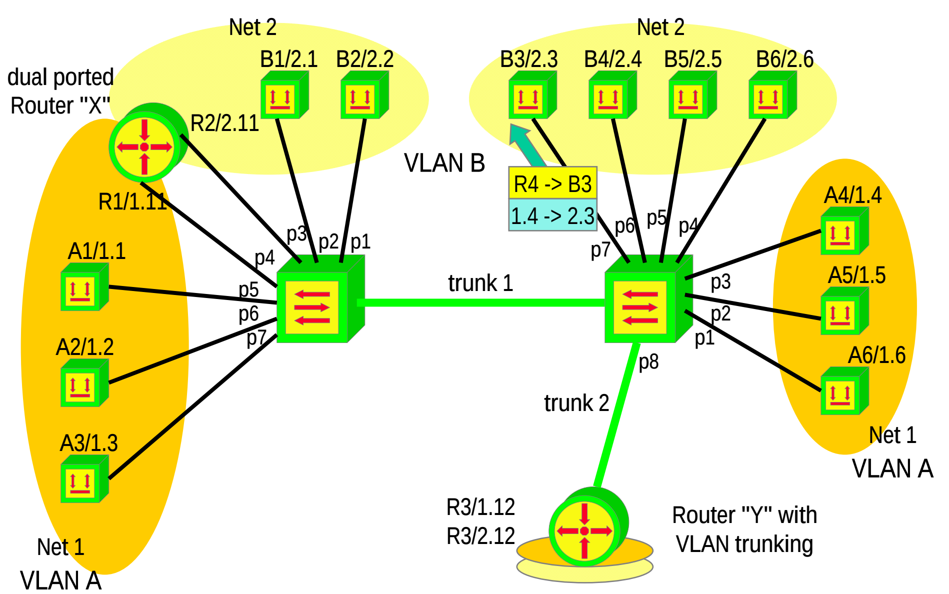
L23 - The Ethernet Evolution

Inter-VLAN-Traffic (7)



© 2011, D.I. Manfred Lindner   Ethernet Evolution, v4.11   61

Inter-VLAN-Traffic (8)



© 2011, D.I. Manfred Lindner   Ethernet Evolution, v4.11   62

Switching and Routing Tables

Switching Table 2

VLAN A	
A4	p3
A5	p2
A6	p1
A1	trunk 1
A2	trunk 1
A3	trunk 1
R3	trunk 2

VLAN B	
B3	p7
B4	p6
B5	p5
B6	p4
B1	trunk 1
B2	trunk 1
R4	trunk 2

Routing Table Y

net	next hop	port (MAC), tagging
1	local	trunk 2, R3, tag=red
2	local	trunk 2, , R4, tag=yellow

Agenda

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

© 2011, D.I. Manfred Lindner

IEEE 802.3 (2002)

- **the latest version of IEEE 802.3 (2005) specifies**
  - operation for 10 Mbit/s, 100 Mbit/s, Gigabit/s and 10Gigabit/sEthernet
  - 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
- **IEEE 802.3ak (2006)**
  - operation for 10 Gigabit/s Ethernet over copper

Full-Duplex Mode

- **full-duplex mode is possible on point-to-point links**
  - except 100BaseT4 (Cat 3 cable), 100BaseVG which can work in half duplex mode only
  - note: 10Base2 and 10Base5 are shared links and by default half duplex medias
- **if a network station is connected to an Ethernet switch via point-to-point link**
  - CSMA/CD is not in necessary and can be switched off
- **now a network station can**
  - <u>send</u> frames immediately (without CS) using the transmission-line of the cable <u>and simultaneously receive</u> data on the other line

© 2011, D.I. Manfred Lindner

Flow Control

- **speed-requirements for switches are very high**
  - especially in full duplex operation
  - also powerful switches can't 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-protocol and 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-ctrl 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)**

© 2011, D.I. Manfred Lindner

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)

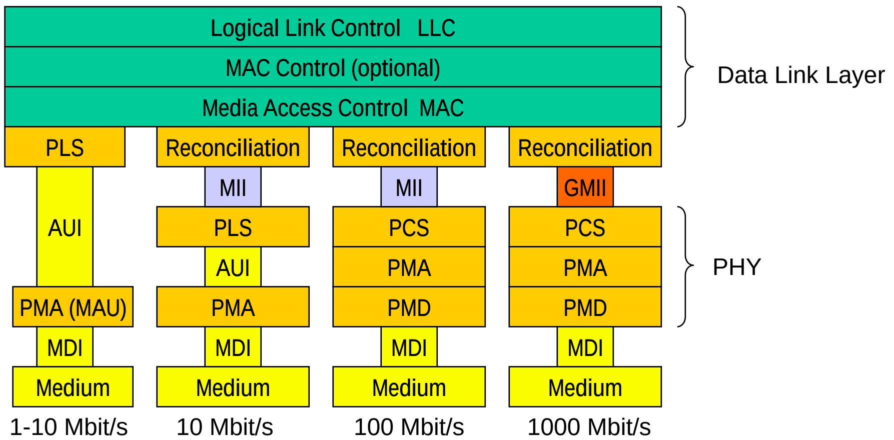
The Pause Command 2

- **destination address is either**
  - address of destination station or
  - broadcast address or
  - special multicast address 01-80-C2-00-00-01
- **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**

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

© 2011, D.I. Manfred Lindner

PHY Sublayers

- **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

- **PLS has been replaced with the Reconciliation sublayer**
  - Reconciliation layer transforms old MAC PLS-primitives into MII control signals
- **MII 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)
  - 4 bit (GMII: 8 bit) parallel transmission channels to the physical layer

© 2011, D.I. Manfred Lindner

PHY Sublayers

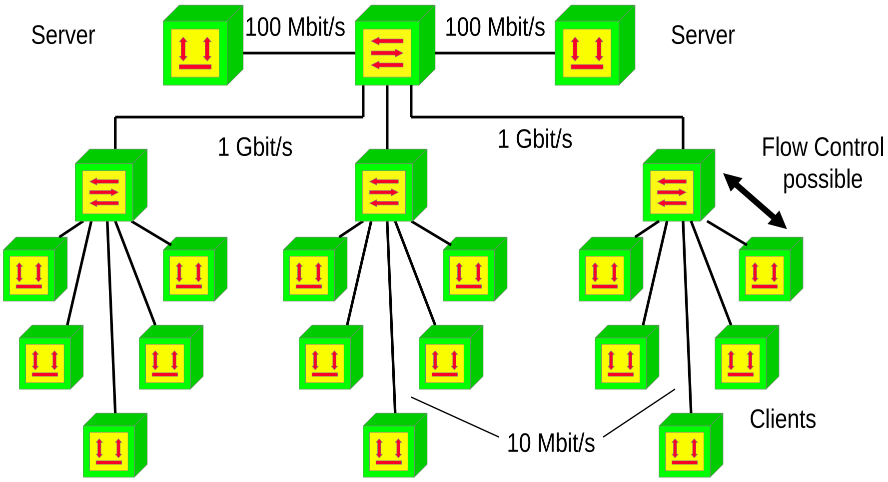
- **Physical Coding Sublayer (PCS)**
  - encapsulates MAC-frame between special PCS delimiters
  - 4B/5B or 8B/10B encoding respectively
  - appends idle symbols
- **Physical Medium Attachment (PMA)**
  - interface between PCS and PMD
  - (de) serializes data for PMD (PCS)
- **Physical Medium Dependent (PMD)**
  - serial transmission of the code groups
  - specification of the various connectors (MDI)

Bridging Aspects

- **new PHY-sublayers preserves old Ethernet MAC frame format**
  - 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 -> slower !!
- **therefore Ethernet L2 switches**
  - can connect Ethernets with 10 Mbit/s, 100 Mbit/s or 1000 Mbit/s easily and fast

© 2011, D.I. Manfred Lindner

Different Speeds – No Problem with Ethernet Switching because of Packet Switching Principle



Today: Gigabit Ethernet

- **continues point-to-point and full-duplex idea**
- **also backward compatible with initial 10 Mbit/s shared media idea -> CSMA/CD capable**
- **but nobody uses it as shared media!**
  - multiport repeater with Gigabit Ethernet seems absurd because of small network diameter (20m)
    - 200m with carrier extension and burst mode
  - bandwidth sharing decreases performance; every collision domain produces an additional delay for a crossing packet
  - full duplex means exclusive, unshared, high performance point-to-point connections between two stations (total 2Gbit/s!)

© 2011, D.I. Manfred Lindner

Gigabit Ethernet becomes WAN

- **point-to-point full-duplex connections do not limit the maximal network diameter as CSMA/CD does**
  - Gigabit over fiber optic cables reach 70 km length (and even more)
- **trend moves towards layer 3 switching**
  - high amount of today's traffic goes beyond the border of the LAN
  - routing decisions enable load balancing and decrease network traffic
- **Gigabit Ethernet becomes WAN technology**

Agenda

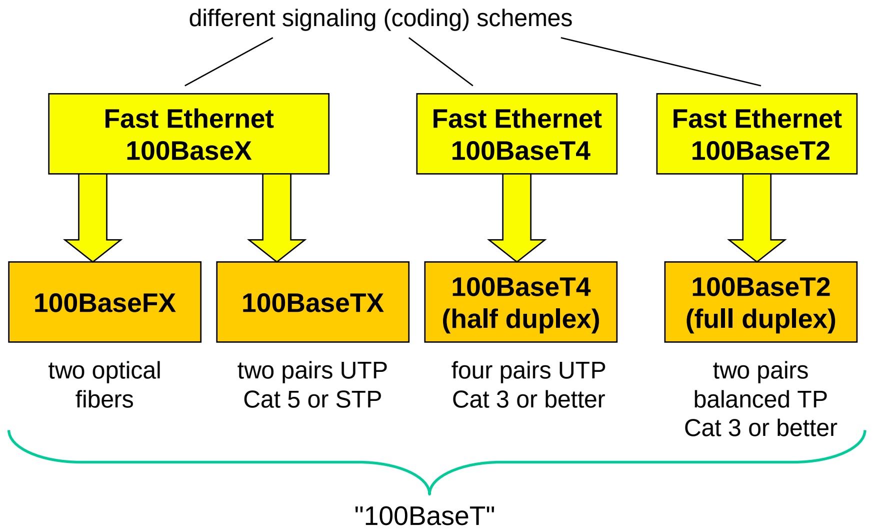
- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

© 2011, D.I. Manfred Lindner

100 Mbit/s Ethernet

- **Access method disagreement split 100 Mbit/s LAN development into two branches:**
  - Fast Ethernet - IEEE-802.3u (today 802.3-2002)
  - 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

Implementation Overview 1



Implementation Overview 2

preserves classical Ethernet

100BaseT

Access method: CSMA/CD

IEEE 802.3

HP and AT&T own specification for time sensitive applications

100VG-AnyLAN

Access method: demand priority

IEEE 802.12

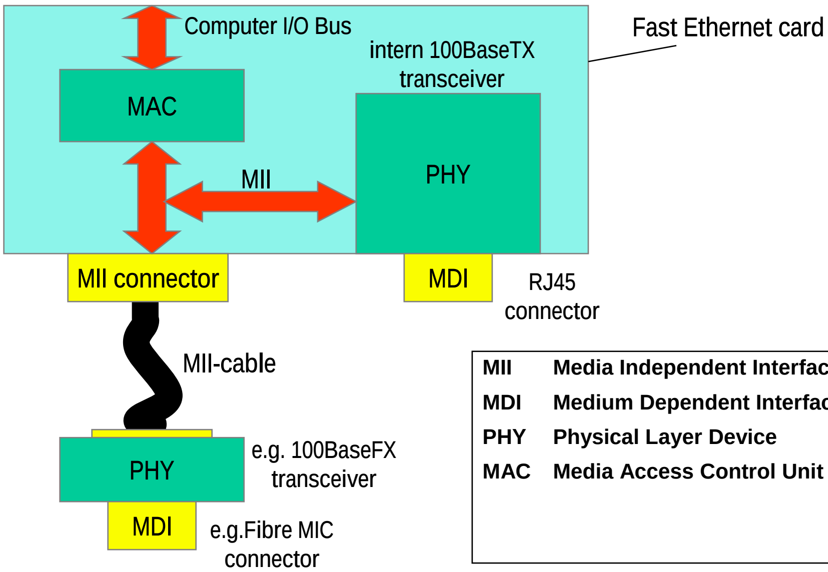
Fast Ethernet

- **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

© 2011, D.I. Manfred Lindner

L23 - The Ethernet Evolution

Typical Fashion



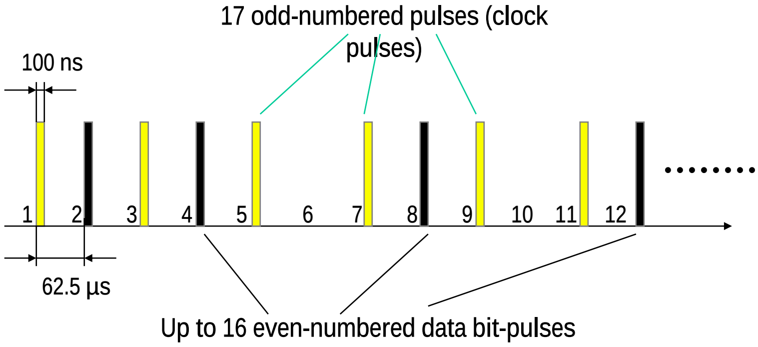
MII	**Media Independent Interface**
MDI	**Medium Dependent Interface**
PHY	**Physical Layer Device**
MAC	**Media Access Control Unit**

Autonegotiation

- **Autonegotiation support enables two 100BaseT devices (copper only) to exchange information about their capabilities**
  – signal rate, CSMA/CD or full-duplex
- **Achieved by Link-Integrity-Test-Pulse-Sequence**
  – Normal-Link-Pulse (NLP) technique is already available in 10BaseT to check the link state
  – 10 Mbit/s LAN devices send every 16 ms a 100ns lasting NLP -> no signal on the wire means disconnected
- **100BaseTX uses bursts of Fast-Link-Pulses (FLP) consisting of 17-33 NLPs**
  – Each representing a 16 bit word

© 2011, D.I. Manfred Lindner

FLP Burst Coding



= 1 1 0 1 0 1 ....

Autonegotiation

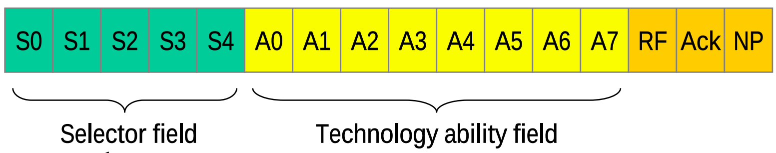
- **To avoid increase of traffic FLP-bursts are only sent on connection-establishments**
- **100BaseT stations recognizes 10 Mbit/s stations by receiving a single NLP only**
- **Two 100BaseT stations analyze their FLP-bursts and investigate their largest common set of features**
- **Last frames are sent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

© 2011, D.I. Manfred Lindner

FLP-Session

- **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



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

© 2011, D.I. Manfred Lindner

Base Page

- **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

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 91

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)

© 2011, D.I. Manfred Lindner Ethernet Evolution, v4.11 92

© 2011, D.I. Manfred Lindner

Next-Pages

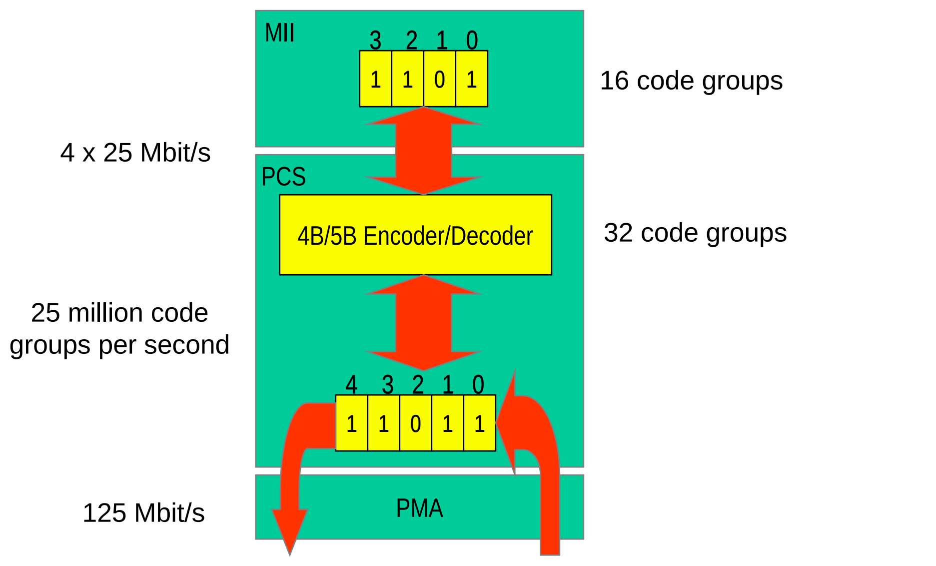
- **Acknowledge 2 (Ack2)**
  - Ack2 is set to 1 if station can perform the declared capabilities
- **Message Page (MP)**
  - Differentiates between message-pages (MP=1) and
  - Unformatted-pages (MP=0)
- **Toggle (T)**
  - Provides synchronization during exchange of next-pages information
  - T-bit is always set to the inverted value of the 11th bit of the last received link-codeword

Coding

- **4B/5B block encoding: each 4-bit group 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 %)

© 2011, D.I. Manfred Lindner

4B/5B Coding



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)

© 2011, D.I. Manfred Lindner

Signaling Types

- **Three signaling types :**
  - 100BaseX:
    - refers to either the 100BaseTX or 100BaseFX specification
  - 100BaseT4
  - 100BaseT2
- **100BaseX**
  - combines the CSMA/CD MAC with the FDDI Physical Medium Dependent layer (PMD)
  - allows full duplex operation on link

Signaling Types

- **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

© 2011, D.I. Manfred Lindner

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 and 100BaseT2

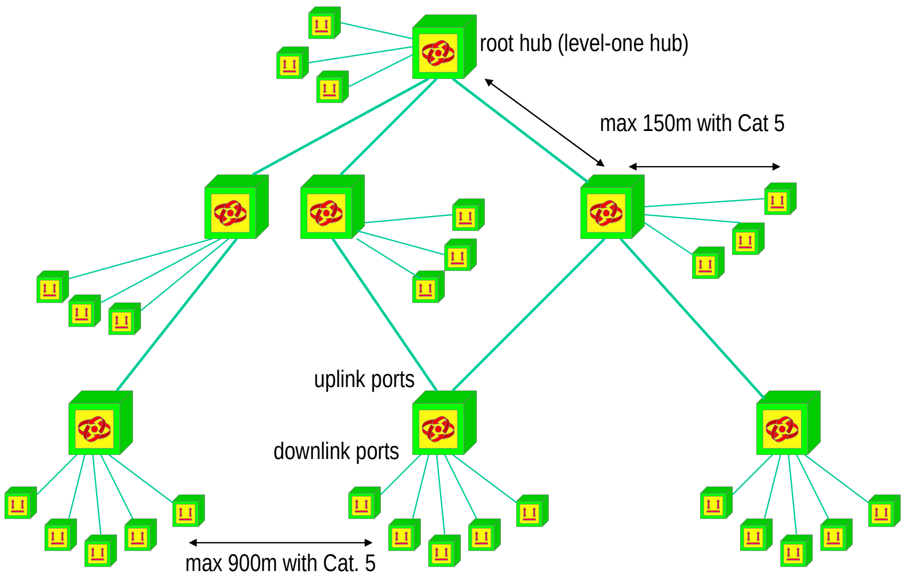
- **100BaseT4:**
  - 25 MBaud, half duplex, ternary encoding
  - Cat3 or better, needs all 4 pairs installed
  - 200 m maximal network diameter
  - maximal 2 hubs
- **100BaseT2:**
  - 25 MBaud, full duplex, quinary encoding
  - 2 pairs Cat3 or better

© 2011, D.I. Manfred Lindner

100VG-AnyLAN

- specified by HP and AT&T, standardized by IEEE 802.12
- uses 802.2 LLC but incompatible with 802.3 MAC
- designed for existing "Voice Grade" cabling (point to point only, unidirectional) in a tree structured net; hubs are arranged hierarchically
- demand priority access method which is more deterministic than CSMA/CD; eliminates collisions and can be more heavily loaded than 100BaseT

100VG-AnyLAN Cabling Structure Example



© 2011, D.I. Manfred Lindner

100VG-AnyLAN Operation 1

- station is in a sending or receiving mode (never both)
- each hub has at least one uplink port and several downlink ports
- hubs can be cascaded 3 levels deep; level 1 hub controls the priority domain and polls its connected hubs
- station signals send-request to the hub
- if network is idle, station gets sending permission immediately; station sends packet to the hub

© 2011, D.I. Manfred Lindner

100VG-AnyLAN Operation 2

- on receiving more than one request: hub schedules sending permissions using a round robin method which can be controlled by priority tags (packet switching task)
- to ensure fairness, a hub does not grant priority access to a port more than twice in a row
- 5B/6B block code
- various cabling types
  - 4 wire pairs of Cat 3 UTP (100m)
  - 2 wire pairs Cat 4 or Cat 5 UTP (150m)
  - STP cable
  - Fibre Optic

© 2011, D.I. Manfred Lindner

© 2011, D.I. Manfred Lindner

Agenda

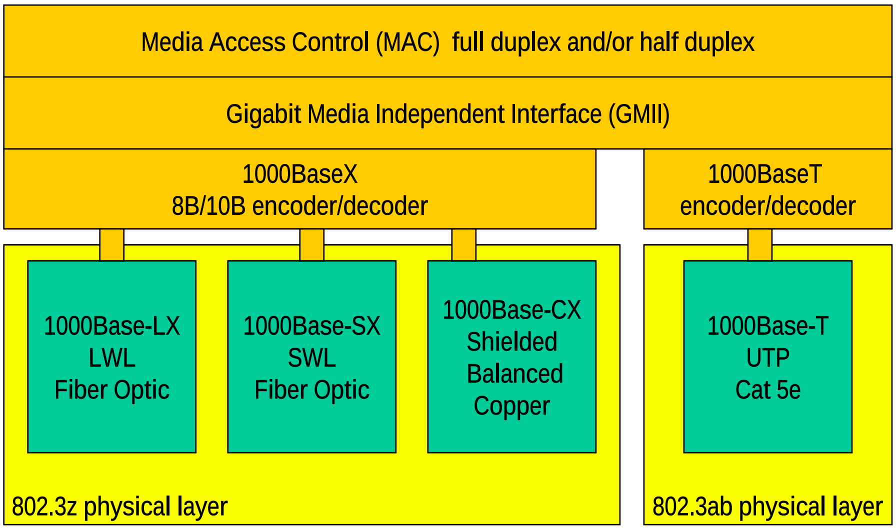
- **Ethernet Evolution**
- **VLAN**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

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

- **Easy integration in existing 802.3 LAN configurations because backwards compatible**
  - Through integration of 3 different transceivers for 10, 100 and 1000 Mbit/s
  - No need to change existing equipment
  - Supports also 10 Mbit/s and 100 Mbit/s (not with fibre)
  - Access methods: CSMA/CD or full duplex
- **Backbone technology; has also WAN capabilities**
  - Reaches 70 km length using fibre optics
  - 1 Gbit/s data rate in both directions (full duplex mode, no collisions)
  - MAC based congestion avoidance (pause frame)

© 2011, D.I. Manfred Lindner

Implementation Overview

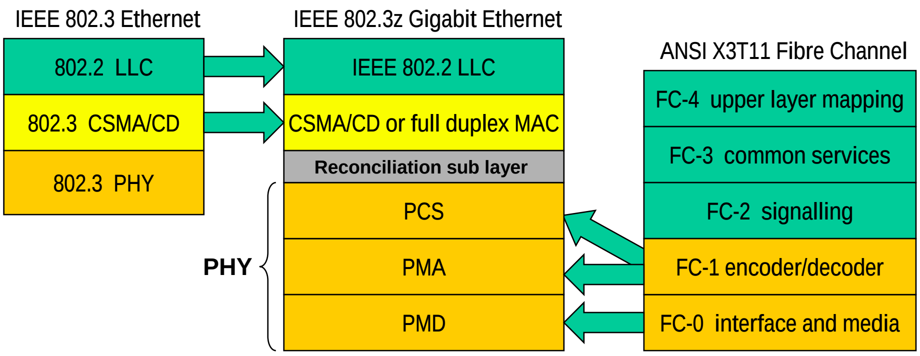


1000BaseX

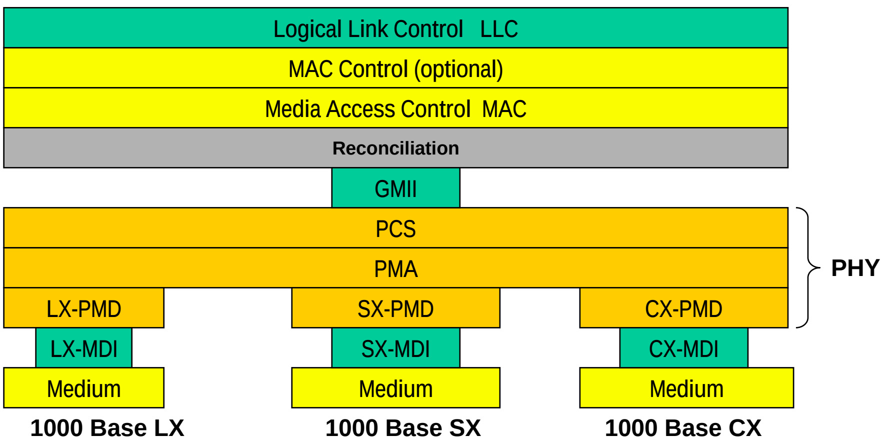
- Looks like 802.3 Ethernet from the data link layer upward
- Physical layer consists of well-tried high-speed components of the Fibre Channel implementation
- Coding is similar to Fibre Channel 8B/10B (FC1 layer) but at higher signal rate of 1.25 Gbaud
- Reconciliation layer translates between the link layer and the physical layer

© 2011, D.I. Manfred Lindner

Protocol Architecture



Physical Sublayers



© 2011, D.I. Manfred Lindner

CSMA/CD Restrictions (Half Duplex Mode)

- **The conventional collision detection mechanism CSMA/CD**
  - Requires that stations have to listen (CS) twice the signal propagation time to detect collisions
  - Collision window of 512 bit times at a rate of 1Gbit/s limits the maximal net expansion to 20m !

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").

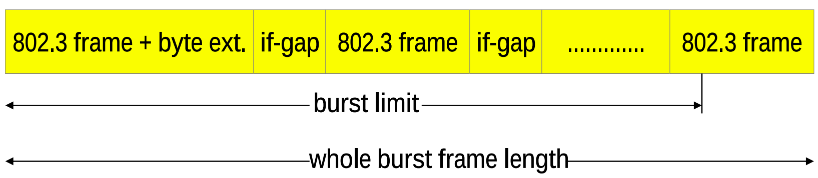
© 2011, D.I. Manfred Lindner

Frame Bursting 1

- **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)**

Frame Bursting 2

- **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

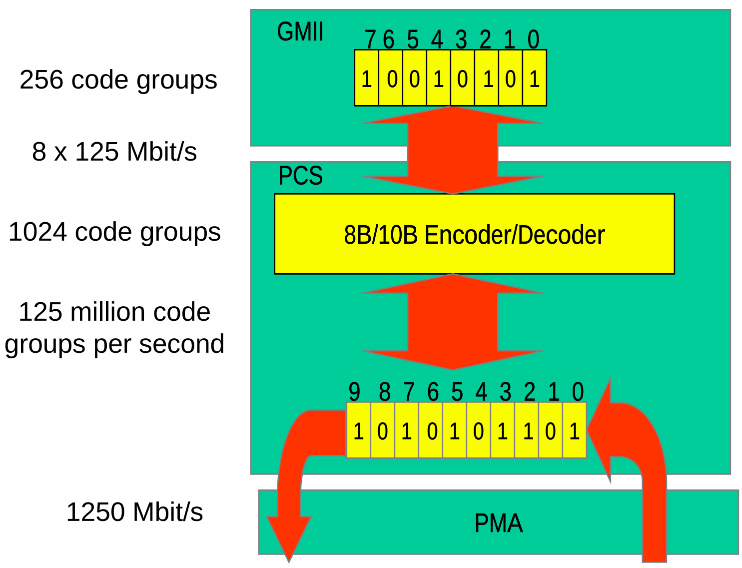


© 2011, D.I. Manfred Lindner

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%

8B/10B Coding



© 2011, D.I. Manfred Lindner

8B/10B Coding

- **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

- **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**

© 2011, D.I. Manfred Lindner

Ordered Sets

- **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

Implementations

- **Actually 2 different wavelengths on fibre media, both full duplex, SC connector**
  - 1000Base-SX: short wave, 850 nm multimode (up to 550 m length)
  - 1000Base-LX: long wave, 1300 nm multimode or monomode (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

© 2011, D.I. Manfred Lindner

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

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

© 2011, D.I. Manfred Lindner

1000BaseX Autonegotiation

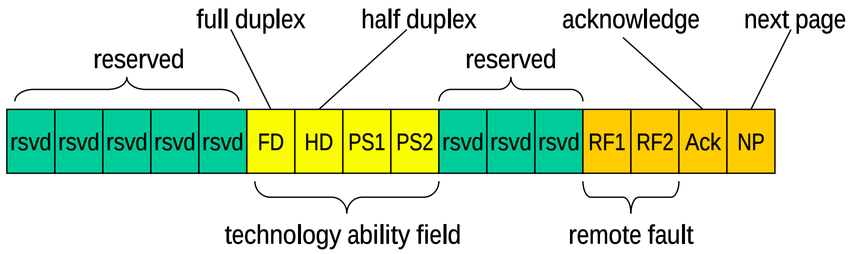
- **1000Base-X autonegotiation uses normal (1000Base-X) signalling !**
  - "Ordered sets" of the 8B/10B code groups
  - No fast link pulses !
    - Autonegotiation had never been specified for traditional fiber-based Ethernet
    - So there is no need for backwards-compatibility
- **1000Base-X does not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

1000BaseX Autonegotiation

- **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**

© 2011, D.I. Manfred Lindner

Base-Page

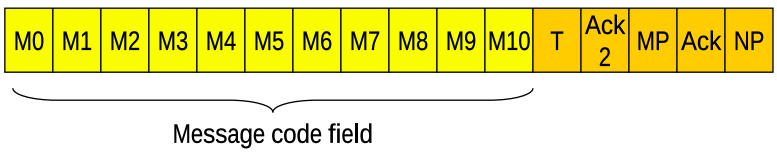


PS1	PS2	description
0	0	no pause
0	1	asymmetrical pause
1	0	symmetrical pause
1	1	symmetrical and asymmetrical pause

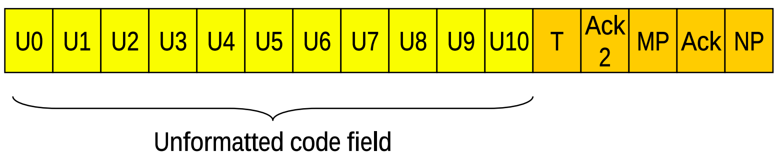
RF1	RF2	description
0	0	no error
0	1	offline
1	0	connection error
1	1	autonegotiation error (no common capabilities)

Next-Pages 1

Normal message-page (predefined codes)



Vendor specific page (non predefined codes)



© 2011, D.I. Manfred Lindner

Next-Pages 2

- **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

© 2011, D.I. Manfred Lindner

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**

© 2011, D.I. Manfred Lindner

© 2011, D.I. Manfred Lindner

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

© 2011, D.I. Manfred Lindner

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

© 2011, D.I. Manfred Lindner

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

© 2011, D.I. Manfred Lindner