Ethernet Primer (v6.1)

Primer Ethernet Technology

From 10Mbit/s to 10Gigabit/s Ethernet Technology
From Bridging to L2 Ethernet Switching and VLANs
From Spanning-Tree to Rapid Spanning-Tree

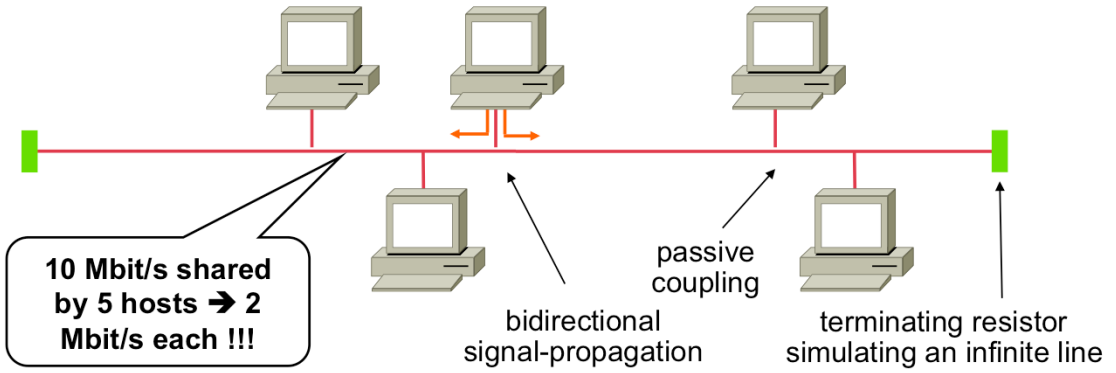
Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed Ethernet

Ethernet Primer (v6.1)

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 Primer, v6.1 3

Basic ideas:

Bus topology based on coax-cables with passive, uninterrupted coupling. Shared media like the „Ether“ of air. Bidirectional signal-propagation -> termination resistors avoid signal reflections. Only half duplex transmission is possible. Control who is allowed to send is performed by media access control implementation. Given transmitting power of a network station limits cable length and number of receiver-stations. 10 Mbit/s baseband transmission with Manchester encoding. To differentiate between network stations we need addresses -> source and destination MAC addresses.

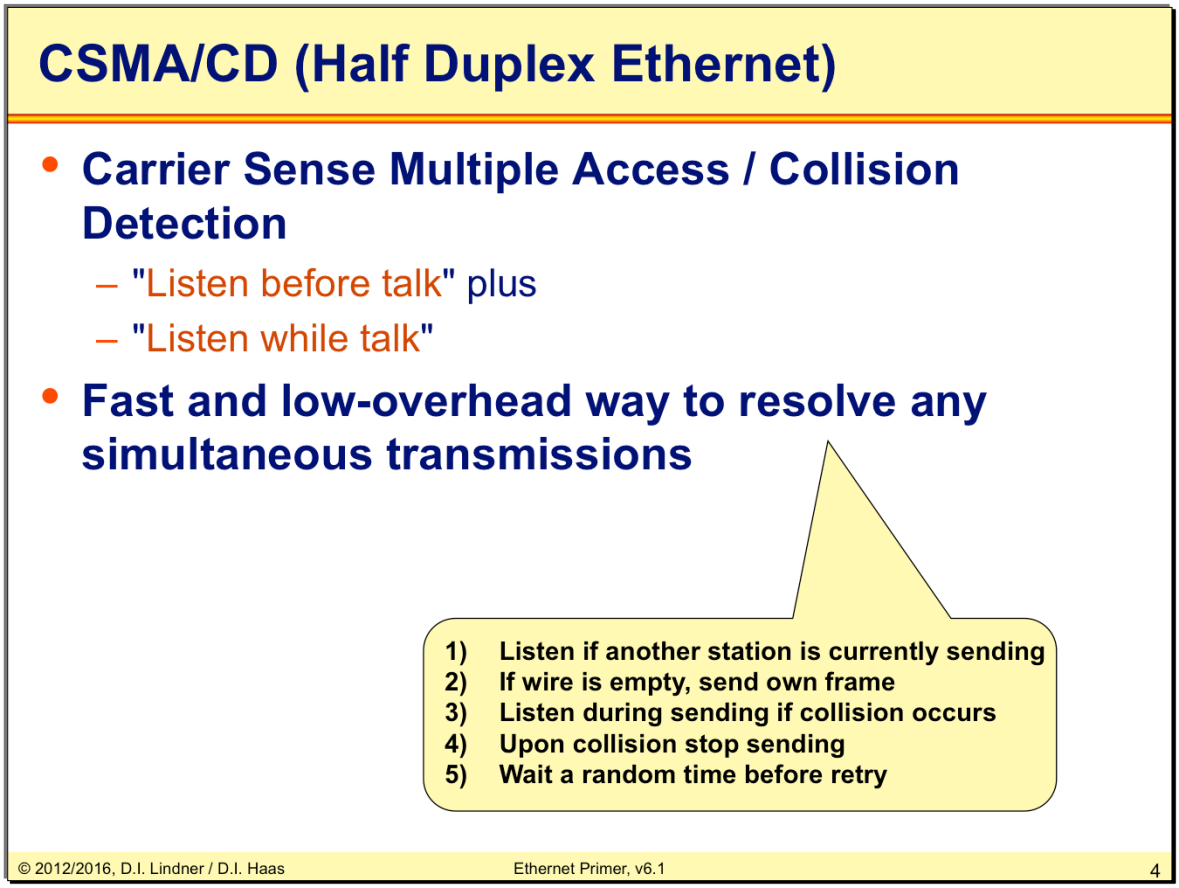
The initial idea of Ethernet was completely different than what is used today under the term "Ethernet". The original new concept of Ethernet was the use of a shared media and an Aloha based access algorithm, called Carrier Sense Multiple Access with Collision Detection (CSMA/CD). Coaxial cables were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

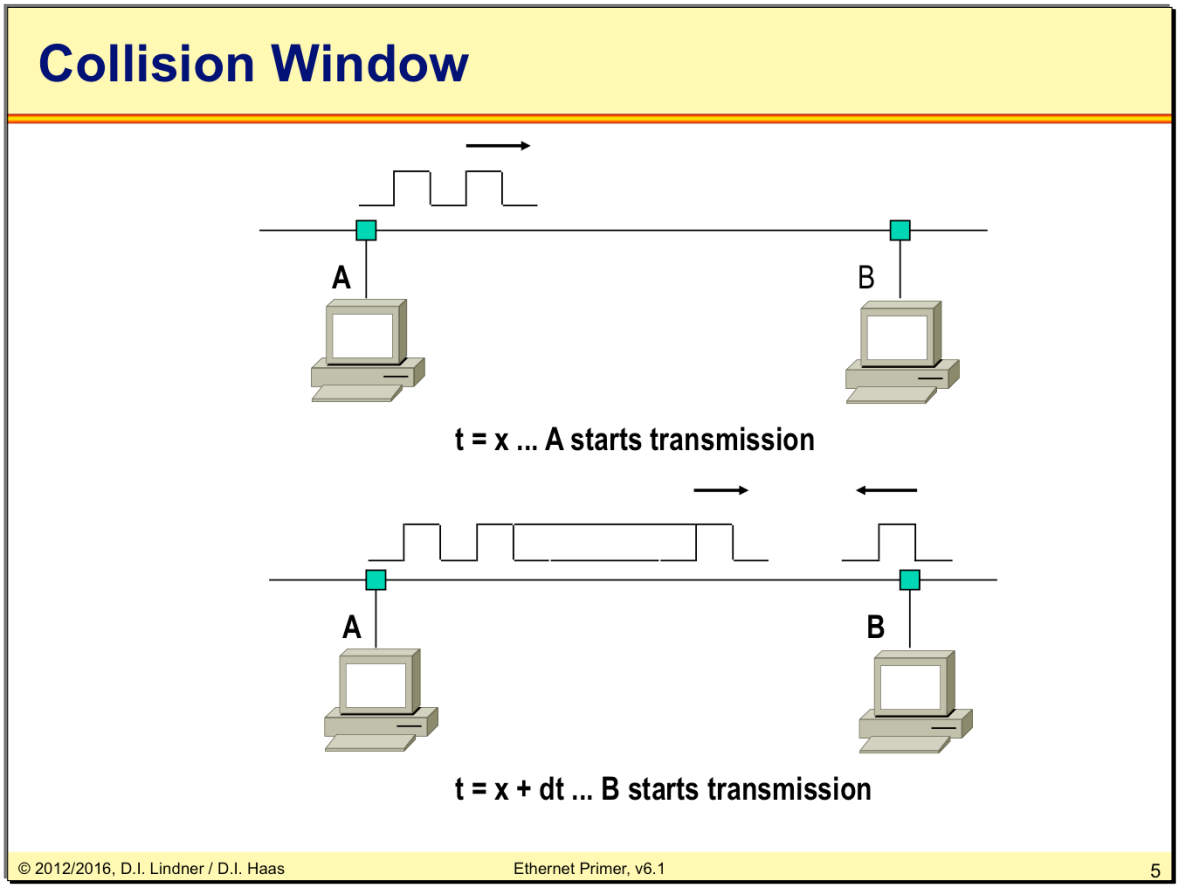
Ethernet technologies at that time (1975-80s): 10Base2 and 10Base5



Ethernet is a shared media technology, so a procedure had to be found to control the access onto the physical media. This procedure was called the Carrier Sense Multiple Access Collision Detection (CSMA/CD) circuit.

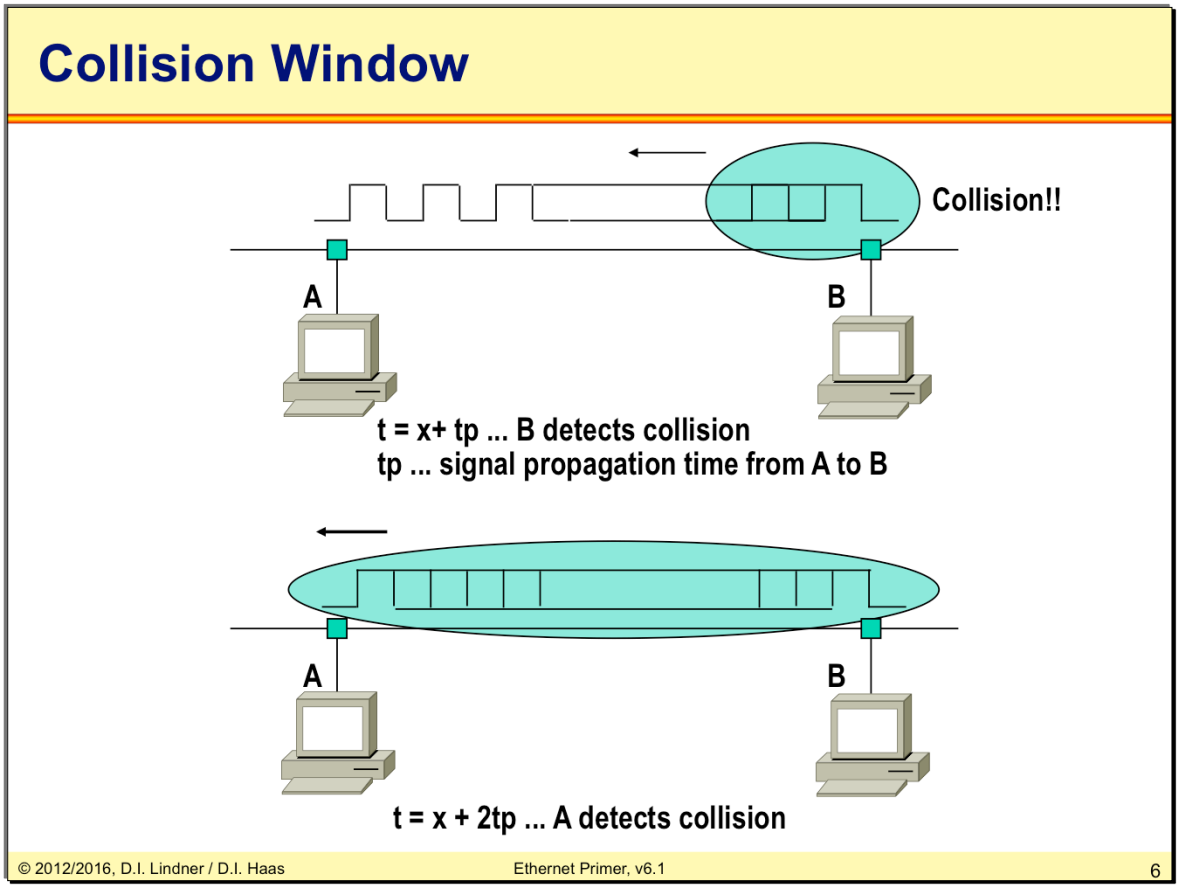
The way it works is quite simple, every stations that wants to send need to do a Carrier Sense to check if the media is already occupied or not. If the media is available the station is allowed to perform an Media Access and may start sending data. In the case that two stations almost at the same time access the media, a collision will happen. To recognize and resolve a collision is the task of the Collision Detect circuit. Every station listens to its own data while sending. In the case of a collision the currently sending stations recognize the collision by the superimposition of the electrical signals on the wire. Collisions are detected (CD) by observing the DC-level on the medium. A jamming signal will be sent out to make sure all involved stations recognize the occurrence of an collision. All stations involved in the collision stop sending and start a randomize timer. When the randomize timer expires the station may try to access the media again.

.



Collision Detection details:

In coaxial Ethernet, transceivers send their Manchester code using the DC offset method. A "high" value is nominally zero current; a "low" value is nominally -80 mA. This results in a DC component to the signal of -40 mA, which creates a voltage of -1 VDC (the transceiver sees a 25 ohm load from the two 50 ohm cables going "left and right" away from the transceiver). When two transceivers send at the same time, their currents add, increasing the DC component of the combined signal to -2 VDC Thus, we can detect collisions by looking for DC signals in excess of the maximum that could possibly be generated by a single transmitter.



In the worst case stations have to send bits twice the maximum signal propagation time (RTT) for reliable collision detection. Otherwise a collision may not be seen by the transmitting station for the currently transmitted frame.

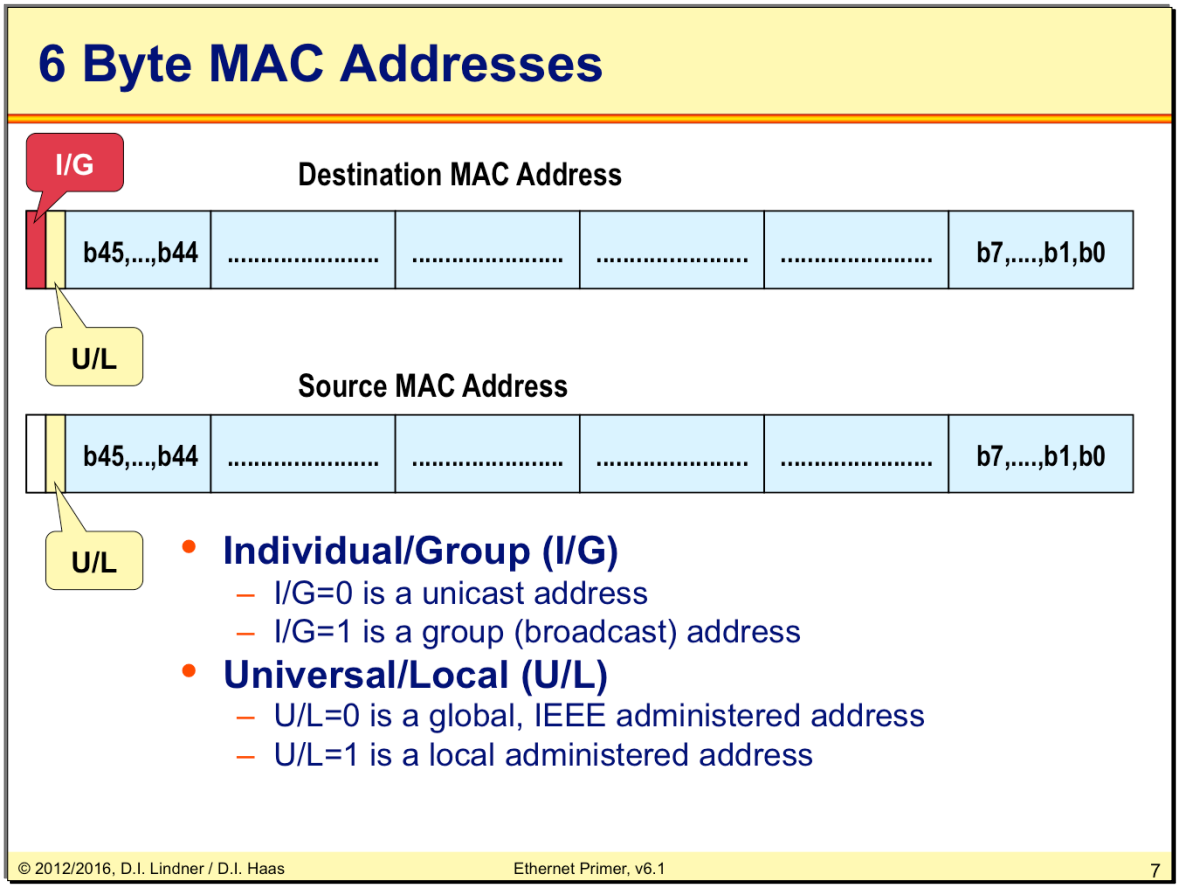
There is a very basic Ethernet rule that says a collision must be detected while a station is transmitting data. Therefore a stations needs to keep on sending at least of the duration of the RTT of the Ethernet system. If collisions occur after expiration of the slot time we talk about “late collisions”, which may cause malfunctions in the network. For example if a station transmits a frame and no collision was detected, the station assumes correct delivery of the frame. Now the station removes the frame from the transmit buffer, leaving no chance to retransmit the frame in the case of a late collision.

The maximum allowed RTT is standardized and is called the slot time. The slot time for 10Mbit/s Ethernet systems is set to 51,2 µs = 512 bit-times. Hence minimal Ethernet frame length is 64 byte. The request for fairness limits the maximum Ethernet frame size, too. 1518 byte is the maximum allowed Ethernet frame size.

With signal speed of 0.6c and the delay caused by electronic circuits such as interface cards and repeaters the slot time allows a maximum network diameter of 2500 meters for 10 Mbit/s. The maximum network diameter of faster Ethernet systems is directly related to their shorter slot times, because of the higher speed e.g. 250 meters for 100 Mbit/s (Fast Ethernet), 25 meters for 1000 Mbit/s (Gigabit Ethernet).

These distance limitations must only be taken into account in shared media environments like

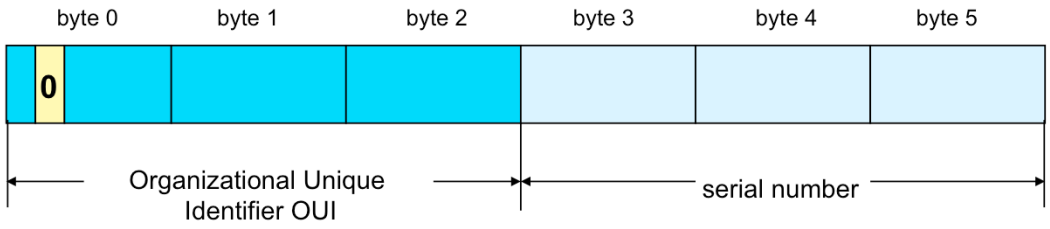
Ethernet Primer (v6.1)



Every station on a LAN is identified by a unique MAC-address which may be used either as source or destination MAC-address in LAN frames.

A MAC address is 6 bytes or 48 bits long and is typically written in hexadecimal notation. The first two bits of a MAC address on the have a special meaning. The first bit (I/G) specifies whether the MAC address is a unicast address (0) or a broadcast/multicast address (1). A multicast is a broadcast for a group whereas broadcast addresses all stations on a single LAN. The broadcast-address is an address with all bits set to 1 (hex FFFF FFFF FFFF, U/L is set to 1). Please recognize that bit 47 (I/G) has no meaning in the source address of a LAN frame. The second bit (U/L) specifies whether it's a global and unique MAC address administered by the IEEE, or a local administered address.

MAC Address Structure



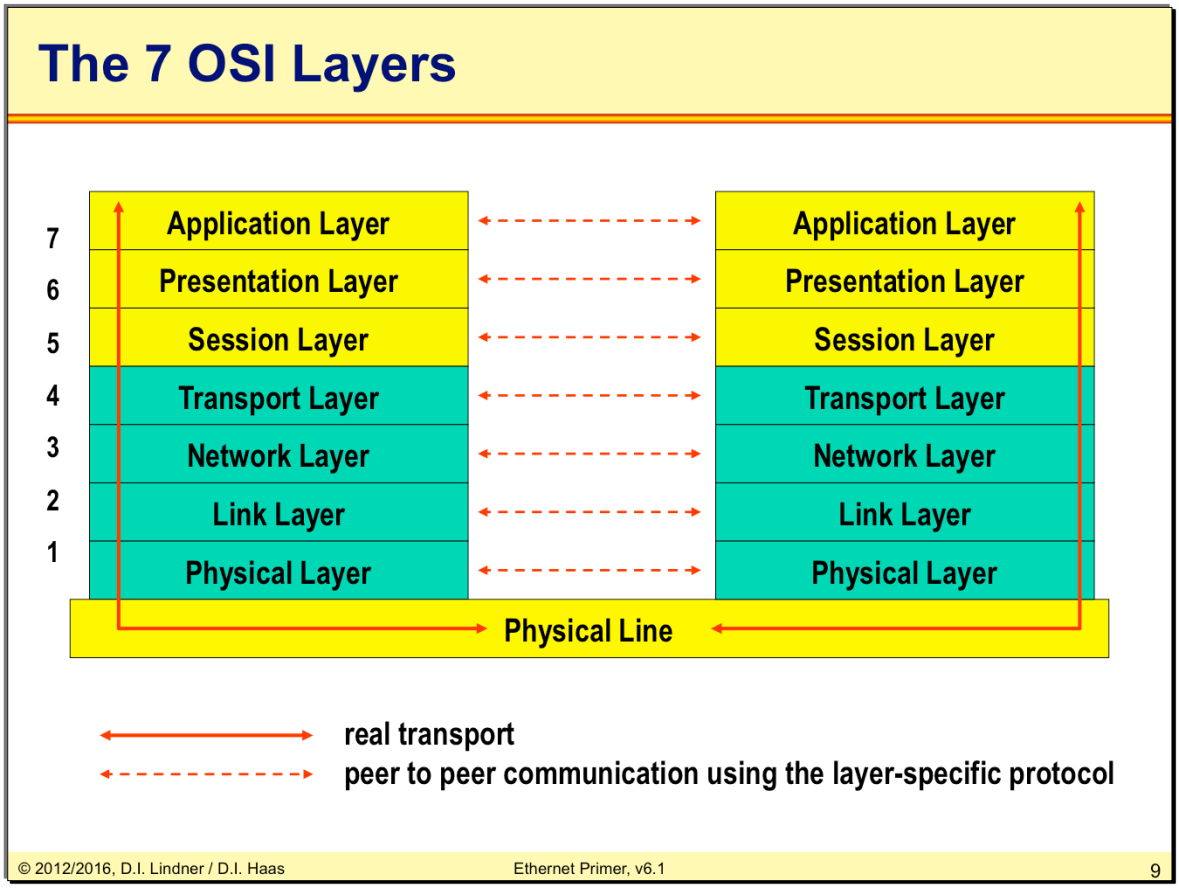
- **Each vendor of networking component can apply for an unique vendor code**
- **Administered by IEEE**
- **Called “Burnt In” Address (BIA)**

 Ethernet Primer, v6.1 8

The MAC addresses are globally administered by the International Electrical and Electronic Engineering (IEEE) standardization organization.

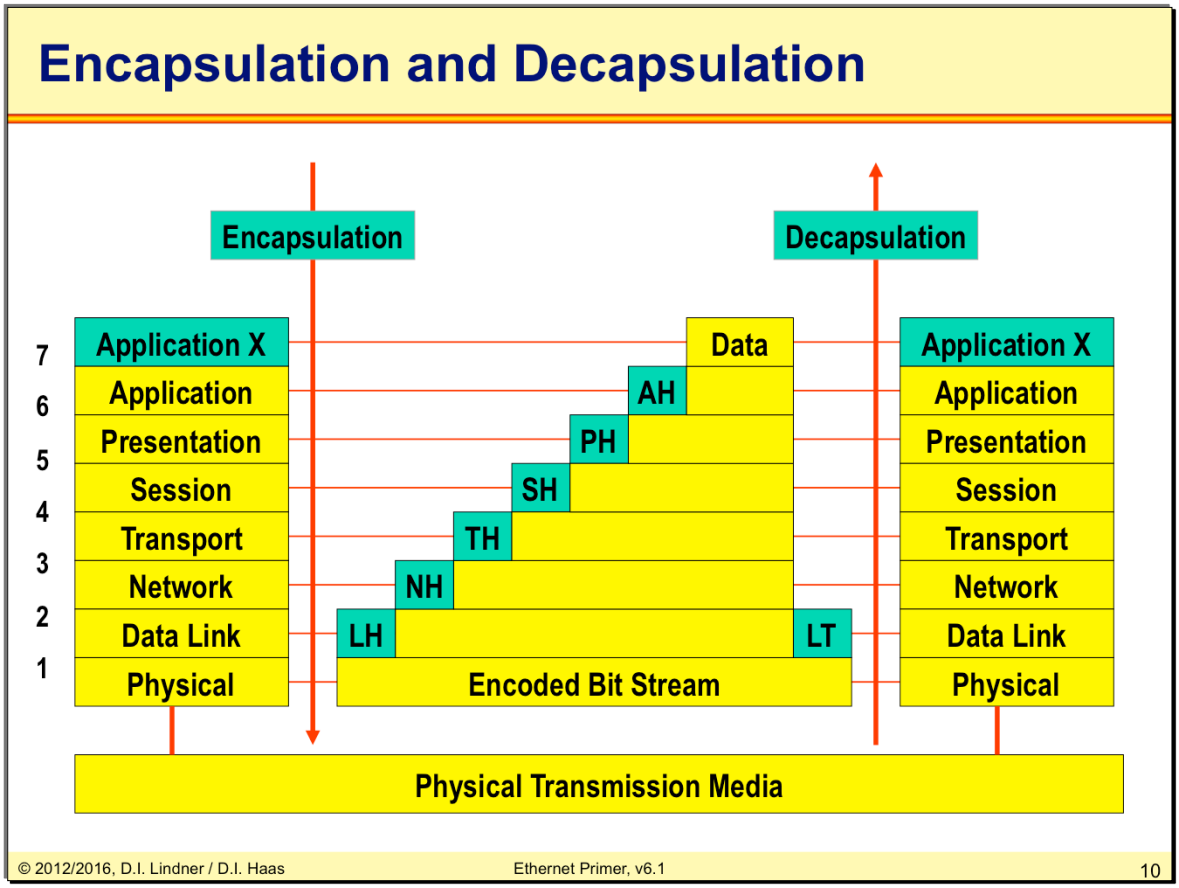
Each vendor of networking components can apply for an globally unique vendor code. The vendor code costs 1000$ and occupies the first three bytes of the MAC address.

The remaining three bytes of the MAC address may be used by the vendor to address its components. Every Ethernet network card has one burnt in MAC address (BIA). Network cards of some vendors even support the use of programmable local administered MAC addresses.



Because the communication between different systems can be a very complex task, OSI splits the communication aspects into smaller tasks. All layering is based on the OSI reference Model, which defines tasks and interactions of seven layers.

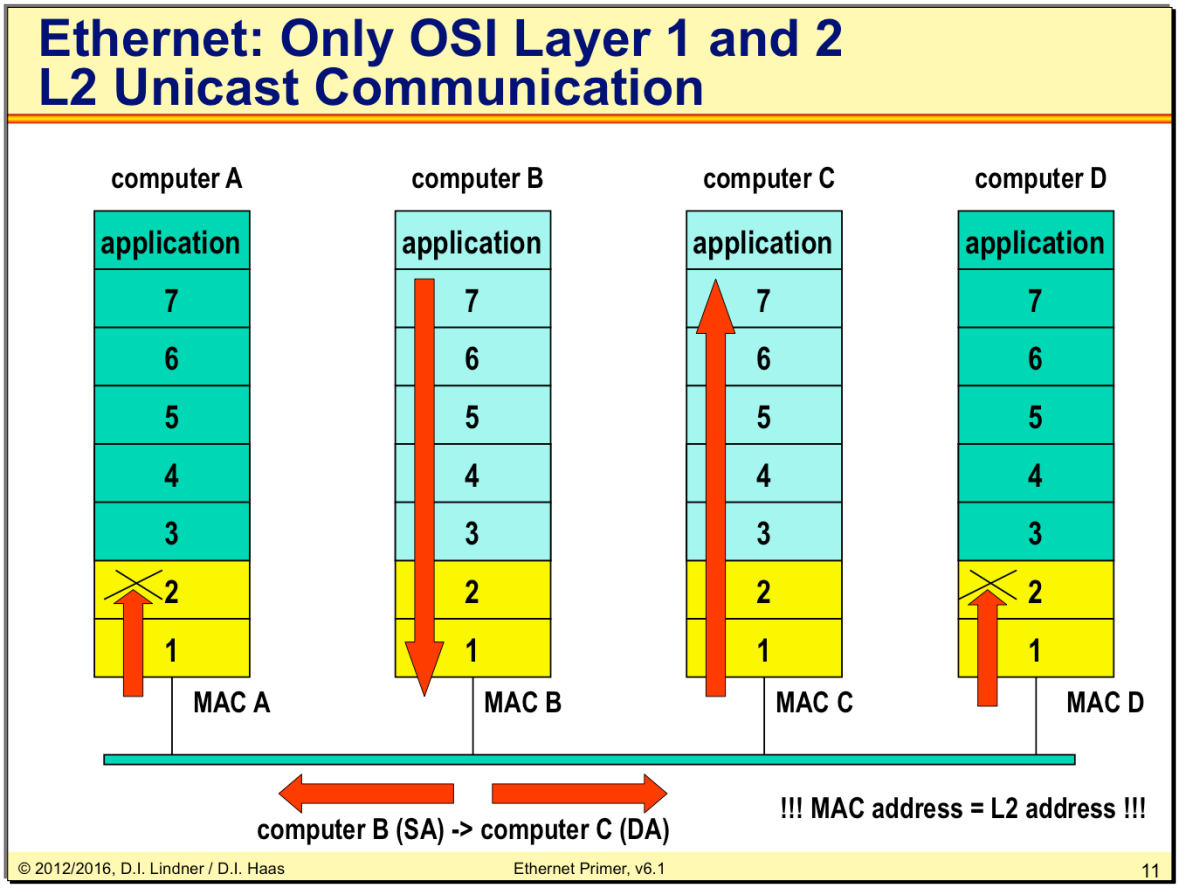
The user's data moves from the first layer (Application Layer) through all other layers. When two systems communicate with each other, then only the different layers talk. The application layer only talk with the application layer or the network layer only communicate with the network layer of system B. We can talk about a parallel communication between the layers. Every layer works for its own, it is not interested what the other layer does.



One of the most important principles:

Every layer adds its own protocol header by going downstairs in the stack -> Encapsulation at the source.

Every layer removes its own protocol header by going upstairs in the stack -> Decapsulation at the destination.



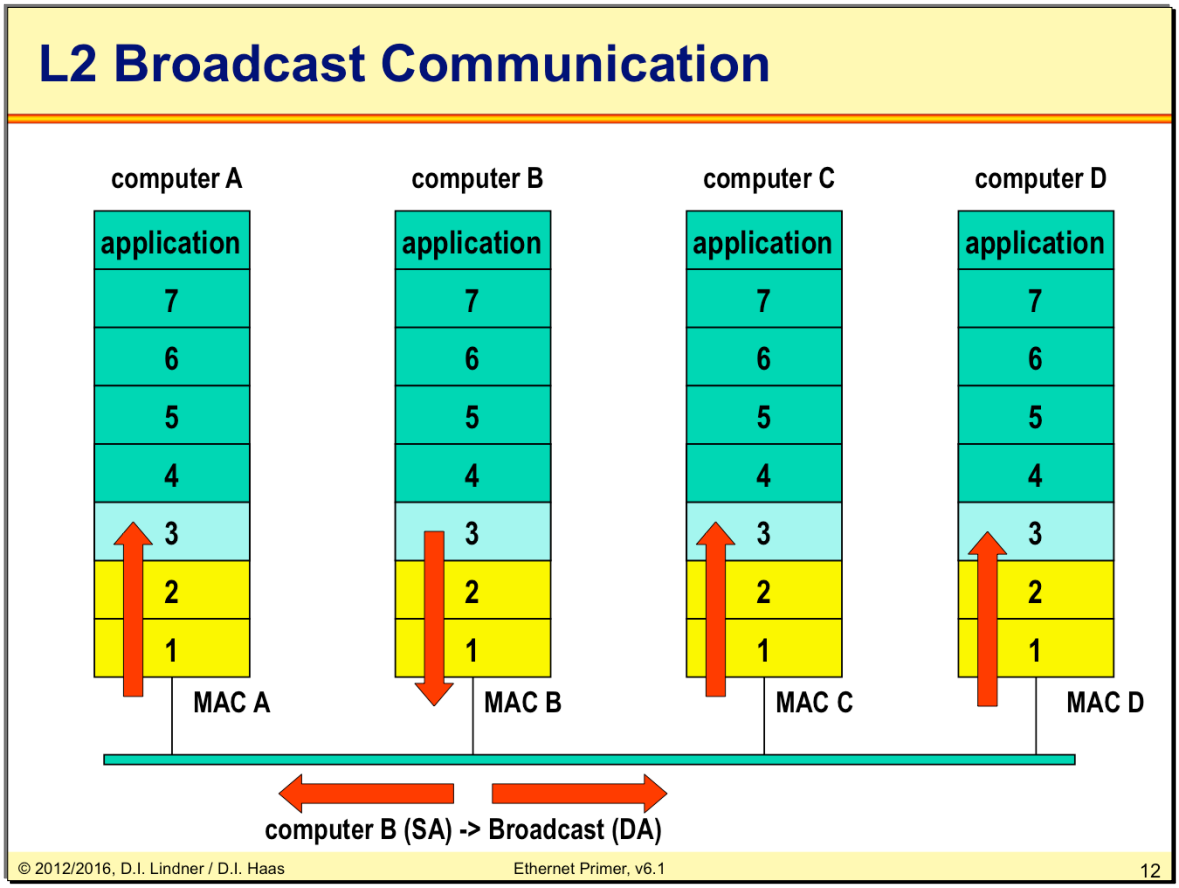
Receipt of frames:

Because of the inherent broadcast behavior of a LAN every frame is received by the Network Interface Card (NIC) of a station. The NIC decides if a frame should be forwarded to the higher layers (3-7) of a station depending on its own BIA and the destination address of the received frame. Usually NIC interrupts the CPU of the station if frame is to be forwarded. Otherwise the received frame is silently discarded by the NIC.

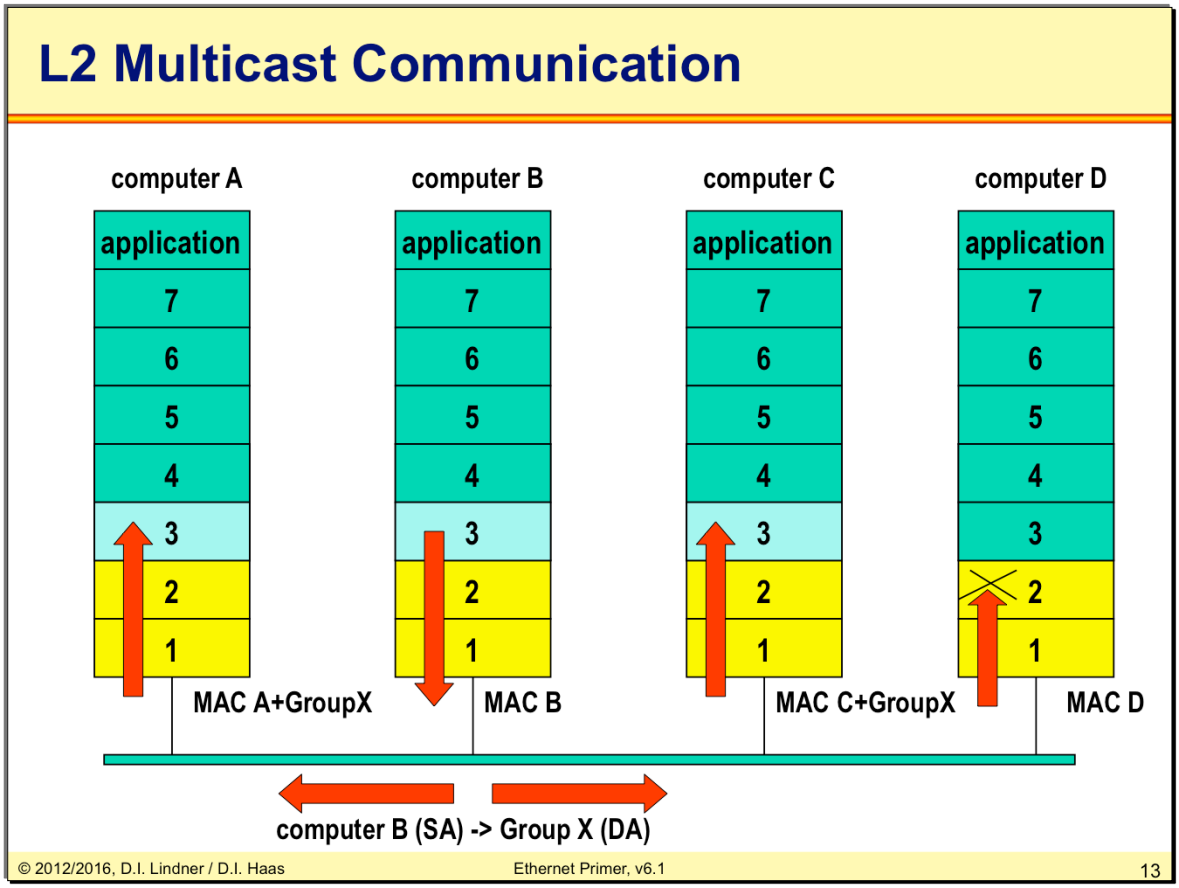
Frame are only forwarded to the higher layers (3-7) :

1. The destination address of the frame is equal with own BIA address
2. The destination address was a broadcast address
3. The destination address was a multicast (group) address and the given station is member of such a group.

The later two are seen on the next two slides.



Keep in mind that for normal operation frames should be destined to a station specific MAC-addresses (direct communication), because broadcast frames will interrupt all stations for further handling by the higher layers. Even if it turns out that a station needs not to act on such a broadcast frame the CPU time of this stations will be wasted. Broadcast should be used in initialization phases of a network only!

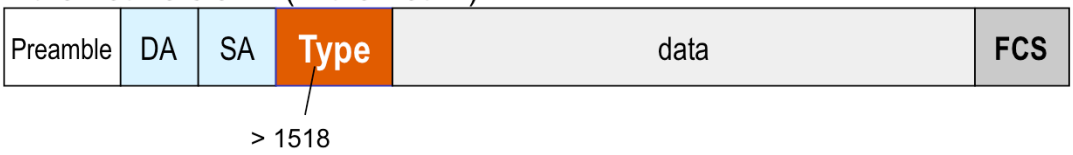


In this example computer A and C are programmed to listen to group address X. Computer D will not be disturbed by this frame because it is not programmed to listen to X.

Ethernet Primer (v6.1)

Ethernet Version 2 Frame

Ethernet Version 2 ("Ethernet II")



> 1518

Preamble ..... For clock synchronization (64 Bit)
DA ..... Destination MAC-Address (48 Bit)
SA ..... Source MAC-Address (48 Bit)
Type ..... Protocol-type field (16 Bit)
Data ..... Payload
FCS ..... Frame Check Sequence (32 Bit)
based on CRC (Cyclic Redundancy Check)

 Ethernet Primer, v6.1 14

Some more Ethernet parameters:

Interframe gap between to Ethernet frames is 9.6 microsecond.

Jam size is 32 bit.

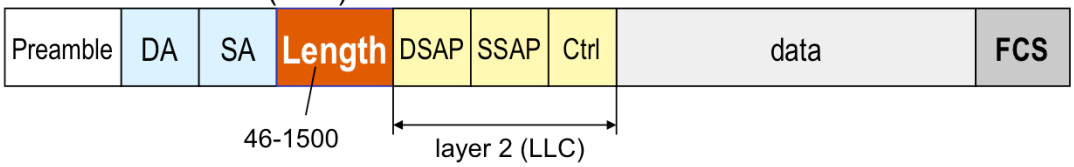
Slot time is 512 bits, minimal frame length 64 Byte

Maximal frame length 1518 byte (6+6+2+1500+4)

Maximum number of frames per second on a 10 Mbit/s Ethernet: 14880 frames of minimal length (6+6+2+46+4, FCS counted, preamble not counted)

IEEE 802.3/802.2 Frame

802.3 with 802.2 (LLC)



Length field instead of Ev2 type field
Additional 802.2 LLC header for implementation of HDLC-like protocol procedures on LAN wire

Length ..... Of IEEE 802.3 frame (16 Bit)
= bytes following without FCS (46-1500)

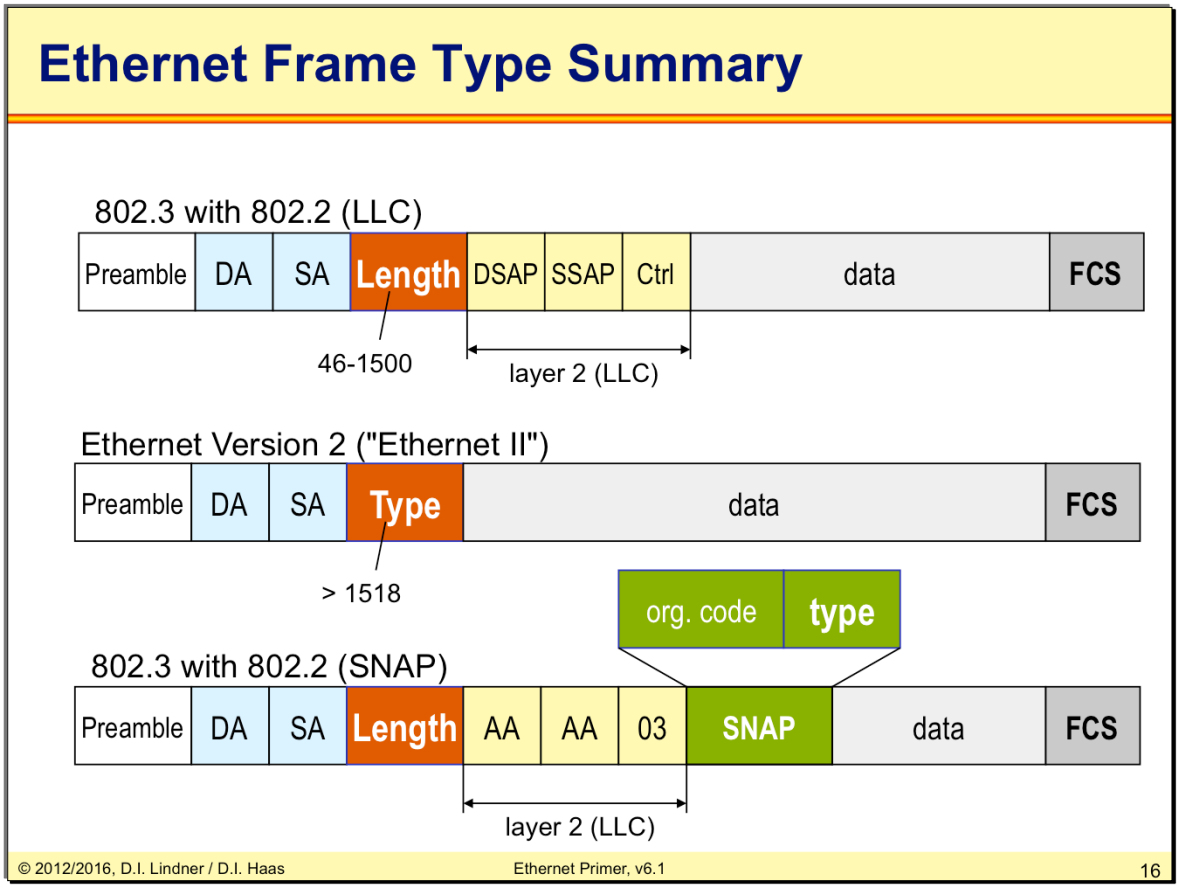
LLC ..... Logical Link Control (HDLC for LAN)
DSAP (Destination Service Access Point) 8 Bit
SSAP (Source Service Access Point) 8 Bit
Ctrl (Control) (8 or 16 Bit)

- **Ethernet II and 802.3 can coexist on the same cable, but each associated sending and receiving station must use the same format.**
- **Fortunately all type-field values are larger than 1518 (max frame length), so any incoming frame can be recognized and handled properly.**

 Ethernet Primer, v6.1 15

Remember IEEE 802.3 relies on LLC (802.2) and SAPs -> the protocol-type is indicated by SSAP and DSAP and the LLC control field can provide connectionless and connection-oriented services to the upper layers.

Ethernet Version 2 uses a protocol-type-field instead of the length field and lacks from any kind of HDLC like control field. Therefore only connectionless services can be provided by Ethernetv2 to the upper layers.



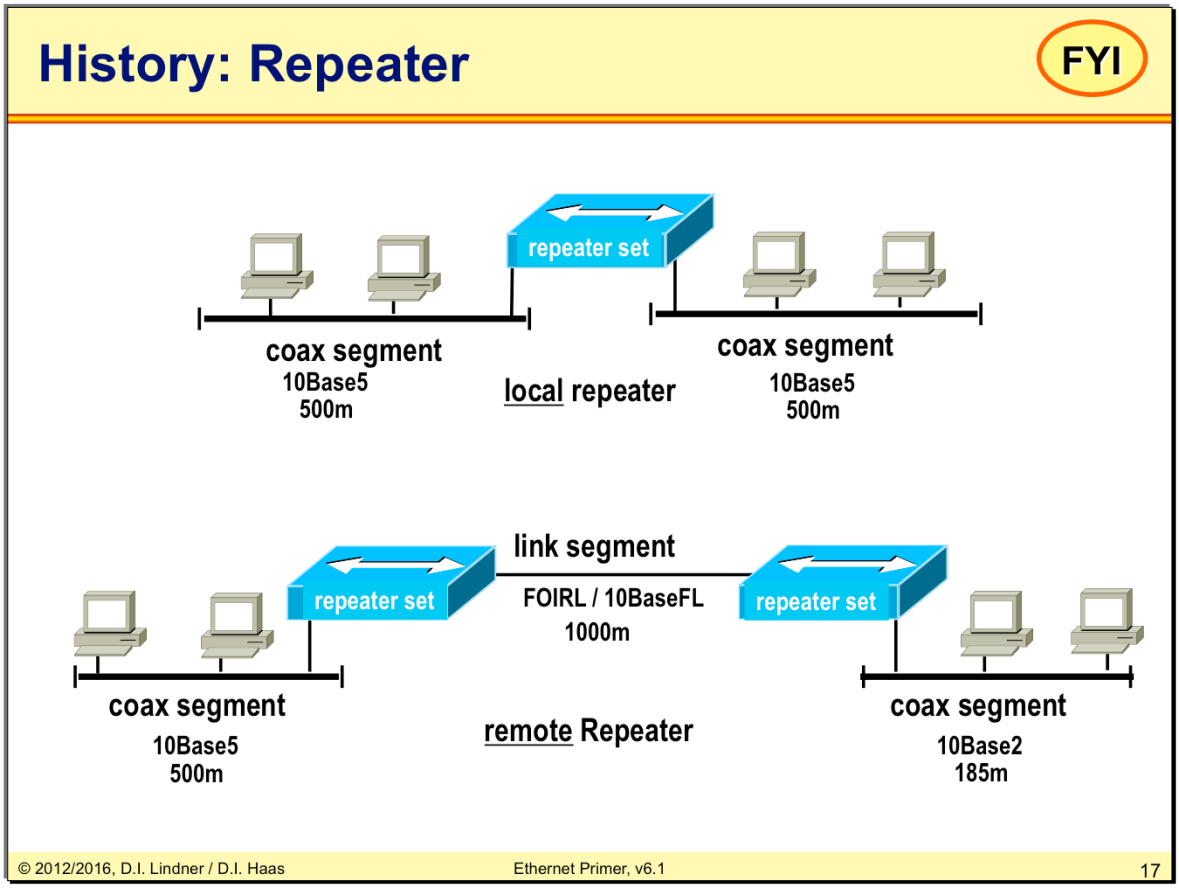
So we end up with three different frame formats used in Ethernet systems.

The 802.3 / 802.2. LLC without SNAP, the Ethernet v2 format and the 802.3/802.2 LLC with SNAP.

The DIX Eth 2 frame format is mainly used for the data transport of protocols that have the functionality of error recovery and flow control implemented in their protocol stack e.g. IP.

The 802.3/802.2 without SNAP frame format is used for protocols that need the functions of error recovery and flow control on layer 2 e.g. NetBeui, SNA.

The 802.3/802.2 with SNAP frame format is used by vendors to implement proprietary protocols, for example Cisco's CDP, VTP, CGMP, etc. protocols. For such purposes the OUI field is used to indicate the vendor and the type field value is chosen vendor specific.



Repeater is an amplifier expanding the maximal distance of an Ethernet-LAN segment. It regenerates signals on the receiving port, amplify them, and send these signals to all other connected net segments (no buffering, just a short delay, which must be taken into account for the collision window / slot-time). A repeater is an active network element which needs electrical power for its operation. In case a collisions is detected all other ports are notified by jam-signal. Optionally auto partition on erroneous ports may be performed by a repeater.

Link segment: a physical point-to point connection between two devices

Coax segment: the shared media as bus for network stations

Local repeaters directly connect two (coax) segments. Remote repeaters are connected by so called linked segments which are point-to-point links between repeaters.

First link segments were used for repeater interconnection only. Several types were defined (fiber based, copper based):

FOIRL (Fibre Optic Inter Repeater Link: maximal length 1000m, for repeater - repeater)

10BaseFL (asynchronous, maximal length 2000m for repeater - repeater, end system - multiport repeater)

10BaseFB (synchronous (idle signals during communication pauses), maximal length 2000m, for repeater - repeater links only)

10BaseFP (passive hub, no active repeater function)

Important: Collision domain is preserved by repeaters.

A repeater with more than two segments and different physics is called a multiport repeater. Multiport repeater in a star like topology is called a “Hub”. Be careful using this expression because it is also used for L2 Ethernet-Switch which is a packet switch but not an amplifier like a repeater.

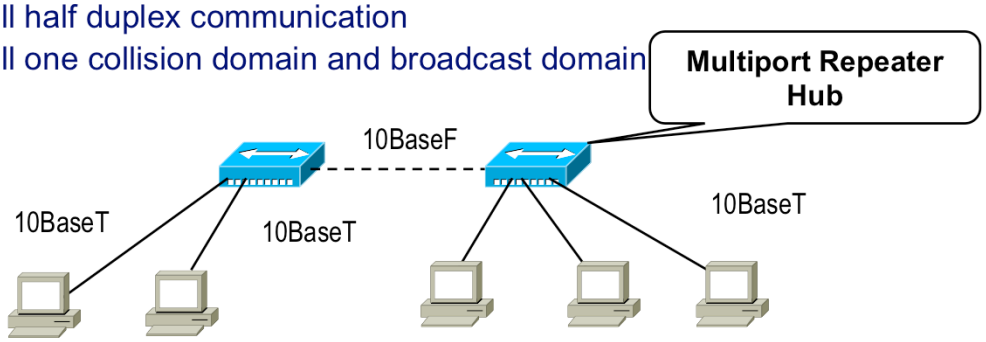
History: Multiport Repeaters

FYI

- Demand for structured cabling
  - 10BaseT (Cat3, Cat4, ...)
  - (voice-grade twisted-pair)
- Star-shaped topology
- 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

——— represents four CU wires 2 for Tmt, 2 for Rcv (e.g. 10BaseT, 100BaseT)

- - - - represents two FO wires e.g. 10BaseF, 100BaseF



 Ethernet Primer, v6.1 18

The link segment was later also defined for connection of a network station (end system) to a multiport repeater using a dedicated physical point-to-point line.

The reason: Ethernet was originally based on coax cabling and bus topology which was hard to wire in a building. Later an international standard for **structured cabling** of buildings was defined which was star wired to (a) central point(s) based on twisted pair cabling. Ethernet standard 10BaseT supporting structured cabling was 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).

10BaseT parameters: unshielded twisted pair, maximal length 100m, 2 lines Tmt+-, 2 lines Rcv+-, RJ45 connector, Manchester-Code with no DC offset

Collision detection details:

In such an environment the collisions can not be detected any longer just by measurement of the DC level as done in 10Base5 because tmt and rcv travels on different physical lines. Now a collision is interpreted, if signals are on the tmt and rcv line at the same time. The repeater has to watch out, if two or more signals are received at the same time (-> that means collision in the LAN). Now the hub has to produce a Jam signal on all ports in order to signal a collision to all systems.

In 10BaseT, the Manchester code is sent symmetrically, with no DC offset. Collisions are detected in the repeater hub, which can observe when two or more devices are transmitting at the same time. Normally, the hub does not repeat a station's own signal back to the station on its receive cable pair. However, when a collision is noted, the hub does send a signal (the so-called "collision enforcement", or "jam") to the transmitting stations. The stations detect collisions by noting when they see a signal on their receive pair at the same time that they are transmitting on their transmit pair.

Hub devices are 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.

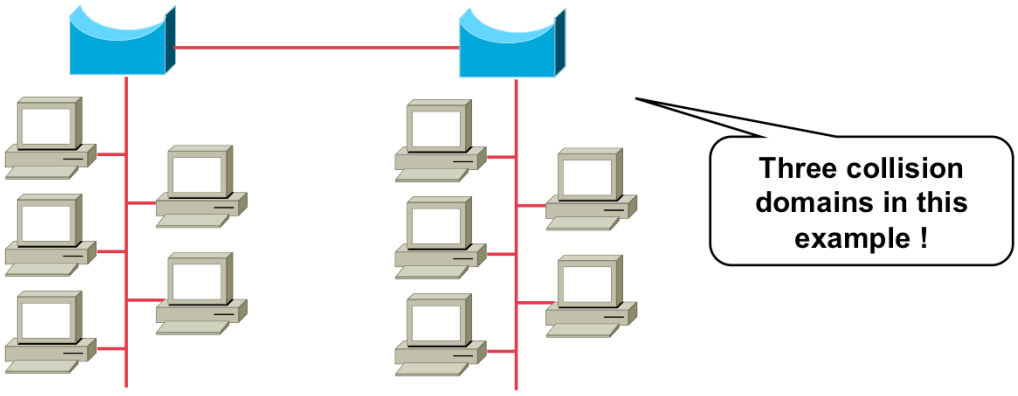
Note that the Ethernet topology became star-shaped.

Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed Ethernet

History: Bridges

- Packet switching principle (Store and Forward) based on MAC address information of Ethernet frames
- Bridge contains a MAC address table used for forwarding decision
  - Signposts to reach a MAC address by pointing to the corresponding physical port
- Bridging
  - Separates collision domains, improves network performance, but still is one broadcast domain



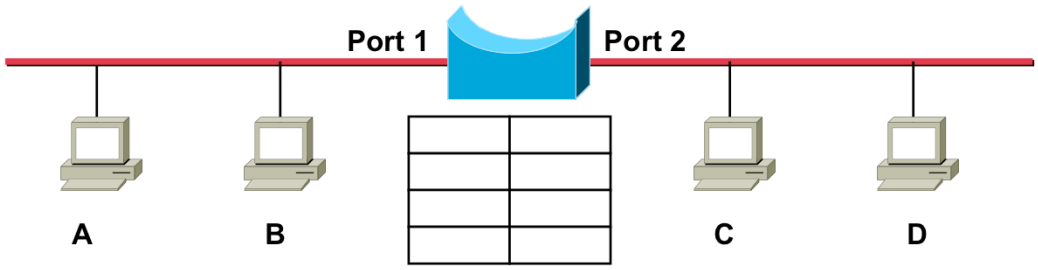
 Ethernet Primer, v6.1 20

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.

Transparent Bridging

- **Designed for "plug & play"**
- **Upon startup a bridge knows nothing**
  - MAC address table is empty
- **Bridge is in learning mode**
  - Dynamic entries are built on the fly



 Ethernet Primer, v6.1 21

The main advantage of transparent bridging is the transparency and "plug & play" capability. No end station notices the presence of bridges. Bridge is invisible for end stations. LAN Left and LAN Right appear to the end systems like one single, logical, big LAN. But because of this transparency a bridge must receive and process every frame on a LAN. This means much more performance is needed for a bridge than for a router which is explicitly addressed. Also flow control between end systems and bridges was not defined in the original implementation.

Transparent bridge uses layer 2 MAC-addresses to decide if a given frame must be a forwarded or not -> destination-address of a frame is used for this decision.

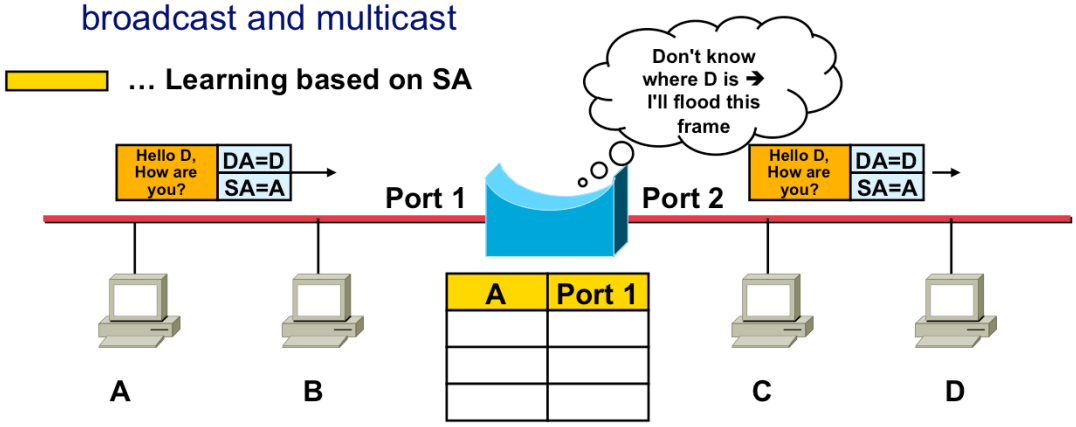
But in order to be invisible, bridges must also learn somehow where end stations are located.

MAC-addresses of all stations are registered in a bridging table either statically done by administrator e.g. for security reasons or dynamically done by a self-learning mechanism.

Following the self-learning mechanism upon startup, a bridge knows nothing and the bridging table is empty. At this time the bridge is in learning mode. For learning the source-address of a frame will be used.

Learning

- **Once stations send frames the bridge notices the source MAC address**
  - Entered in bridging table
- **Frames for unknown destinations are flooded**
  - Forwarded on all ports. Same rule applies to Ethernet broadcast and multicast



 Ethernet Primer, v6.1 22

Assume we have a bridge with only two ports, each attached at one Ethernet segment. Assume the left station "A" sends one frame to "D" on the right side. Obviously the bridge learns the location of A but has no idea where D is. Thus the MAC address of A is entered in the bridging table and also the port number "1", on which A is reachable. Since the location of D is unknown, the bridge loods this frame over all ports, in our case only to port two (as there are no other ports).

This way, connectivity is granted even if there is no entry in the bridging table.

Destination address of a frame is used for MAC address table look up in order to decide what have to be done with the received frame. The following actions are possible:

Filtering: frame will be rejected if destination‘s home is on the LAN segment of the receiving port

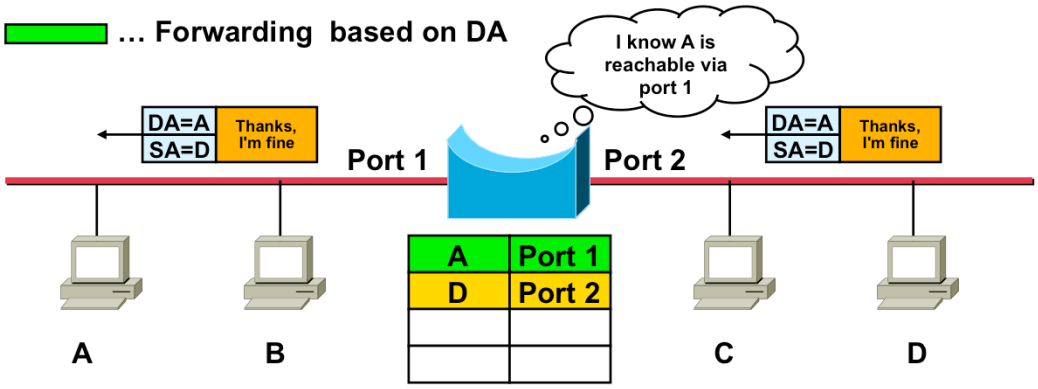
Forwarding: a duplicate of the frame will be forwarded to the appropriate port if destination‘s home is registered in the table of another port

Flooding: during learning time; frame will be forwarded to all other ports (multiport-bridge) if there is no entry in the table (unknown destination).

Frames with broadcast/multicast-address are always flooded.

Learning ➔ Table Filling (1)

- **If the destination address matches a bridging table entry, this frame can be actively**
  - Forwarded if reachable via other port
  - Filtered if reachable on same port



 Ethernet Primer, v6.1 23

Now assume D replies to the message which has been received from A. The bridge knows already the port number over which A can be reached and forwards the frame accordingly. If A would be located on the same port as D then this frame would be filtered.

Ethernet Primer (v6.1)

Table Filling (2)

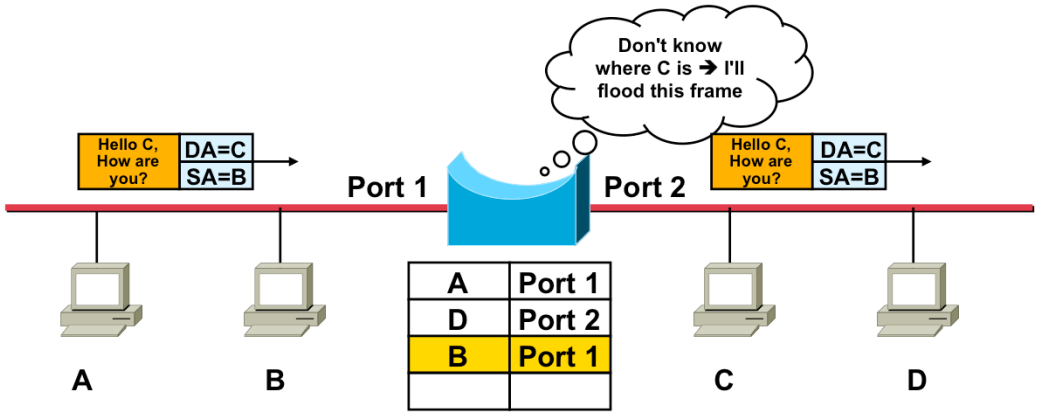
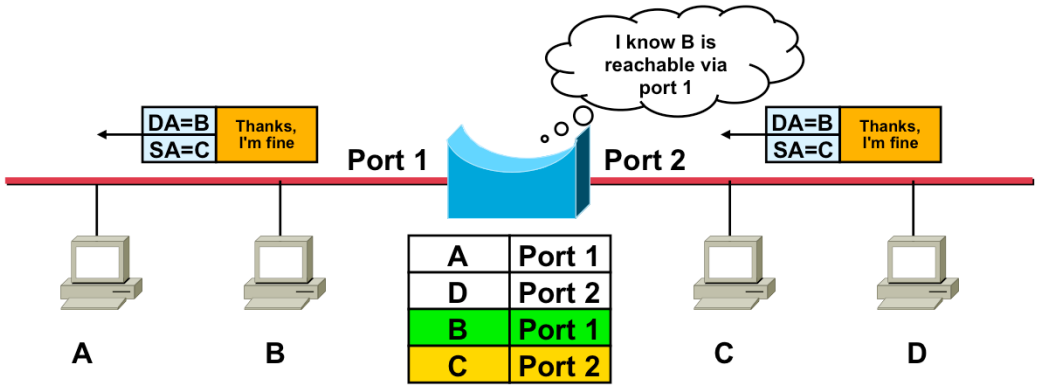


Table Filling (3)

- **After some time the location of every station is known – simply by listening!**
- **Now only forwarding and filtering of frames**
  - Based on destination address

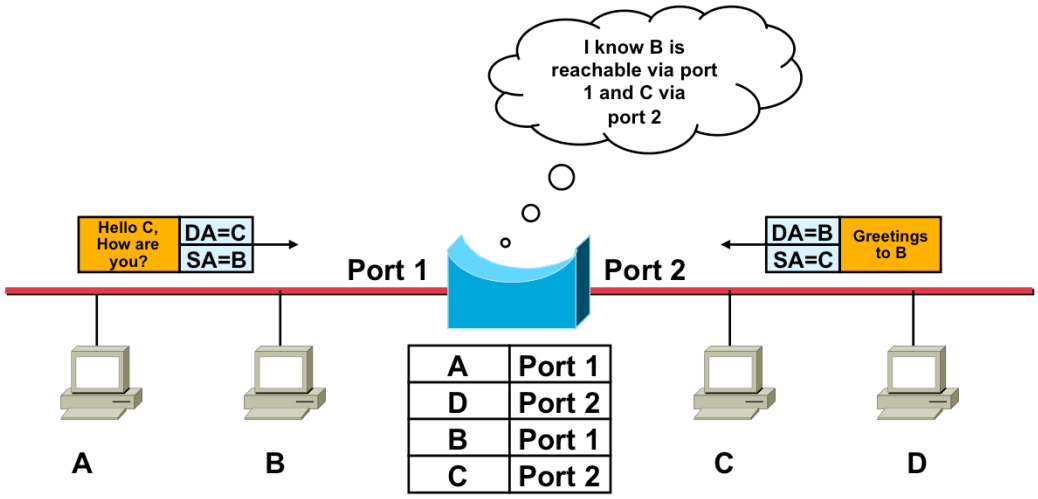


 Ethernet Primer, v6.1 25

After some traffic observing time, the bridging table contains all host locations (addresses and port numbers). At this time the bridge enters the forwarding and filtering mode.

Table Filling – Forwarding

- **Collision domains are separated**
  - Frames can travel in their LAN segments at the same time

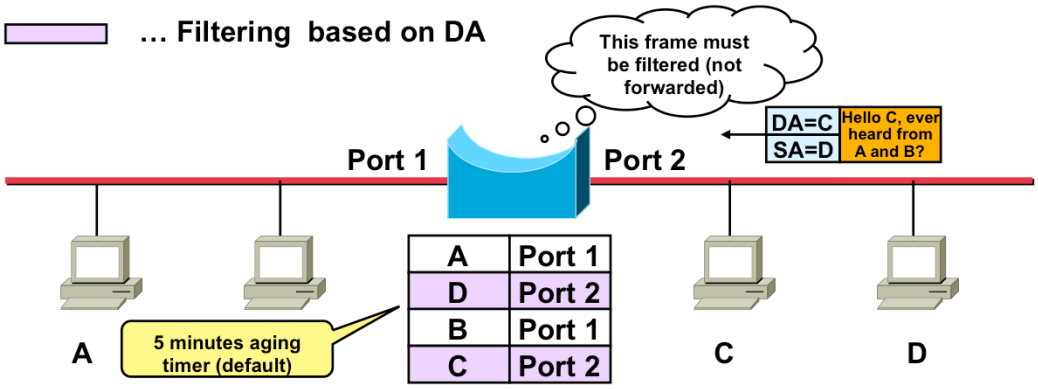


 Ethernet Primer, v6.1 26

Since only frames are forwarded to other ports whose destination is really located there, the LAN is separated into as many collision domains as ports are available (and attached to a LAN segment).

Table Filled – Filtering

- **Frames whose source and destination address are reachable over the same bridge port**
  - Are filtered
- **Entries times out**
  - If not refreshed within 5 minutes



 Ethernet Primer, v6.1 27

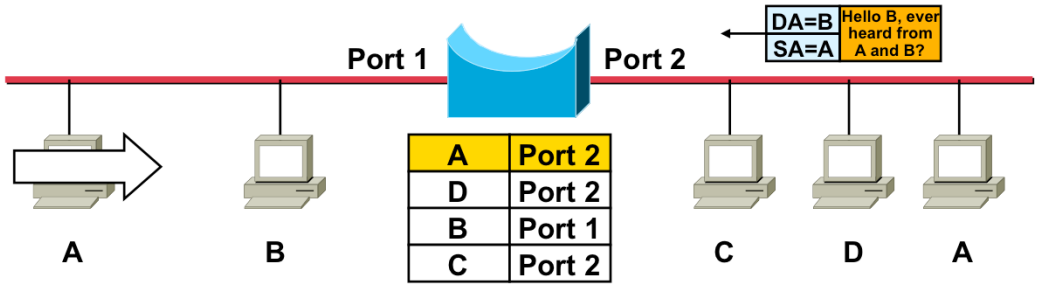
What if a host is removed from its location and attached at another place in the LAN? Obviously frames could be forwarded to the wrong port. Therefore each entry in the bridging table ages out after some time. The default aging time is 300 seconds or 5 minutes.

In case of a dynamic bridging table an aging mechanism allows for changes of MAC addresses in the network which may caused either by change of network card or by location change of end system. If an already registered MAC address is not seen within e.g. 5 minutes as source address of a frame the corresponding bridging table entry is deleted.

.

Last Seen – Last Win

- **Now assume notebook with MAC A is moved**
- **Address is immediately relearned**
  - With the first frame containing source MAC A on the other port
- **Imagine the problem**
  - If there are duplicated MAC address in the LAN !!!



 Ethernet Primer, v6.1 28

Duplicated MAC addresses would cause continuous table rewriting on a last seen – last win base.

Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed Ethernet

Bridging Problems

- **Redundant paths lead to**
  - Endless cycling of frames
  - Continuous table rewriting
  - Blocking of buffer-resources
  - Stagnation of the LAN
  - Broadcast storms
- **To eliminate these unwanted effects**
  - Spanning Tree Protocol (STP) is necessary
- **STP**
  - A must in bridged networks with redundant paths
  - Only one purpose: To cut off redundant paths

 Ethernet Primer, v6.1 30

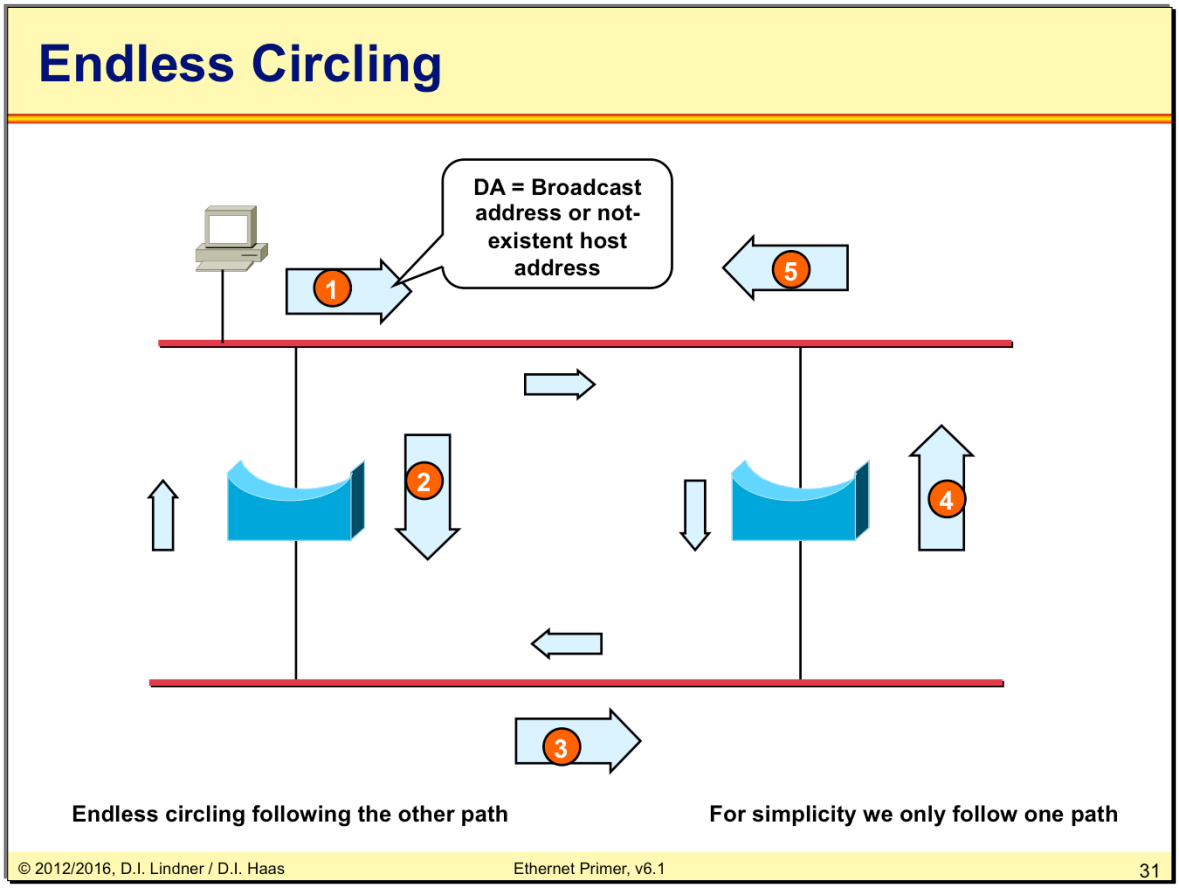
You might have noticed that bridges do not really learn the network topology. They only learn a simple destination to port association! Because of this there is no means to determine the best path, and furthermore frames might be caught in a loop.

Especially broadcast frames have no defined destination and would be forwarded over all parallel paths—endlessly! This results in endless circling of frames, or more dangerous, in a so-called "broadcast storm".

Also a continuous table rewriting might occur (this is not so widely known but also explained in the next pages).

Most people are not aware that frames might be stored up to 4 seconds inside the buffer of a switch—and it still complies to the IEEE standard. Although this would happen only in rare cases of congestion, transparent bridging is not suitable for hard realtime applications. Today the situation has changed, QoS features are included to assure bounded delays.

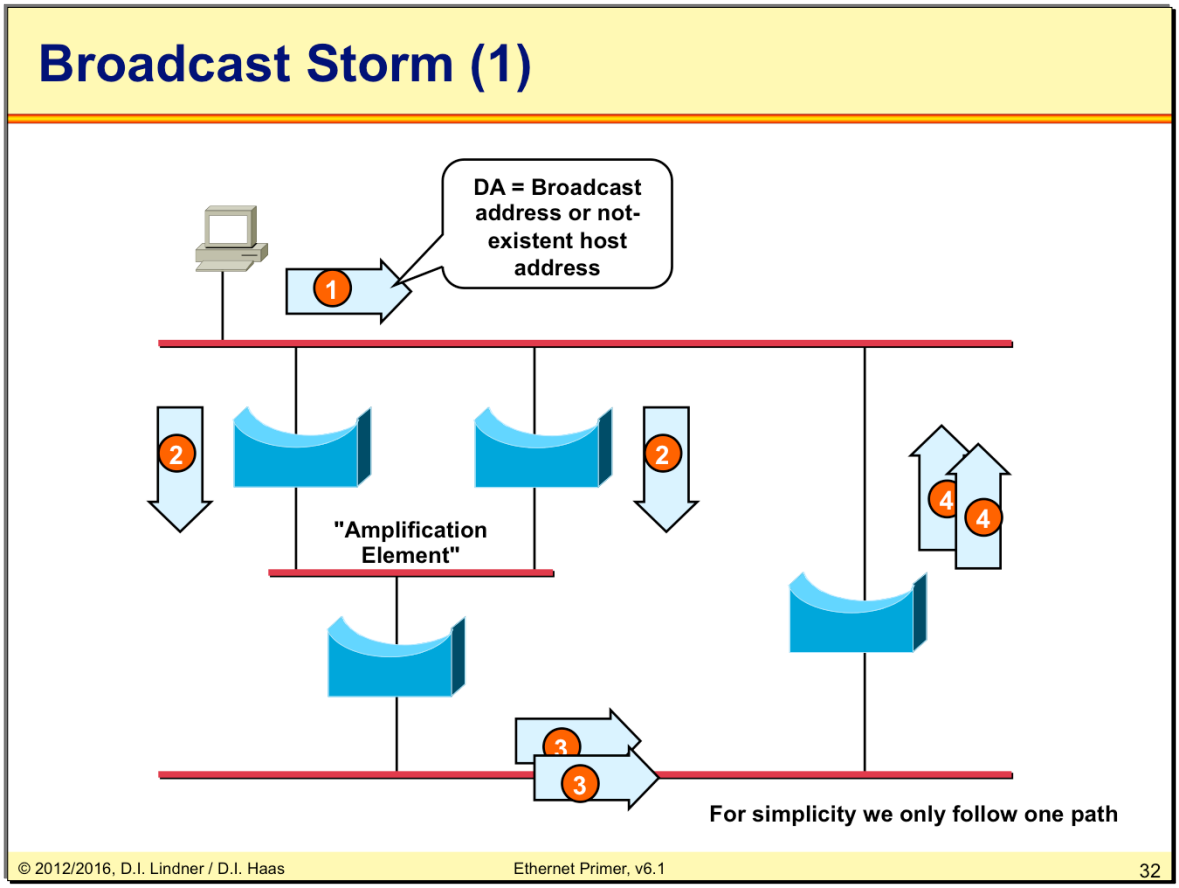
Ethernet Primer (v6.1)



The picture above illustrates the endless circling phenomena. Assume a network with parallel paths between two LAN segments, realized by two bridges. Any frame with a broadcast destination address would be forwarded by both bridges to the other segment and back and forth and so on.

Obviously endless circling leads to congestion problems an is not desired. Remember that there is not hop count or time-to-live number within the Ethernet header.

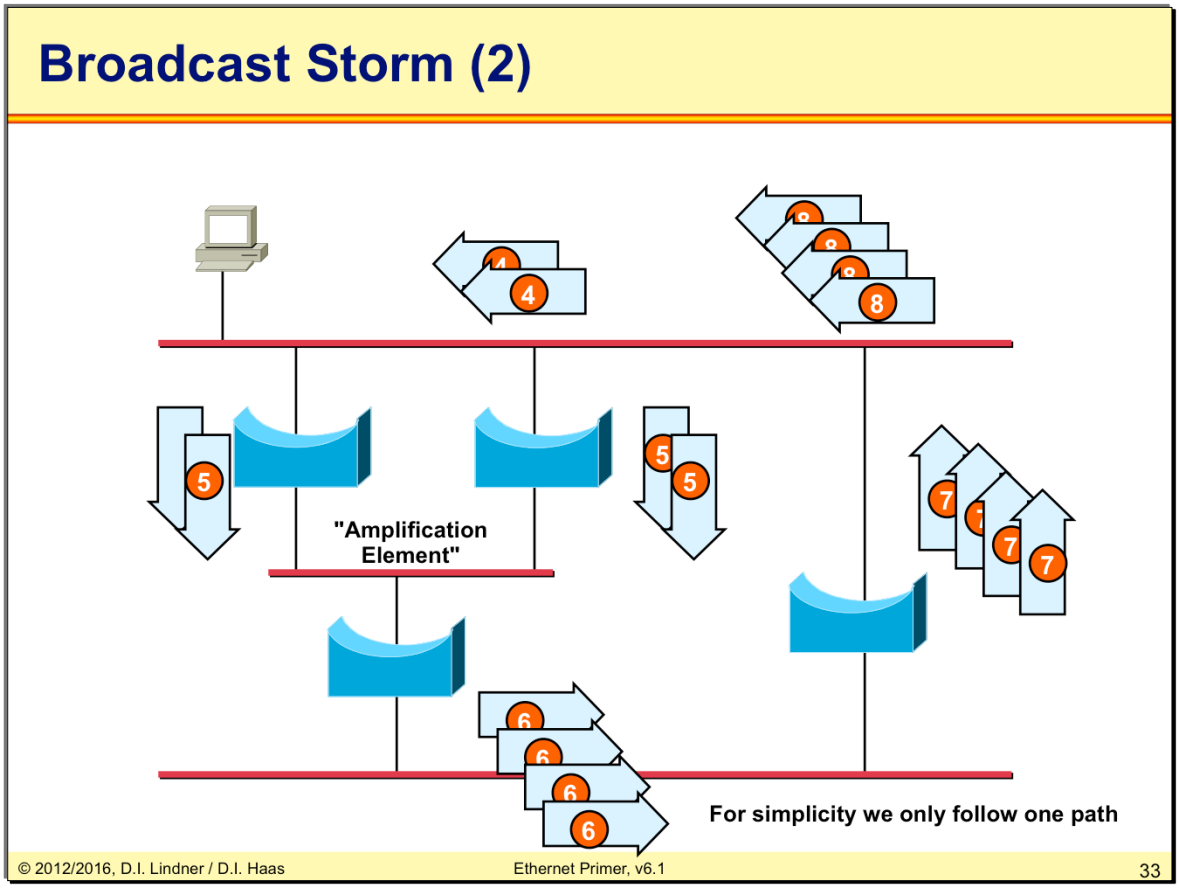
But endless circling is not the main problem... (see next slide)



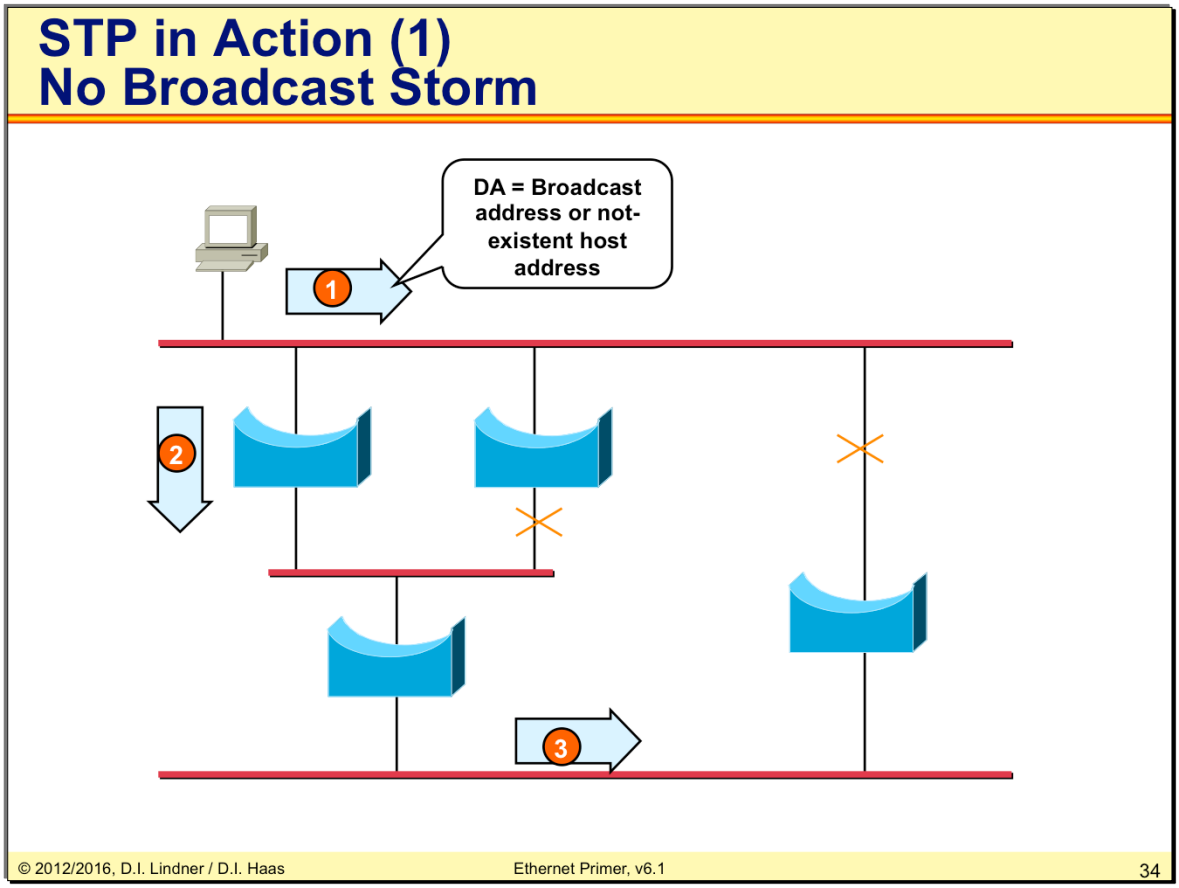
The most feared issue with bridging are broadcast storms. Broadcast storms can be considered as a dramatically "enhanced" endless circling problem. Broadcast storms appear when there is an "amplification" element within the network, such as those threefold parallel paths in the diagram above.

Within a very short time (e.g. 1 second) the whole LAN is overloaded with broadcast frames and nobody could transmit any useful frame anymore.

Ethernet Primer (v6.1)

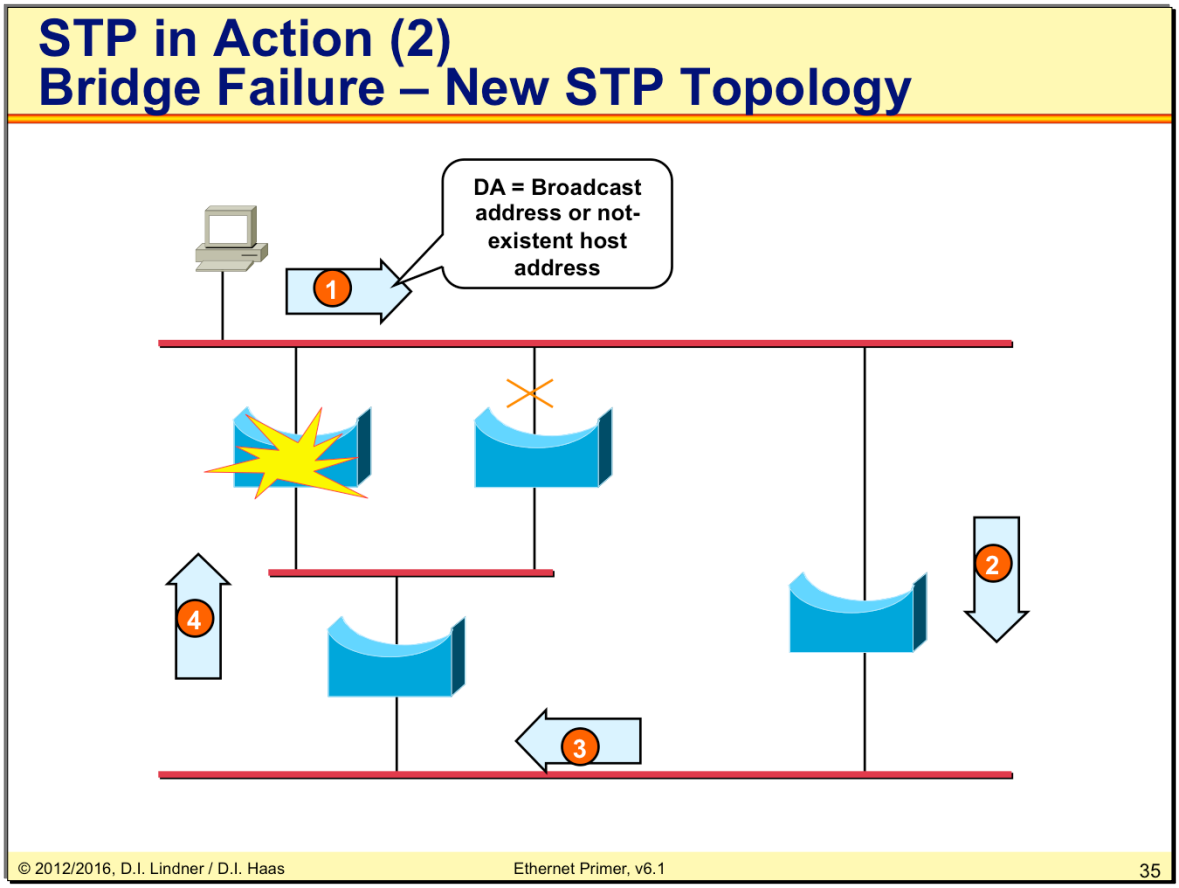


The picture above shows the amplification effect mentioned on the previous page.



STP eliminates redundancy in a LAN bridged environment by cutting of certain paths which are determined by the STP parameters Bridge ID, Bridge Priority and interface Port Costs. An easy way to achieve this is built a tree topology. A tree has per default no redundancy or have you ever seen leafs of a tree which are connected via two or more branches to the same tree?

Spanning Tree Protocol (STP) takes care that there is always exact only one active path between any 2 stations implemented by a special communication protocol between the bridges using BPDU (Bridge Protocol Data Unit) frames with MAC-multicast address. The failure of an active path causes activation of a new redundant path resulting in new tree topology.



Additional task of STP is to recognizes any failures of bridges and to automatically build a new STP topology allowing any-to-any communication again.

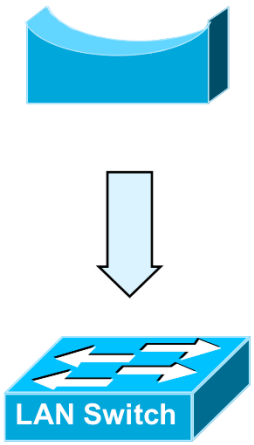
Here you can also see one main disadvantage of STP: Redundant lines or redundant network components cannot be used for load balancing. Redundant lines and components come only into action if something goes wrong with the current active tree.

Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed Ethernet

What is an Ethernet Switch?

- **A switch *is* basically a bridge, differences are only:**
  - Faster because implemented in HW
  - Multiple ports
  - Improved functionality (e.g. VLAN)
  - Different data rates supported simultaneously
    - 10, 100, 1000, 10000 Mbit/s depending on switch
  - QoS (802.1p)
    - Queuing mechanisms
    - Flow control
  - Security features
  - VLAN support (tagging, trunking, 802.1Q)
  - Spanning Tree (RSTP, MSTP, PVST+)
  - SPAN (for monitoring traffic)



 Ethernet Primer, v6.1 37

Now what is the difference between a bridge and a switch? Logically there is no difference. Ethernet switching is much more a marketing term to express something new to the customers but basically it is the same as a transparent bridge. An Ethernet switch is just faster and has much more ports than the good old bridge. Ethernet switches typically employ more than two ports, and the bridging functionality is implemented in hardware. Additionally other features like VLAN (Virtual LAN) support or NAC (Network Access Control) are added in modern Ethernet switches, depending on the vendor

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

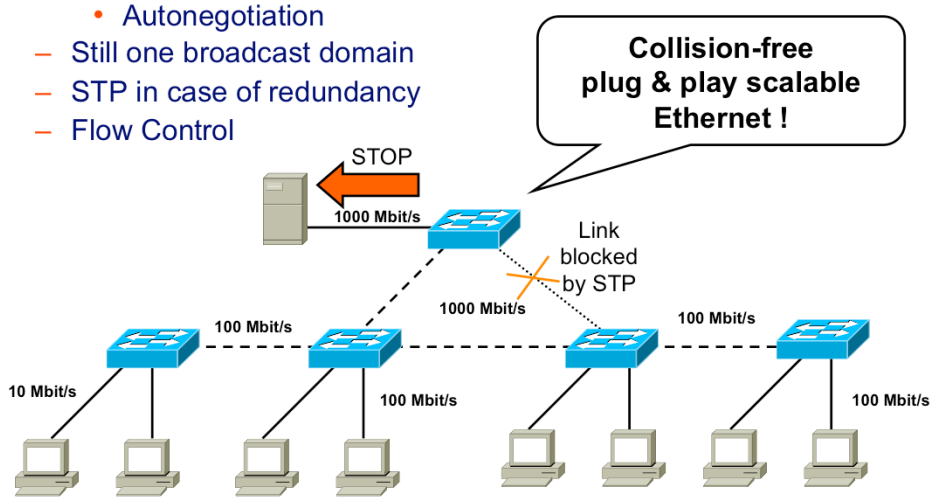
Security is provided by statically entered switching tables and port locking (port secure), that is only a limited number or predefined users are allowed at some designated ports.

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.

Ethernet Switch

- 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



 Ethernet Primer, v6.1 38

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.

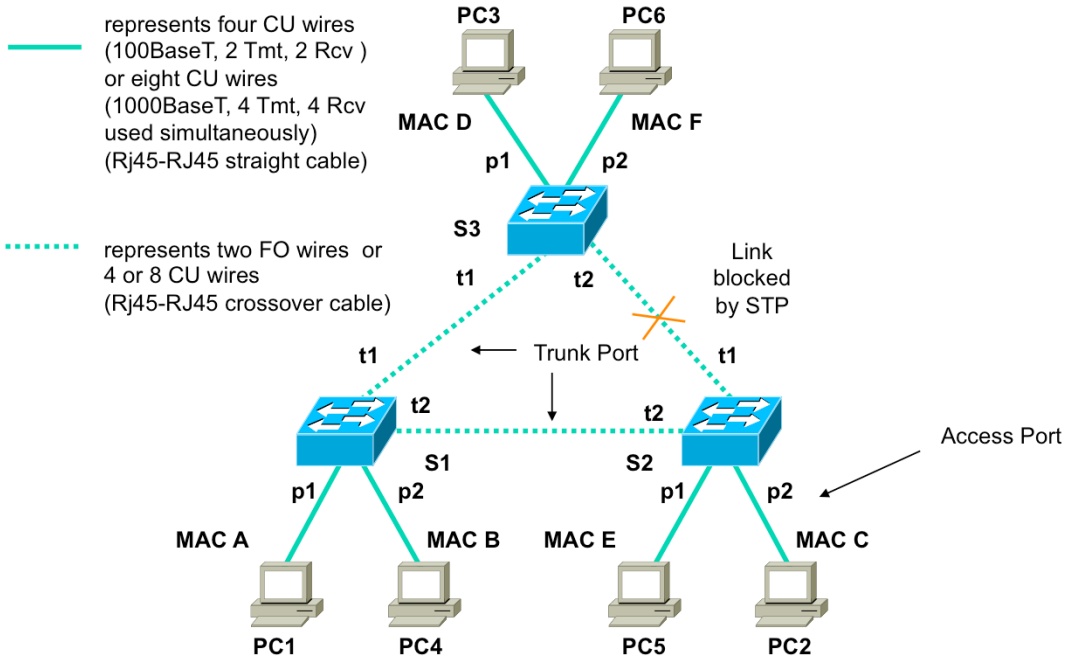
The next benefit of store and forward performed by L2 switches is that now different speeds on different links. Clients may use 10 Mbit/s, servers may use 100 Mbit/s or 1000Mbit/s and Interswitch links may go up to 10GBit/s speed nowadays. That was one reason for success of the Ethernet family. Even 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.

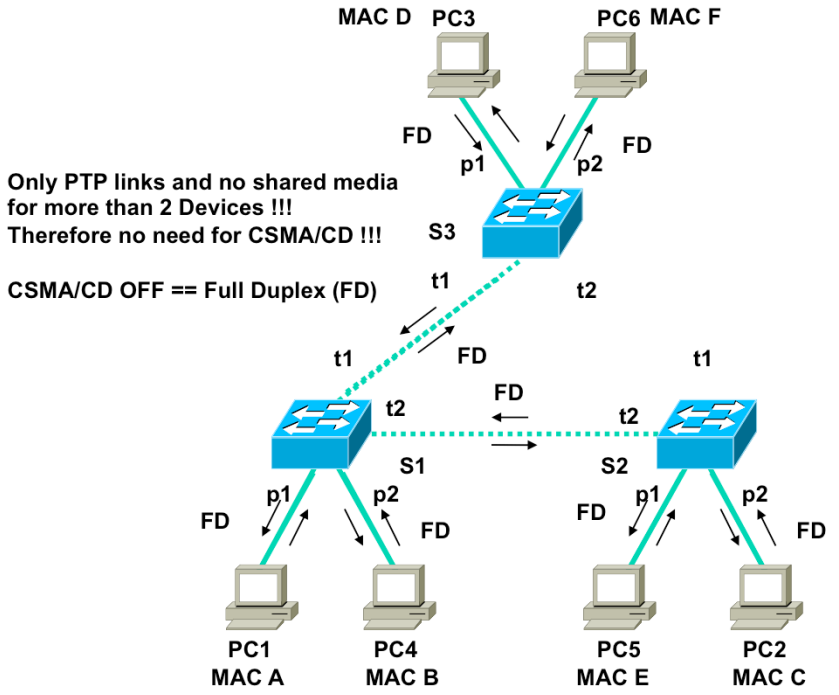
Ethernet Primer (v6.1)

Ethernet Switching Topology
(MAC Address Table - Empty)



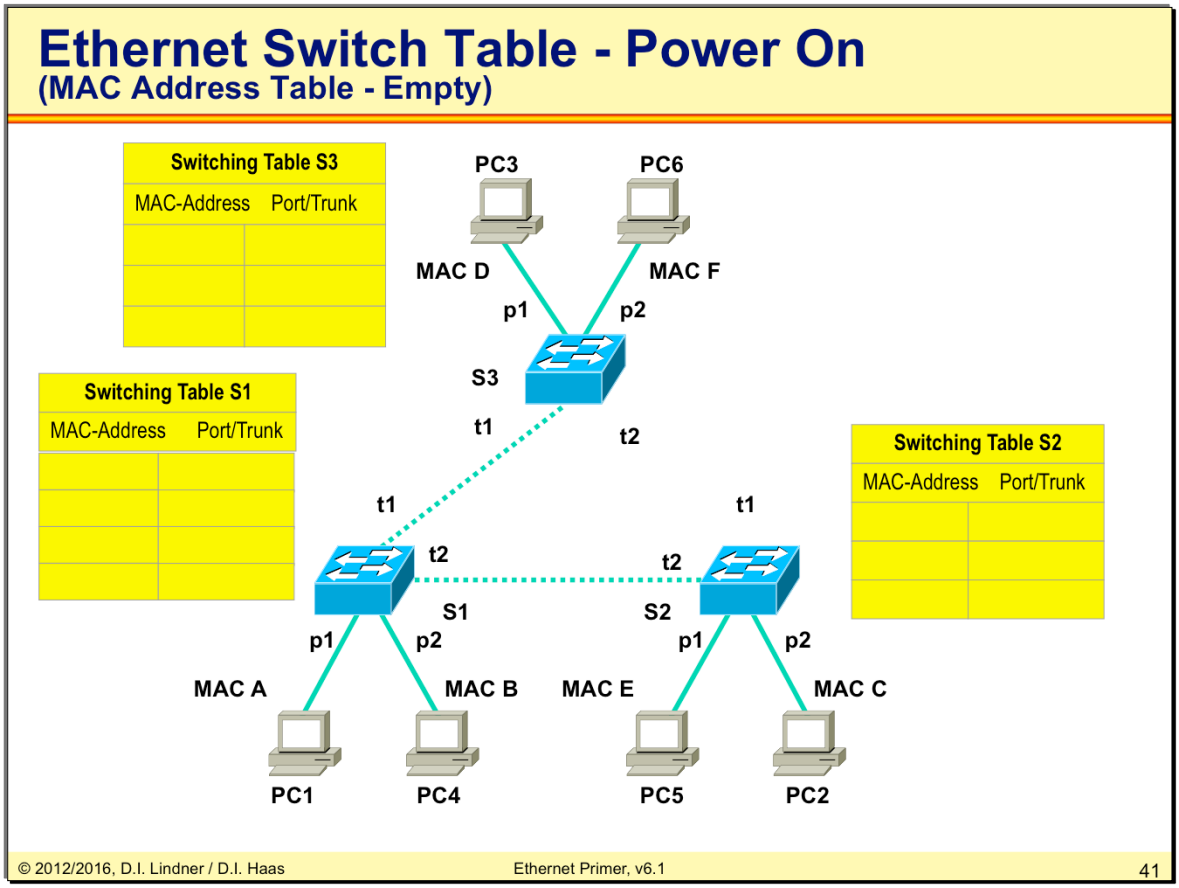
 Ethernet Primer, v6.1 39

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

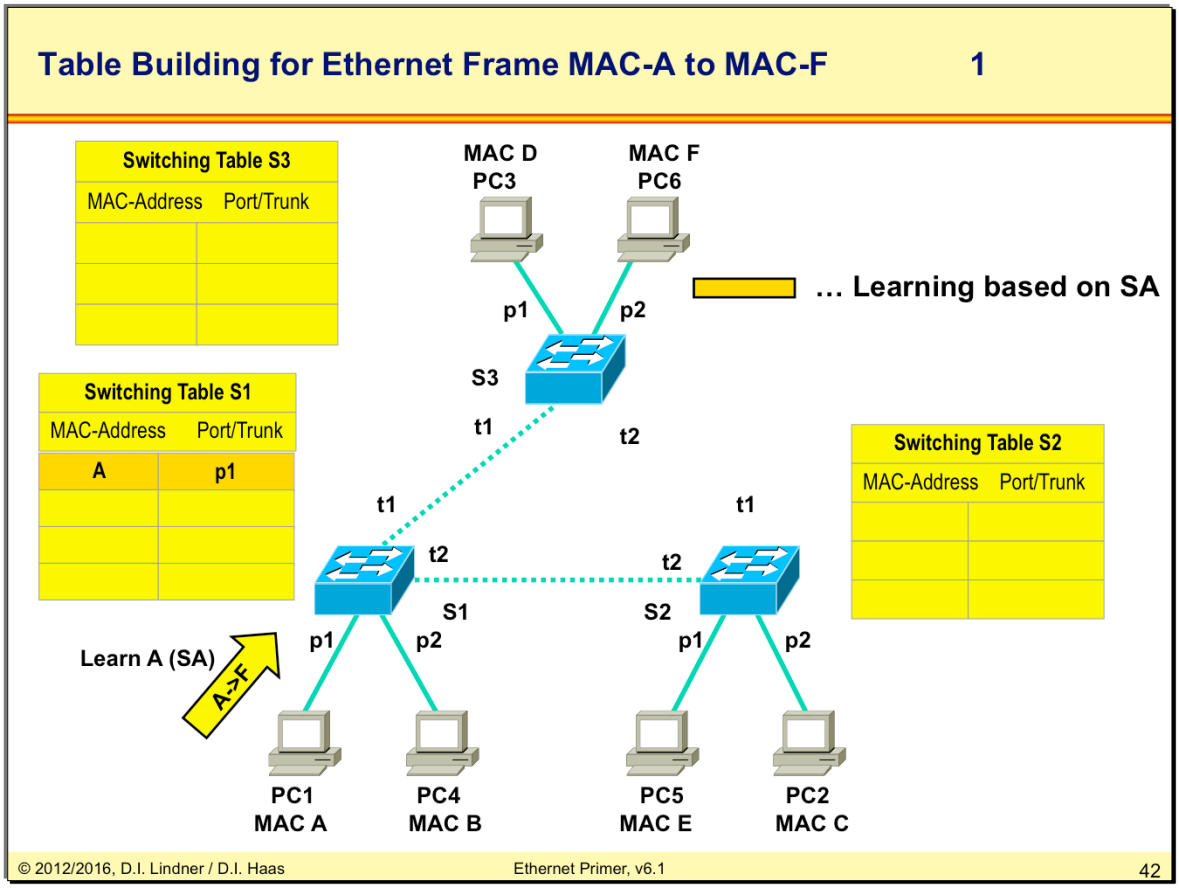


 Ethernet Primer, v6.1 40

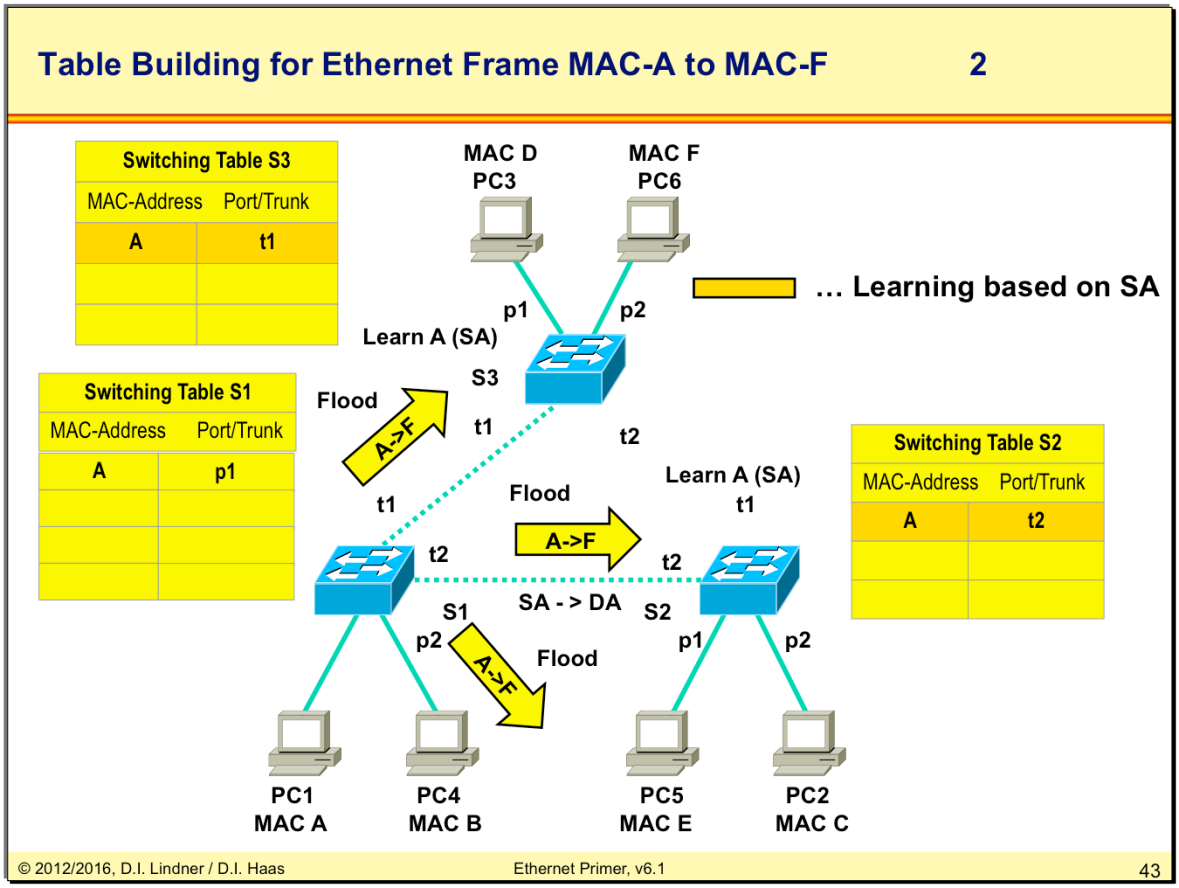
Ethernet Primer (v6.1)



Ethernet Primer (v6.1)

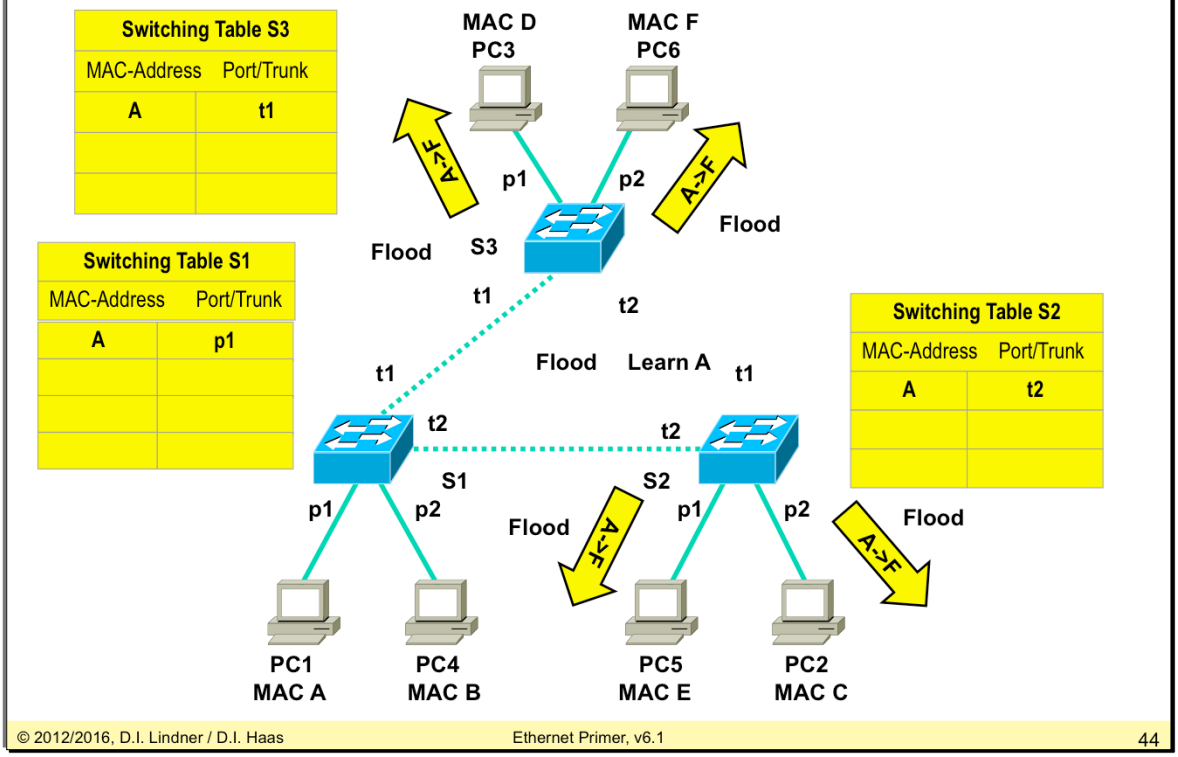


Ethernet Primer (v6.1)



Ethernet Primer (v6.1)

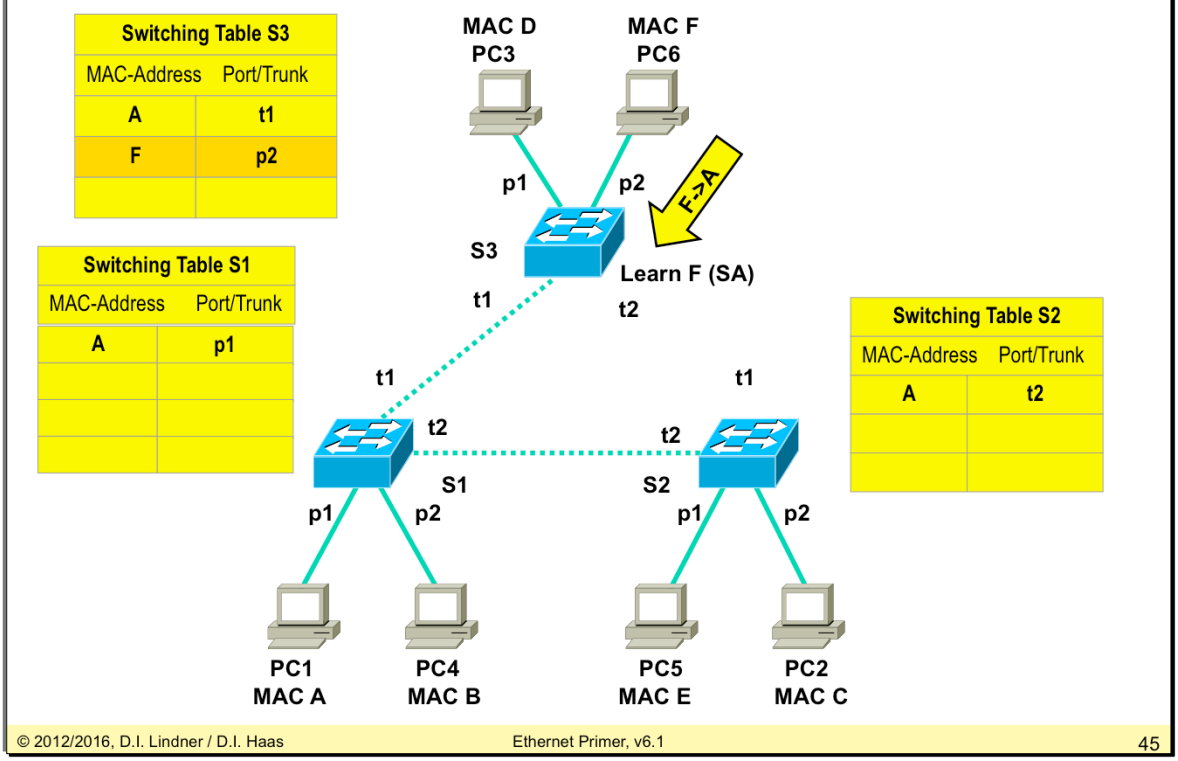
Table Building for Ethernet Frame MAC-A to MAC-F 3



 Ethernet Primer, v6.1 44

Ethernet Primer (v6.1)

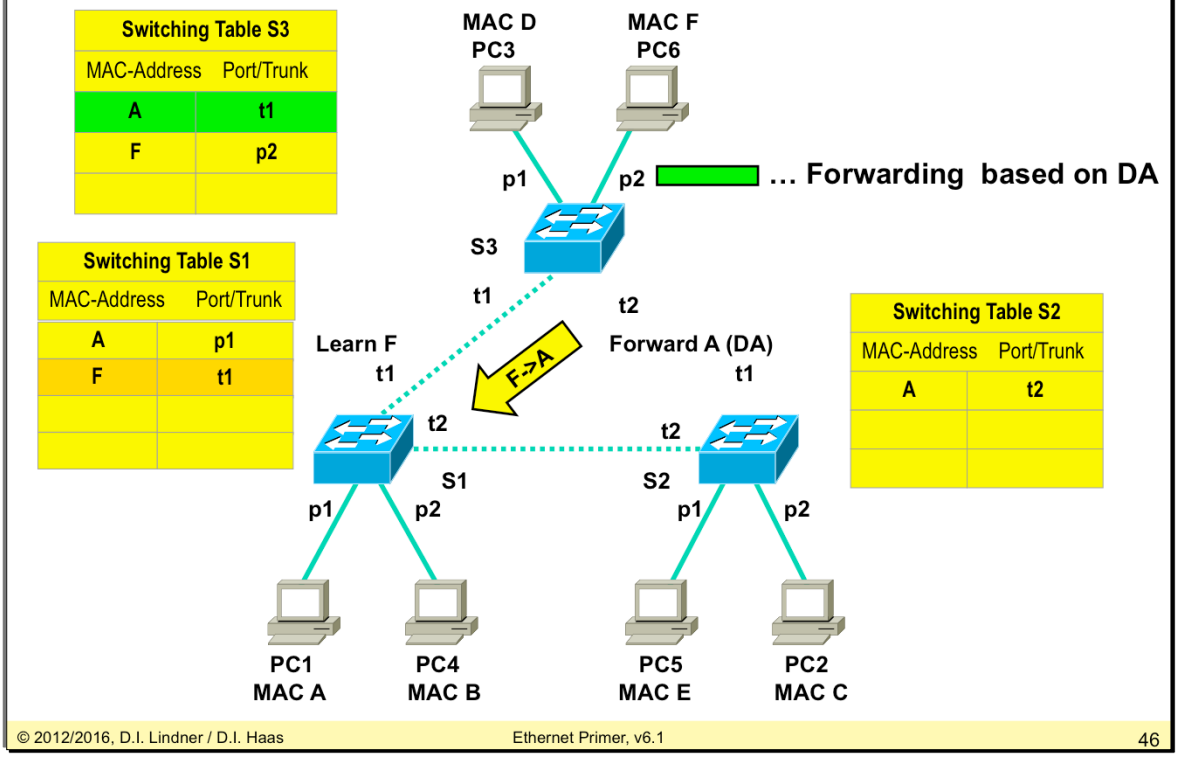
Table Building / Table Usage for Ethernet Frame MAC-F to MAC-A 1



 Ethernet Primer, v6.1 45

Ethernet Primer (v6.1)

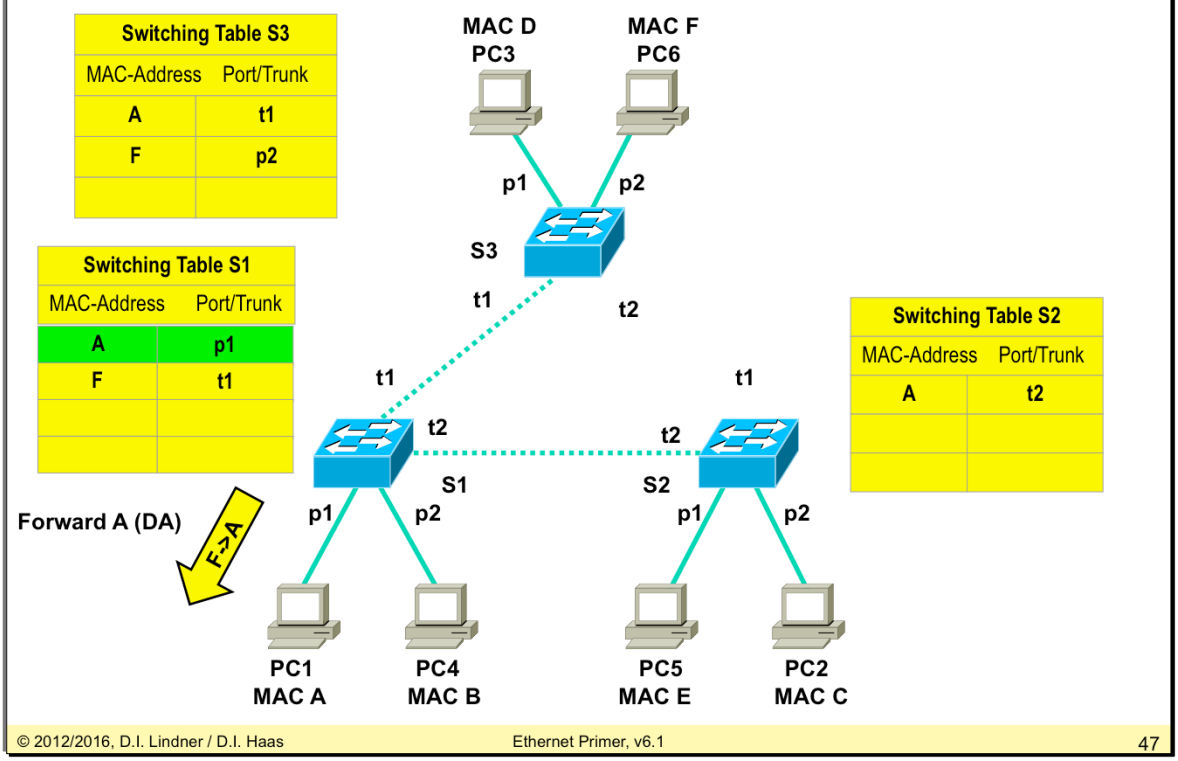
Table Building / Table Usage (Forwarding Decision) for Ethernet Frame MAC-F to MAC-A 2



 Ethernet Primer, v6.1 46

Ethernet Primer (v6.1)

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



 Ethernet Primer, v6.1 47

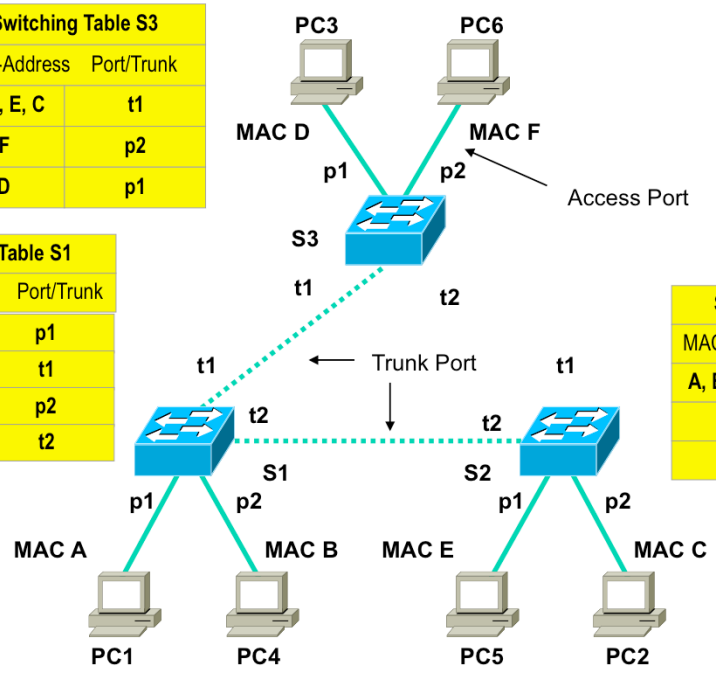
Ethernet Primer (v6.1)

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

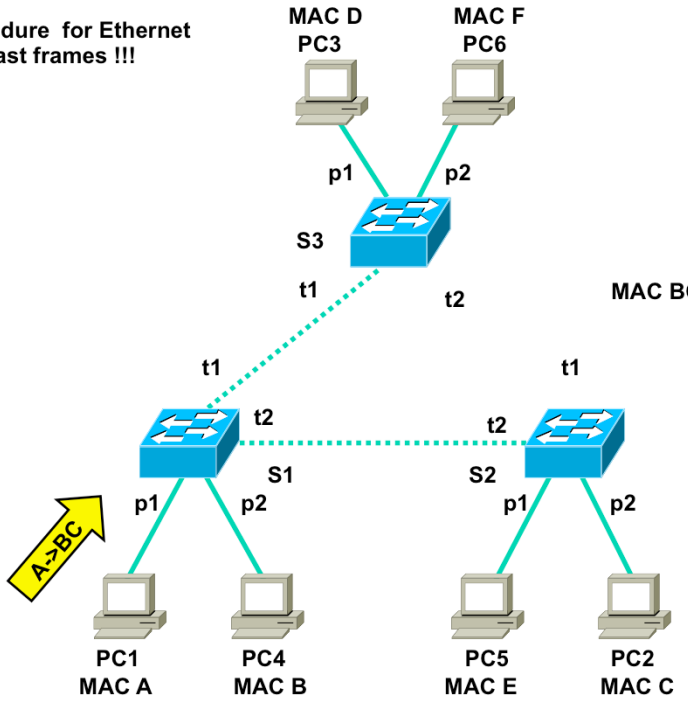


 Ethernet Primer, v6.1 48

Ethernet Primer (v6.1)

Ethernet Broadcast (BC) 1

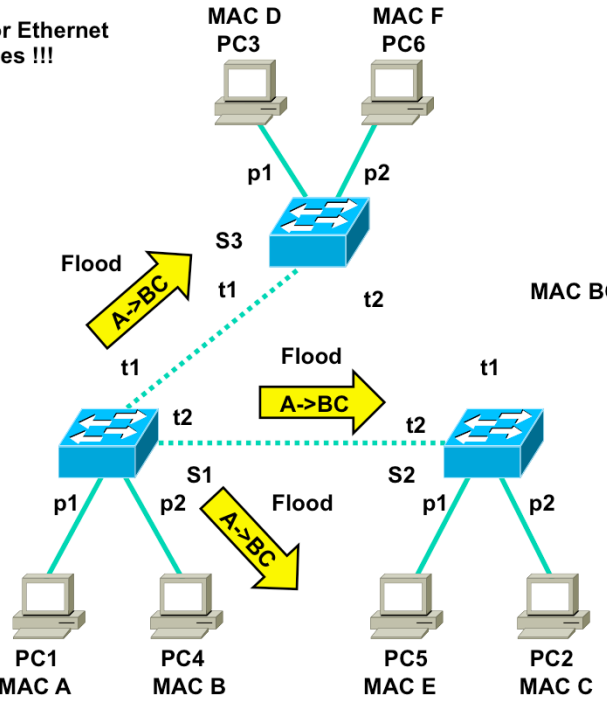
Same procedure for Ethernet multicast frames !!!



Ethernet Primer (v6.1)

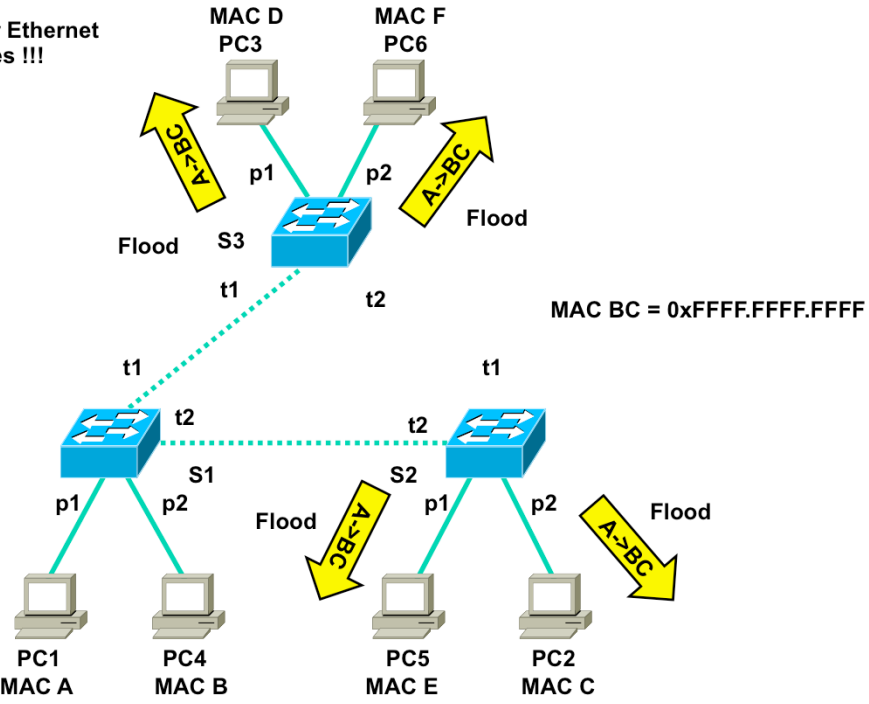
Ethernet Broadcast (BC) 2

Same procedure for Ethernet multicast frames !!!



Ethernet Broadcast (BC) 3

Same procedure for Ethernet multicast frames !!!



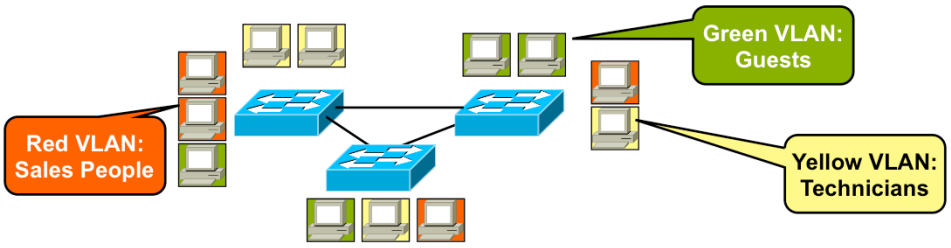
 Ethernet Primer, v6.1 51

Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed 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 Primer, v6.1 53

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 Primer, v6.1 54

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:

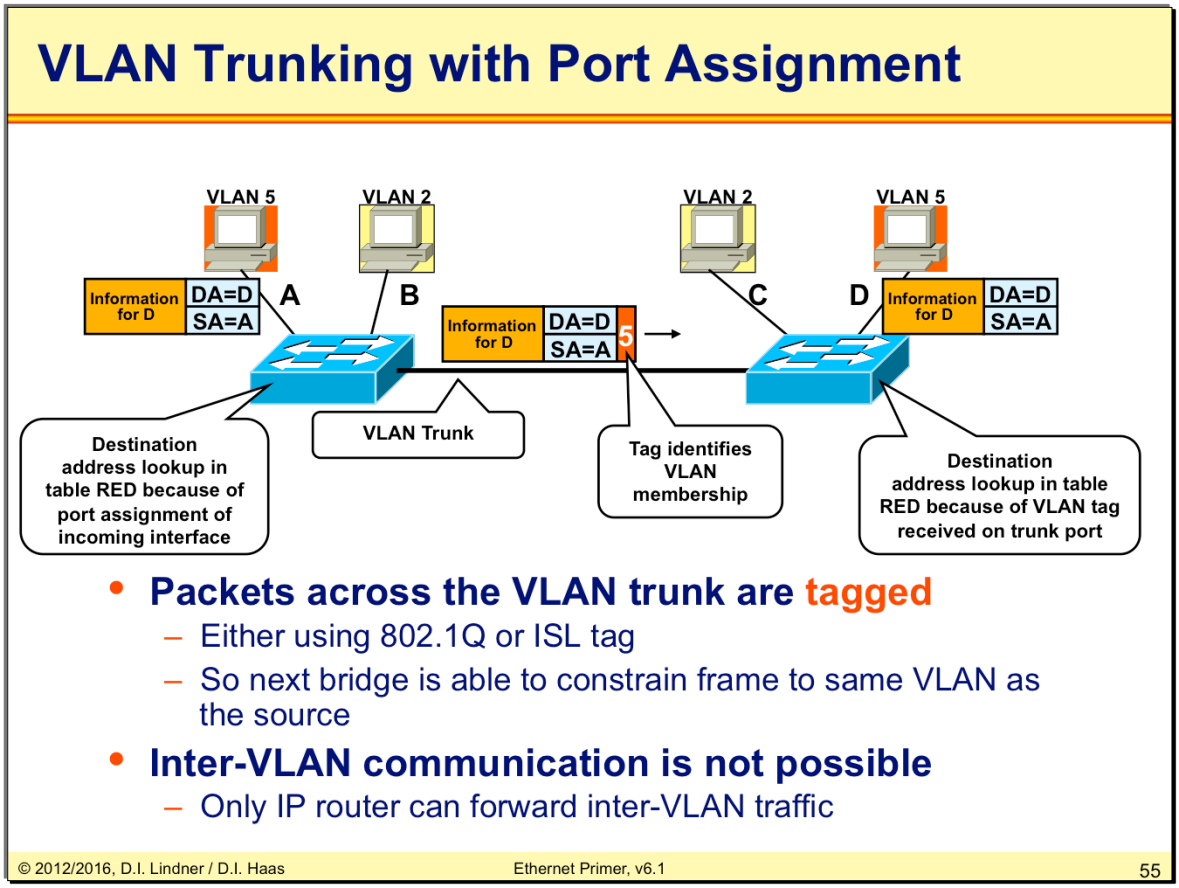
Port-based: fixed assignment port 4 -> VLAN x, most common approach, a station is member of one specific VLAN only, administrator has to reconfigure a switch in order to support a location change of a user.

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

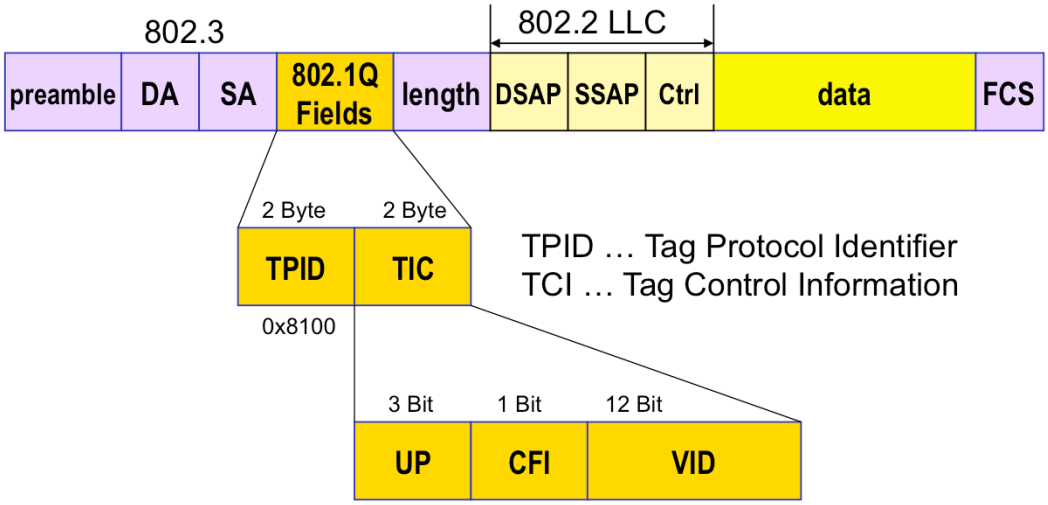
802.1X-based: User A -> VLAN engineering, User B -> VLAN finance, automatic location change support.

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.



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

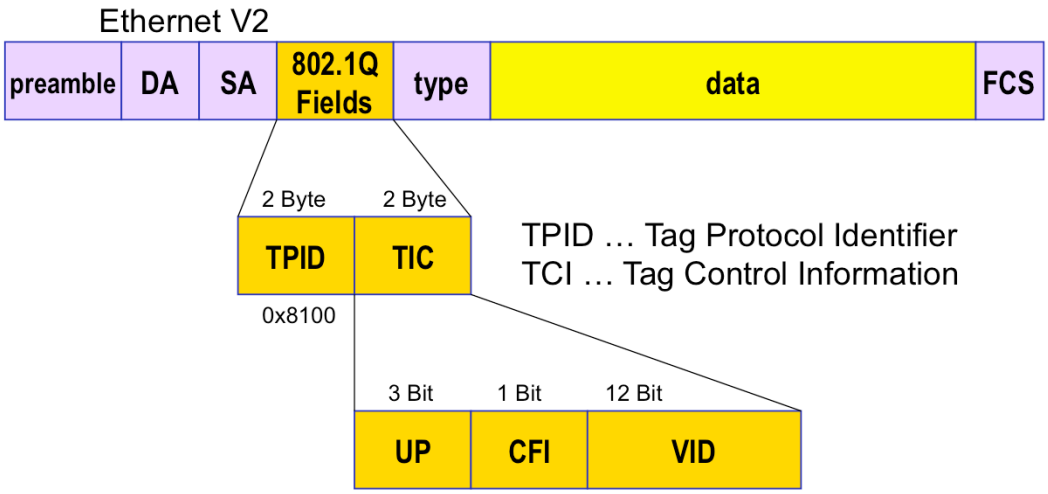
802.1Q VLAN Tagging – LLC (1)



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

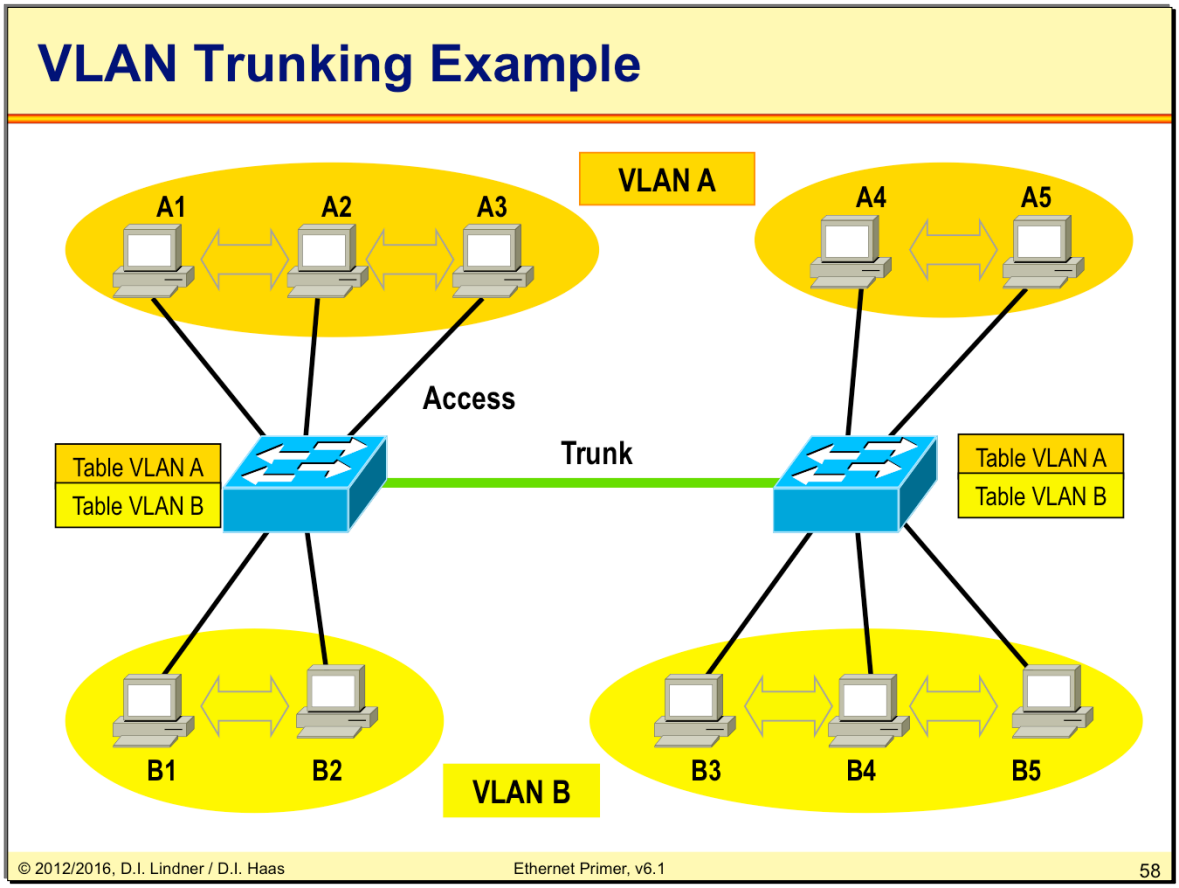
802.1Q VLAN Tagging – Ev2 (2)



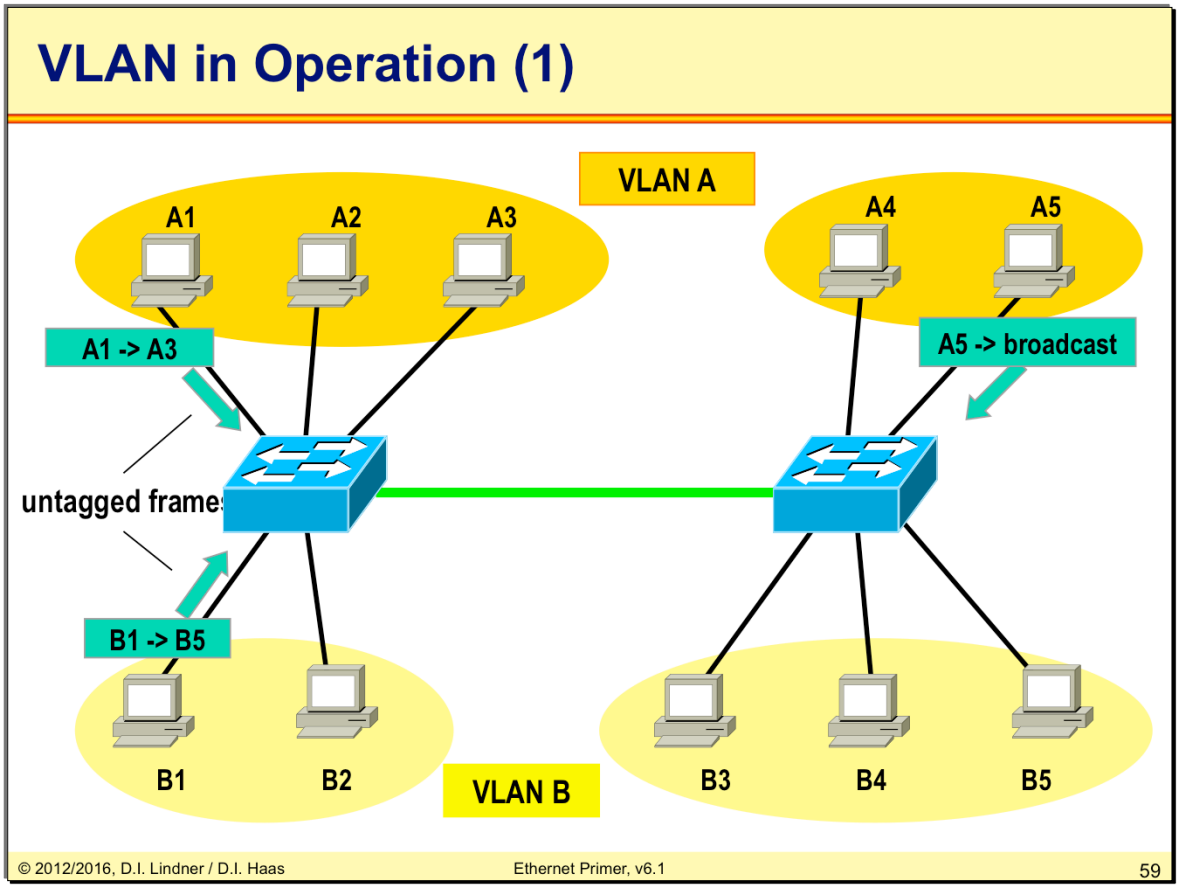
note: With tagging Ethernets maximal frame length = 1522, minimal frame length = 68

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

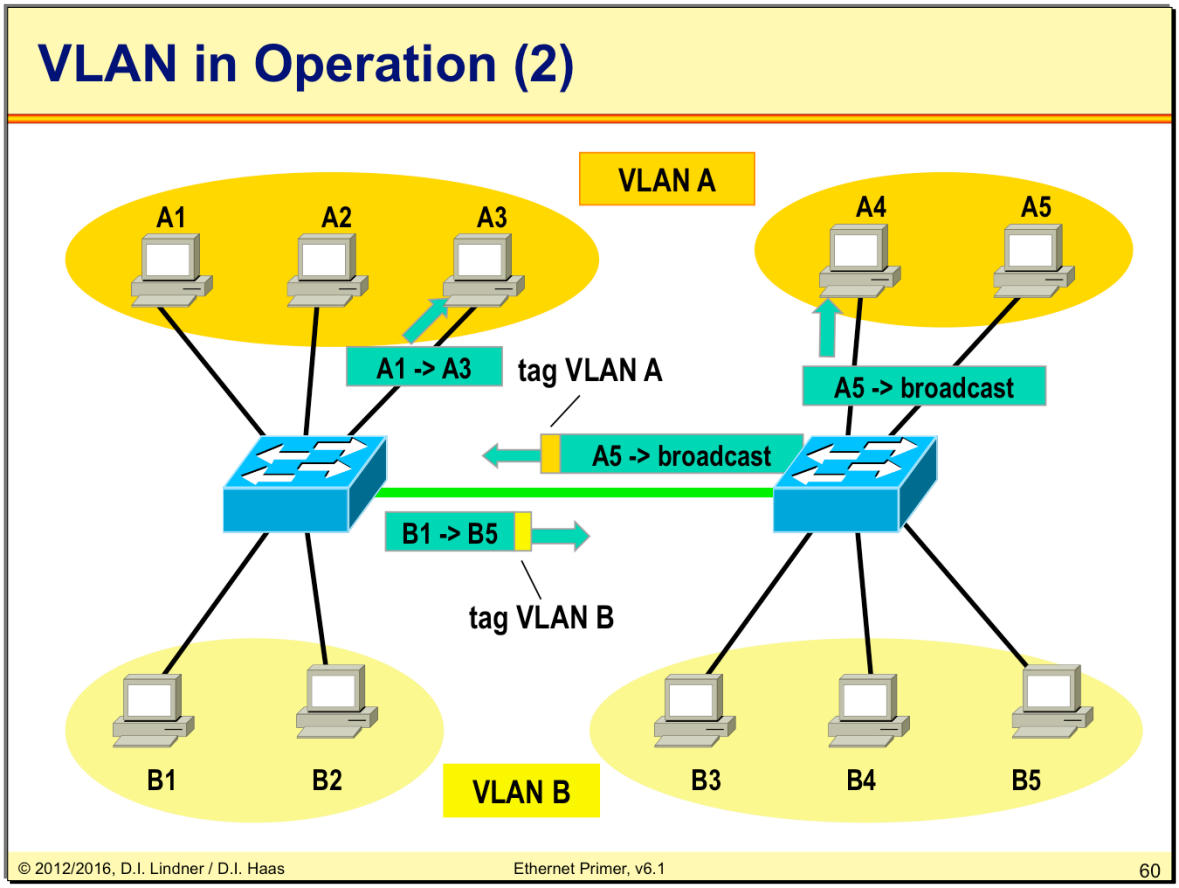
Ethernet Primer (v6.1)

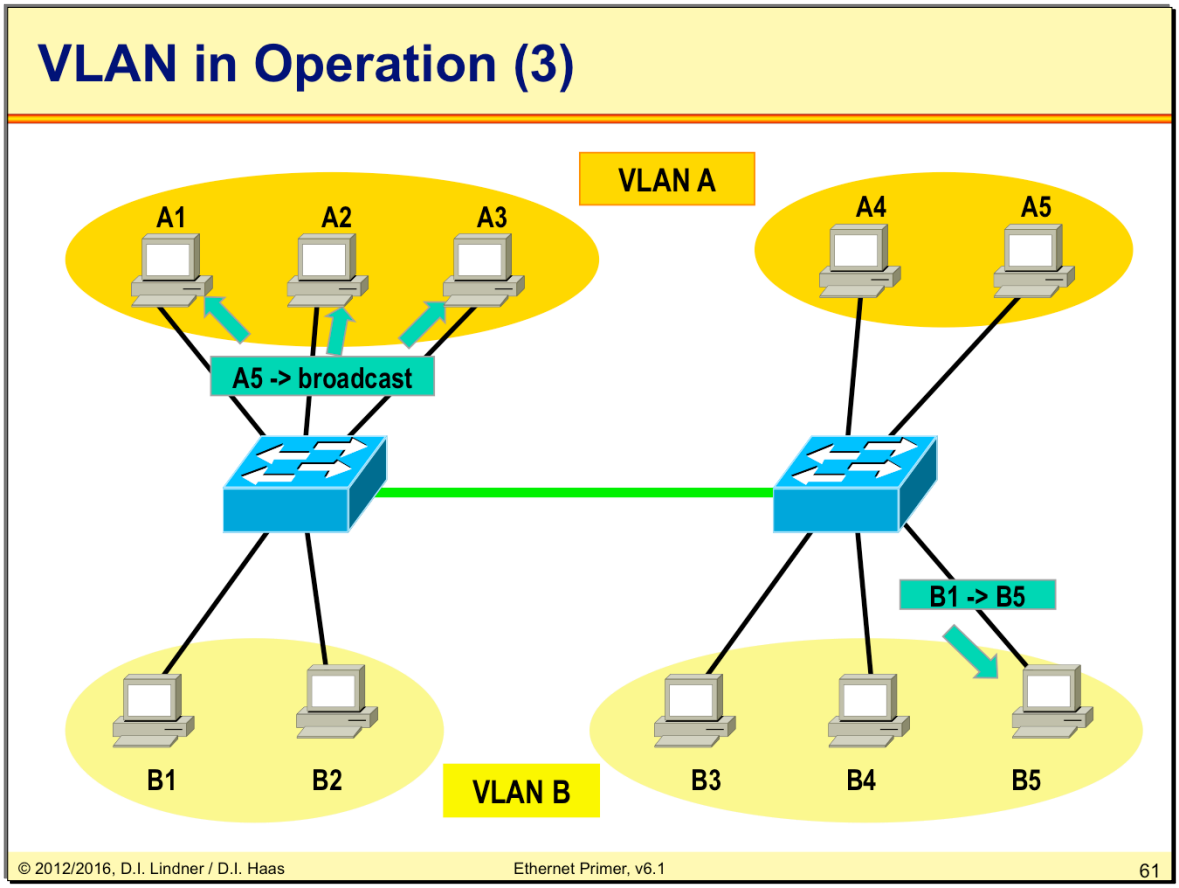


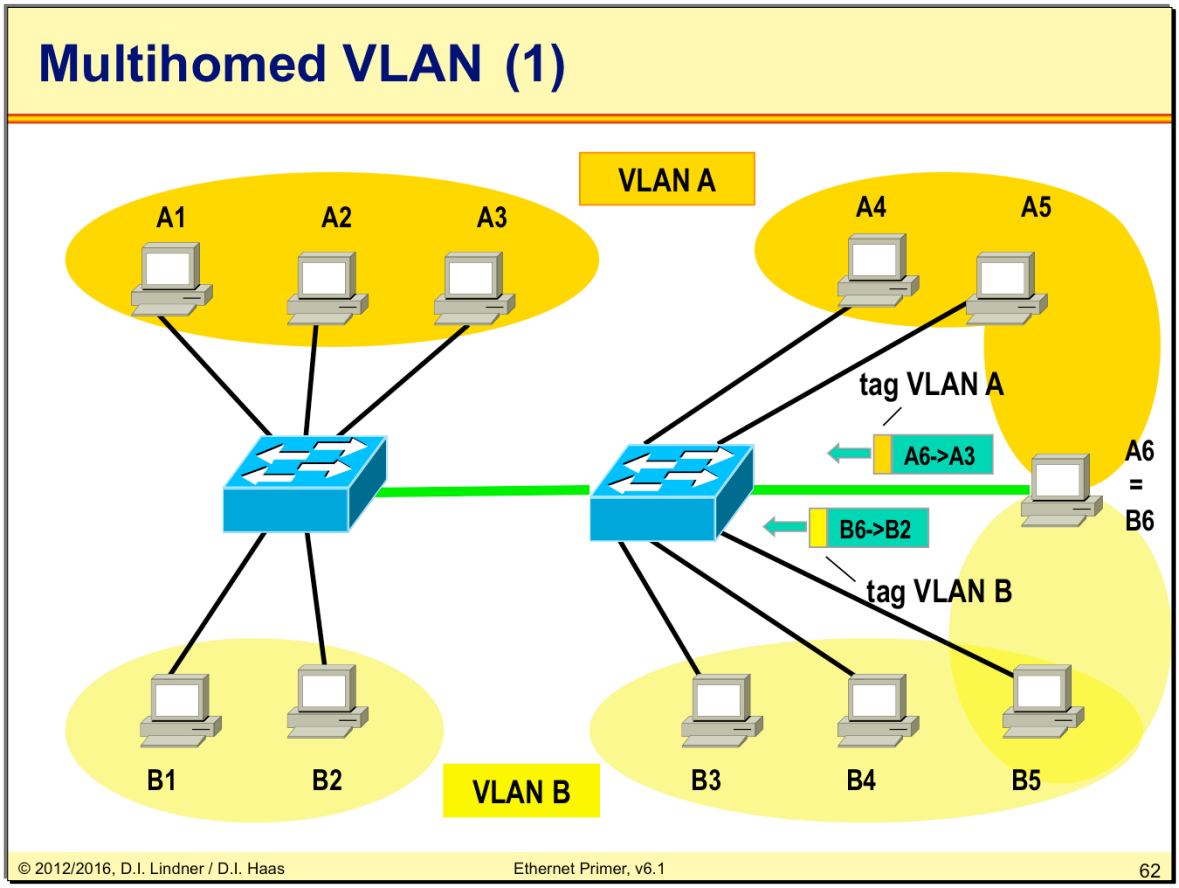
Ethernet Primer (v6.1)



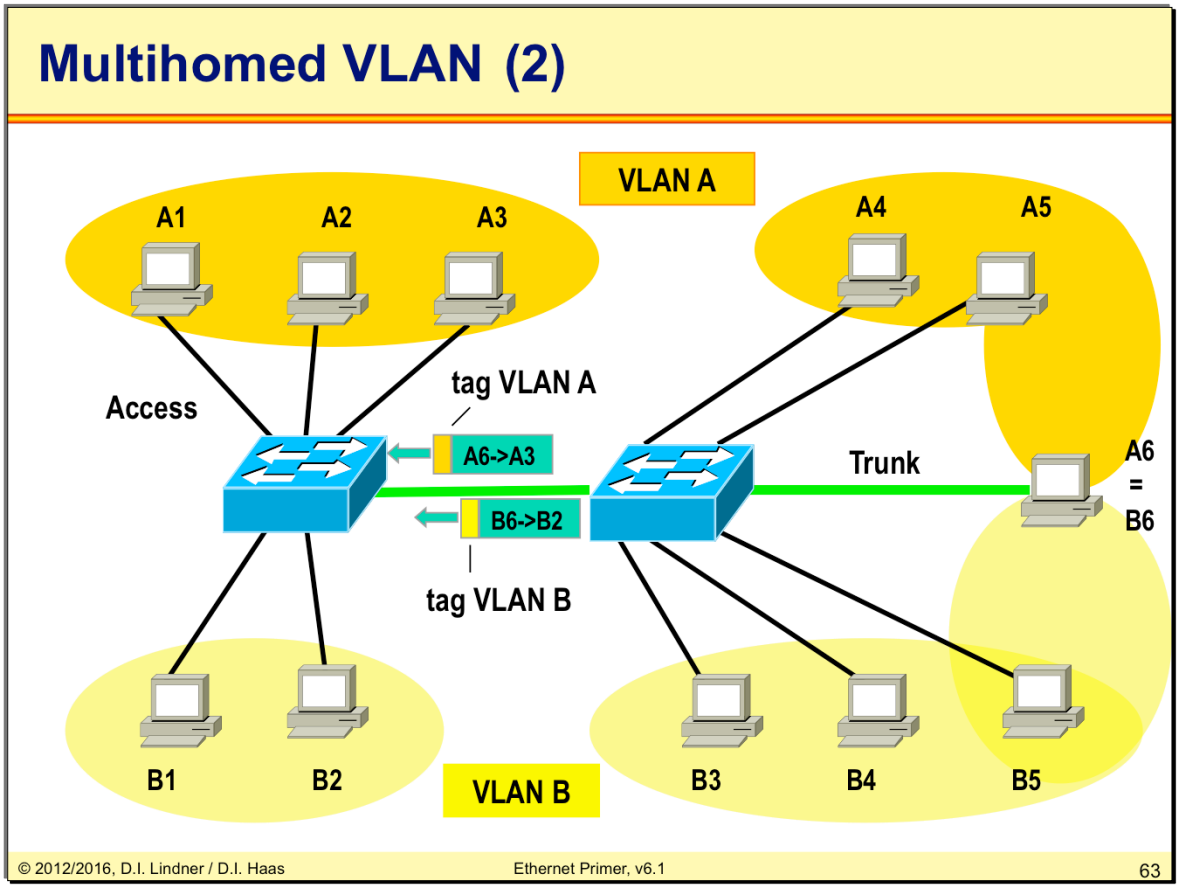
Ethernet Primer (v6.1)



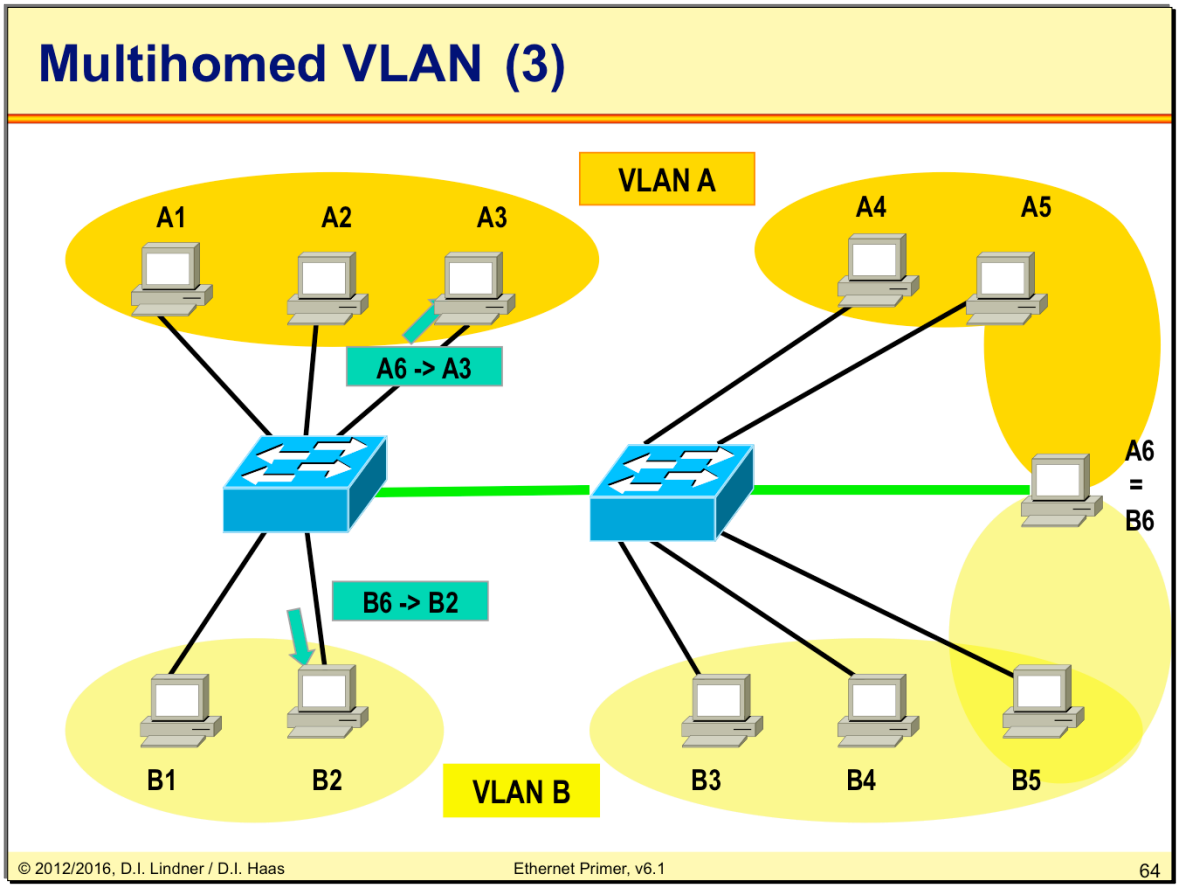




Ethernet Primer (v6.1)

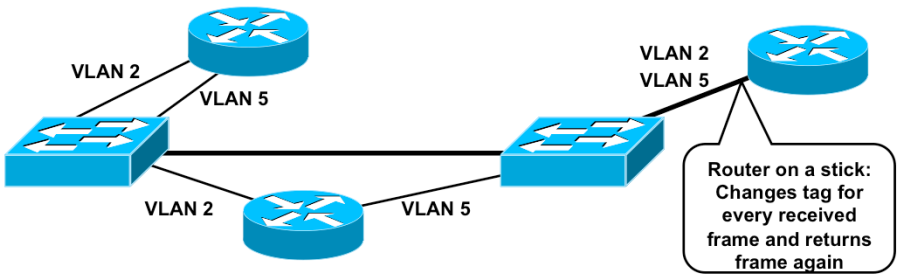


Ethernet Primer (v6.1)



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")

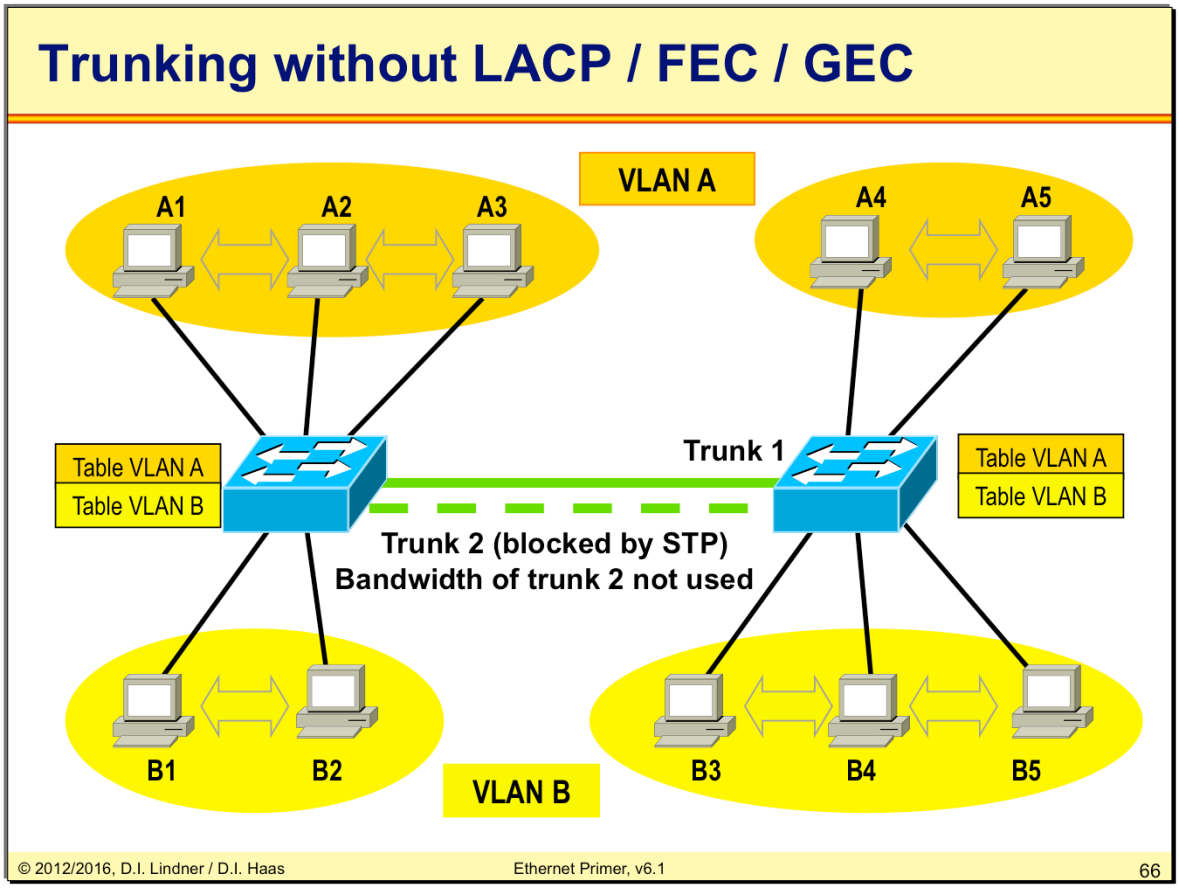


 Ethernet Primer, v6.1 65

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.



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.

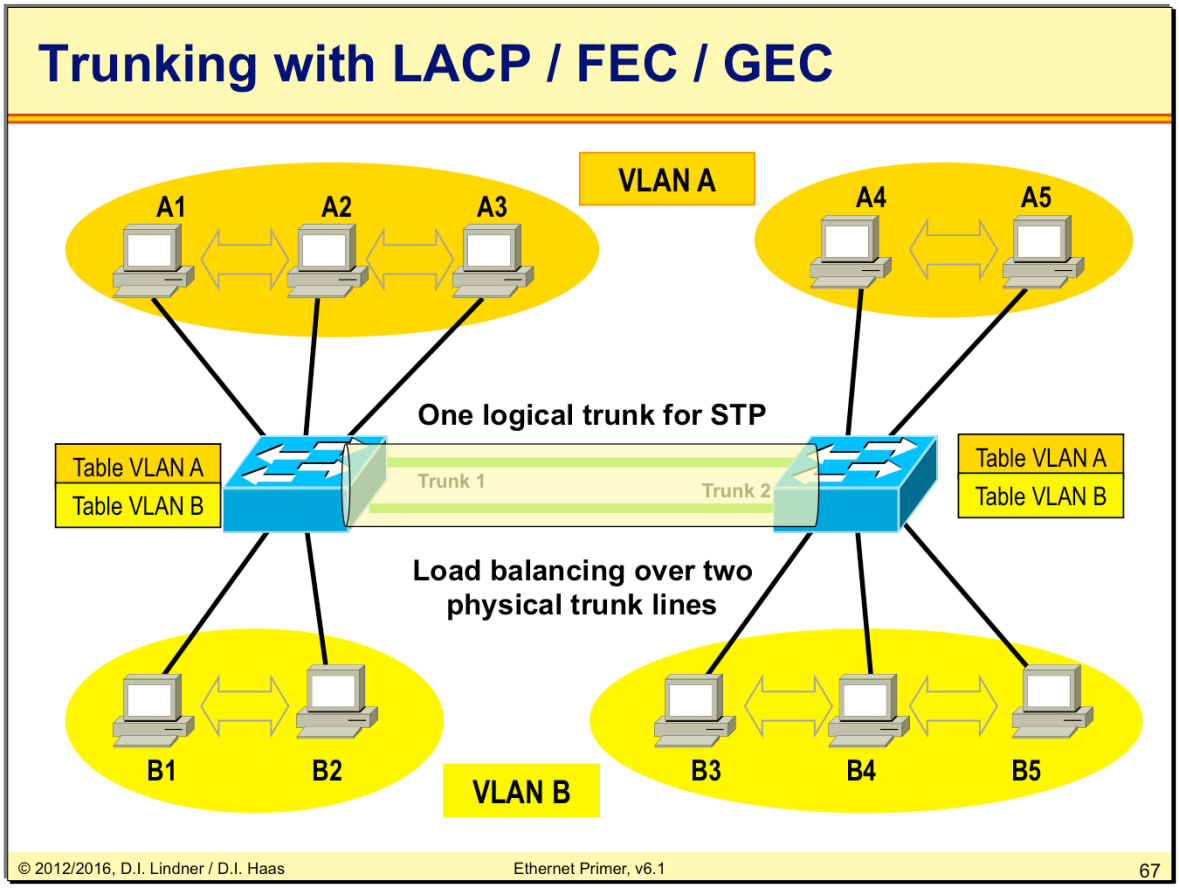
Bundling (aggregation) of physical links to one logical link – which is seen by STP - can be done with:

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
VLAN A
A1
A2
A3
A4
A5
One logical trunk for STP
Table VLAN A
Table VLAN B
Trunk 1
Trunk 2
Table VLAN A
Table VLAN B
Load balancing over two
physical trunk lines
B1
B2
B3
B4
B5
VLAN B
© 2012/2016, D.I. Lindner / D.I. Haas
Ethernet Primer, v6.1
67

Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
  - Old STP
  - Convergence
  - Rapid Spanning Tree Protocol (RSTP)
- High Speed Ethernet

Problem Description

- **We want redundant links in bridged networks**
- **But transparent bridging cannot deal with redundancy**
  - Broadcast storms and other problems
- **Solution: STP (Spanning Tree Protocol)**
  - Allows for redundant paths
  - Ensures non-redundant active paths
- **Invented by *Radia Perlman* as general "mesh-to-tree" algorithm**
- **Only one purpose:**
  cut off redundant paths with highest costs

Ethernet Primer (v6.1)

Algorhyme

I think that I shall never see
a graph more lovely than a tree
a graph whose crucial property
is loop-free connectivity.
A tree which must be sure to span
so packets can reach every lan.
first the root must be selected
by ID it is elected.
least cost paths to root are traced,
and in the tree these paths are place.
mesh is made by folks like me;
bridges find a spanning tree.

Radia Perlman

 Ethernet Primer, v6.1 70

Radia Perlman, PhD computer science 1988, MIT * MS math 1976, MIT * BA math 1973, MIT

Radia Perlman specializes in network and security protocols. She is the inventor of the spanning tree algorithm used by bridges, and the mechanisms that make modern link state protocols efficient and robust. She is the author of two textbooks, and has a PhD from MIT in computer science.

Her thesis on routing in the presence of malicious failures remains the most important work in routing security. She has made contributions in diverse areas such as, in network security, credentials download, strong password protocols, analysis and redesign of IPSec IKE protocols, PKI models, efficient certificate revocation, and distributed authorization. In routing, her contributions include making link state protocols robust and scalable, simplifying the IP multicast model, and routing with policies.

Spanning Tree Protocol

- **Takes care that there is always exact only one active path between any 2 stations**
- **Implemented by a special communication protocol between the bridges**
  - Using BPDU (Bridge Protocol Data Unit) frames with MAC-multicast address as destination address
- **Three important STP parameters determine the resulting tree topology in a meshed network:**
  - Bridge-ID
  - Interface-Cost
  - Port-ID

 Ethernet Primer, v6.1 71

What do we need for STP to work? First of all this protocol needs a special messaging means, realized in so-called **Bridge Protocol Data Units (BPDUs).** BPDUs are simple messages contained in Ethernet frames containing several parameters described in the next pages.

Parameters for STP 1

- **Bridge Identifier (Bridge ID)**
  - Consists of a priority number and the MAC-address of a bridge
    - Bridge-ID = Priority# (2 Byte) + MAC# (6 Byte)
  - Priority number may be configured by the network administrator
    - Default value is 32768
  - Lowest Bridge ID has highest priority
  - If you keep default values
    - The bridge with the lowest MAC address will have the highest priority

 Ethernet Primer, v6.1 72

Each bridge is assigned one unique **Bridge-ID** which is a combination of a 16 bit priority number and the lowest MAC address found on any port on this bridge. The Bridge-ID is determined automatically using the default priority 32768. Note: Although bridge will not be seen by end systems, for bridge communication and management purposes a bridge will listen to one or more dedicated (BIA) MAC addresses. Typically, the lowest MAC-address is used for that. The Bridge-ID is used by STP algorithm to determine root bridge and as tie-breaker to when determine the designated port.

Parameters for STP 2

- **Port Cost (C)**
  - Costs in order to access local interface
  - Inverse proportional to the transmission rate
  - Default cost = 1000 / transmission rate in Mbit/s
    - With occurrence of 1Gbit/s Ethernet the rule was slightly adapted
    - May be configured to a different value by the network administrator
- **Port Identifier (Port ID)**
  - Consists of a priority number and the port number
    - Port-ID = port priority#.port#
    - Default value for port priority is 128
    - Port priority may be configured to a different value by the network administrator

 Ethernet Primer, v6.1 73

Each port is assigned a **Port Cost**. Again this value is determined automatically using the simple formula Port Cost = 1000 / BW, where BW is the bandwidth in Mbit/s. Of course the Port Cost can be configured manually. Port Cost are used by STP algorithm to calculate Root Path Cost in order to determine the root port and the designated port

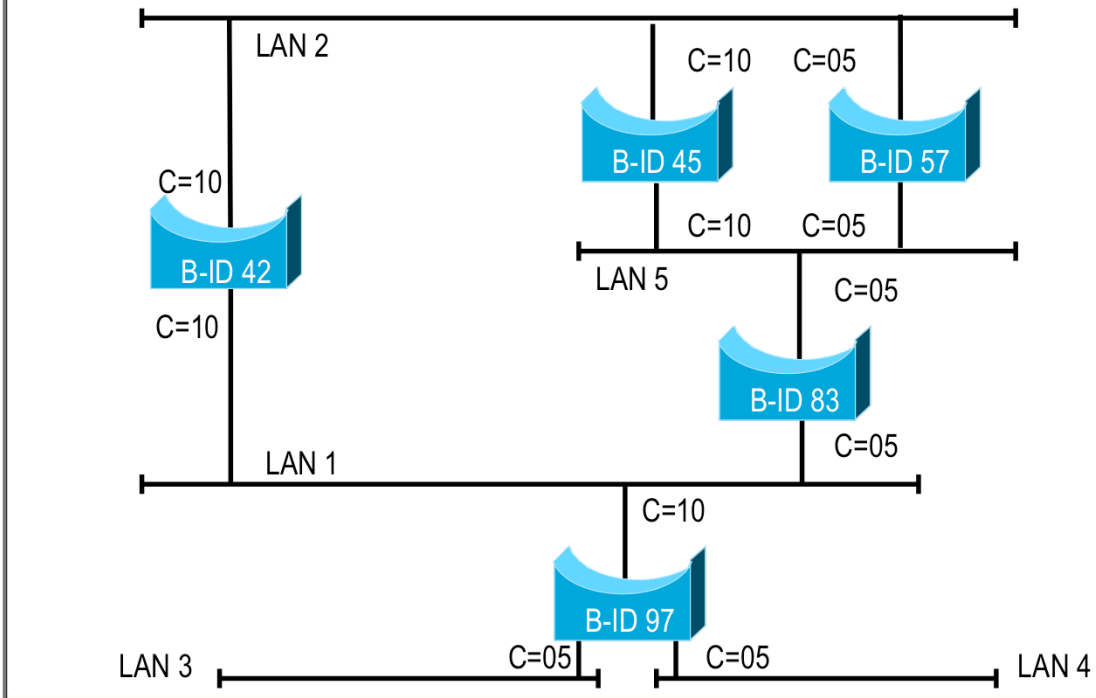
Each port is assigned a **Port Identifier.** Only used by STP algorithm as tie-breaker if the same Bridge-ID and the same Path Cost is received on multiple ports.

Comparison Table For Port Costs:

Speed [Mbit/s]	OriginalCost (1000/Speed)	802.1D-1998	802.1D-2004
10	100	100	2000000
100	10	19	200000
155	6	14	(129032 ?)
622	1	6	(32154 ?)
1000	1	4	20000
10000	1	2	2000

- **Also different cost values might be used**
  - See recommendations in the IEEE 802.1D-2004 standard to comply with RSTP and MSTP
  - 802.1D-2004 operates with 32-bit cost values instead of 16-bit

STP Parameter Example (1)

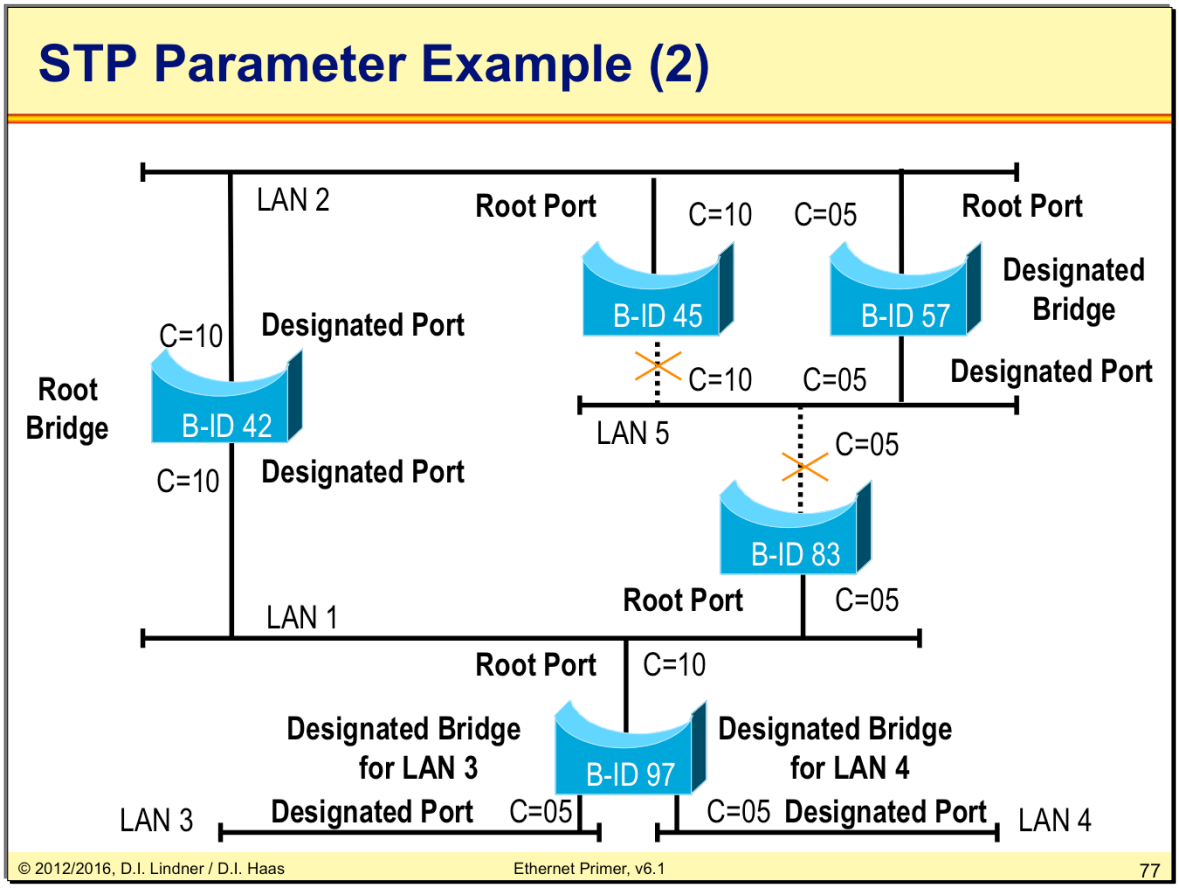


Spanning Tree Algorithm Summary

- **Select the root bridge**
  - Bridge with the lowest Bridge Identifier
- **Select the root ports**
  - By computation of the shortest path from any non-root bridge to the root bridge
  - Root port points to the shortest path towards the root
- **Select one designated bridge for every LAN segment which can be reached by more than one bridge**
  - Bridge with lowest root path costs on the root port side
  - Corresponding port on other side is called designated port
- **Set the designated and root ports in forwarding state**
- **Set all other ports in blocking state**

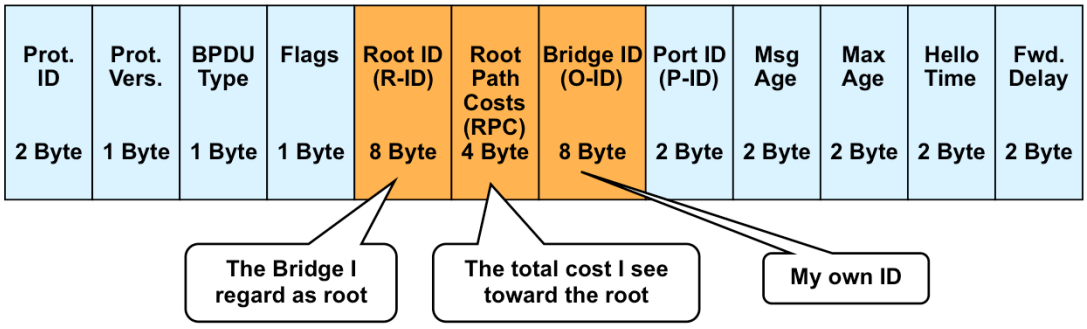
 Ethernet Primer, v6.1 76

These creates single paths from the root to all leaves (LAN segments) of the network.


STP Parameter Example (2)
LAN 2
Root Port
C=10
C=05
Root Port
B-ID 45
B-ID 57
Designated Bridge
C=10
Designated Port
Root Bridge
B-ID 42
C=10
C=05
Designated Port
LAN 5
C=05
B-ID 83
Root Port
C=05
LAN 1
Root Port
C=10
Designated Bridge for LAN 3
B-ID 97
Designated Bridge for LAN 4
LAN 3
Designated Port
C=05
C=05
Designated Port
LAN 4
© 2012/2016, D.I. Lindner / D.I. Haas
Ethernet Primer, v6.1
77

BPDU Format

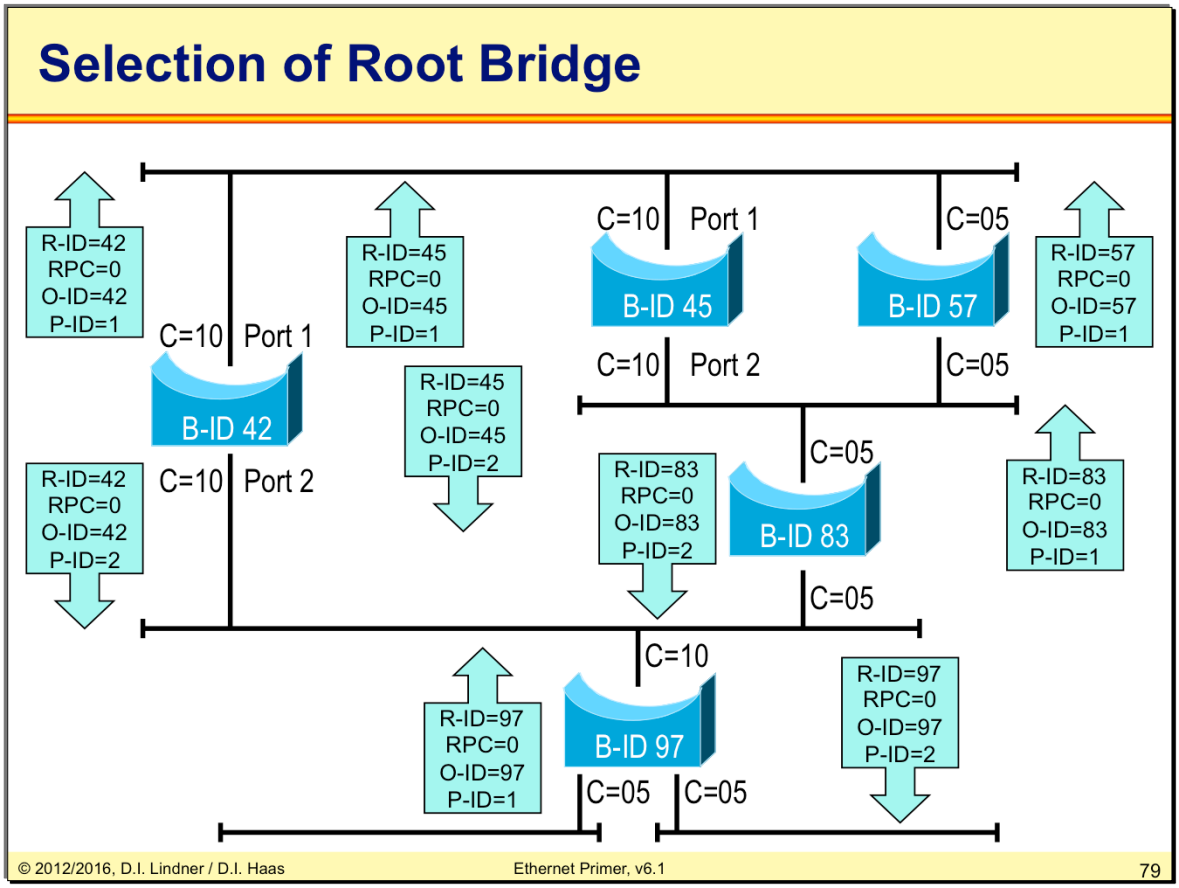
- **Each bridge sends periodically BPDUs carried in Ethernet multicast frames**
  - Hello time default: 2 seconds
- **Contains all information necessary for building Spanning Tree**



 Ethernet Primer, v6.1 78

Just for your interest, the above picture shows the structure of BPDUs. You see, there is no magic in here, and the protocol is very simple. There are no complicated protocol procedures. BPDUs are sent periodically and contain all involved parameters. Each bridge enters its own "opinion" there or adds its root path costs to the appropriate field. Note that some parameters are transient and others are not.

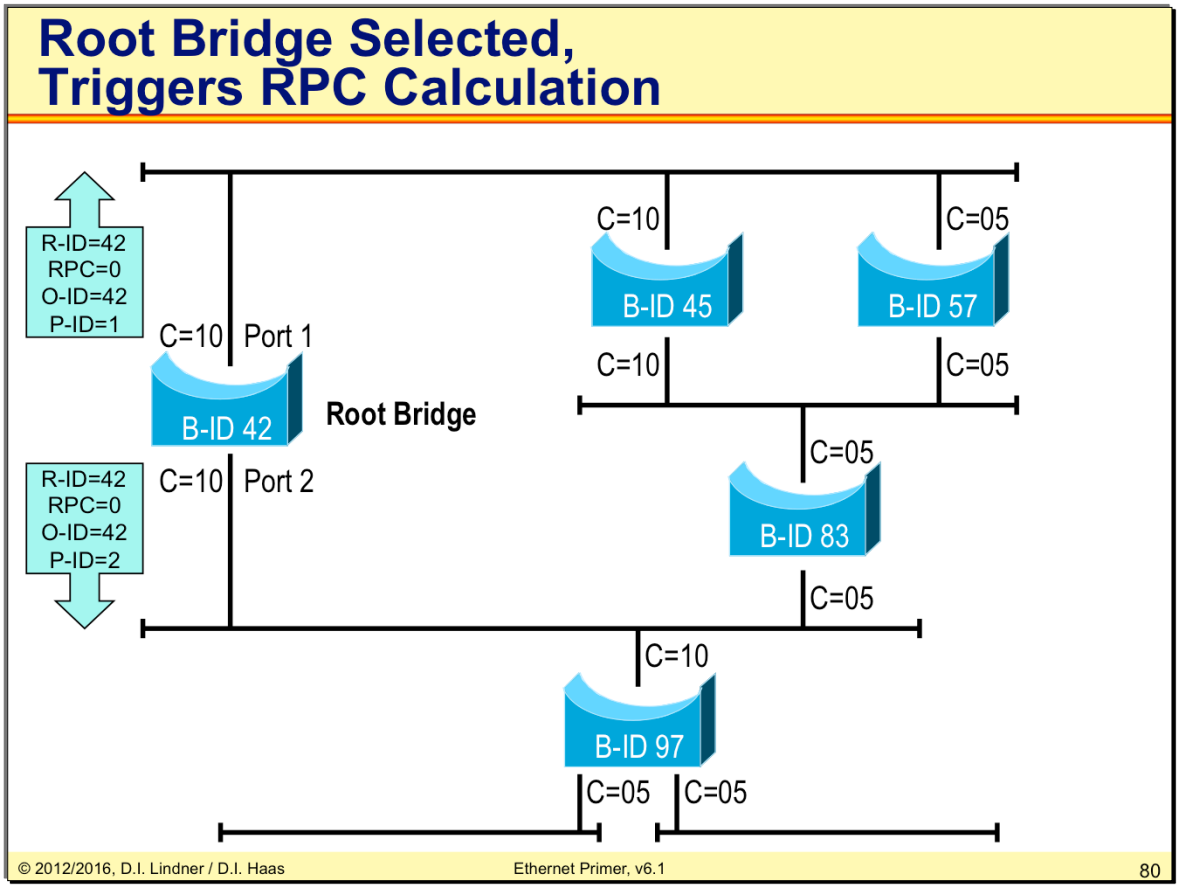
The other parameters will be explained in the next slides.



After power up all ports are set in a Blocking State and every bridge tries to become the Root Bridge (RB) of the Spanning Tree by sending Configuration BPDUs.

Blocking state means: End station Ethernet frames are not received and forwarded on such a port but BDPU frames can still be received, manipulated by the bridge and transmitted on such port. BPDU frames are actually filtered based on the well-known multicast address and are given to the CPU of the bridge.

Using such Configuration BPDUs, a bridge tells, which bridge actually is seen as RB, which path costs exist to this RB (Root Path Cost) and its own Bridge ID and Port ID.

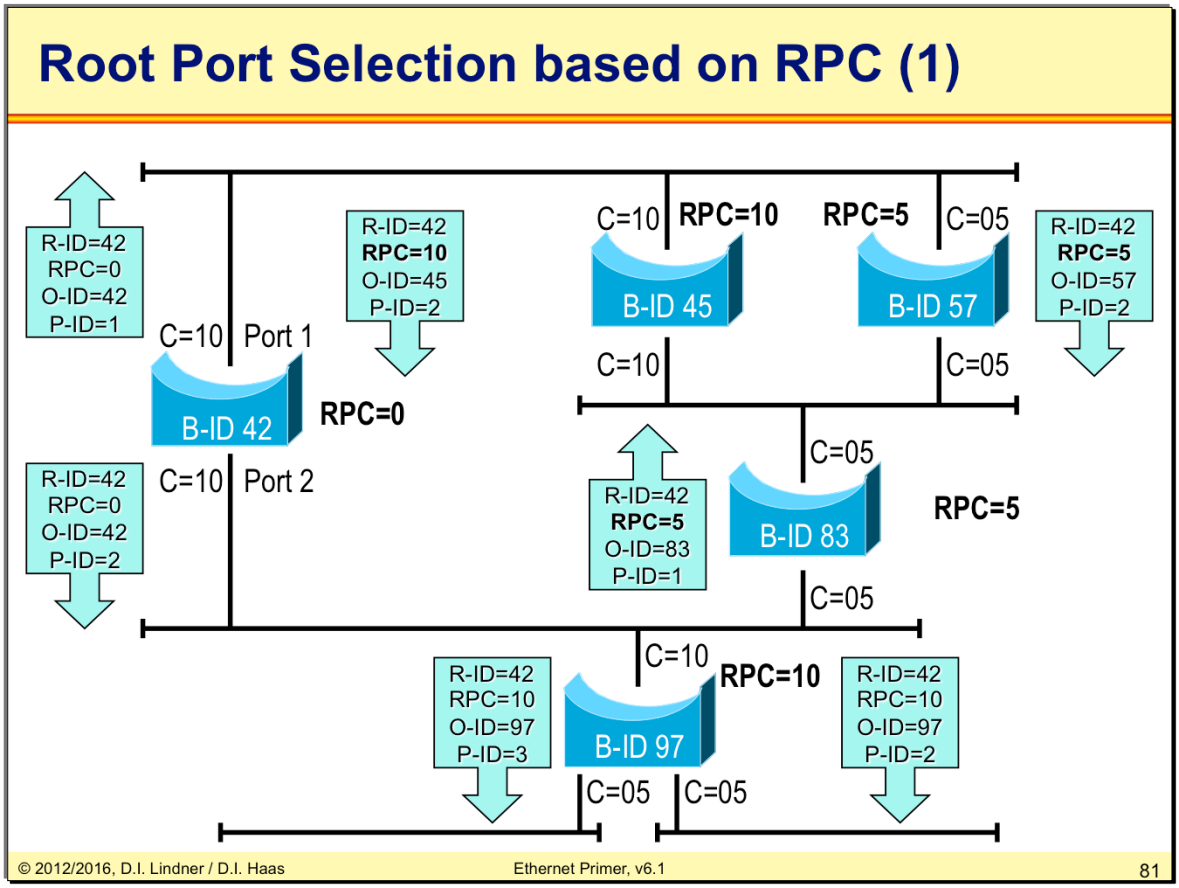


Bridge with the lowest Bridge ID becomes RB. after selection of the RB all sending of Configuration BPDUs are exclusively triggered by the RB. Other bridges just move such BPDUs on after actualizing the corresponding BPDU fileds.

Strategy to determinate the RB :

If bridge receives a Configuration BPDU with *lower* Root Bridge ID as own Bridge ID the bridge stops sending Configuration BPDUs on this port and the received and adapted Configuration BPDU is forwarded to all other ports.

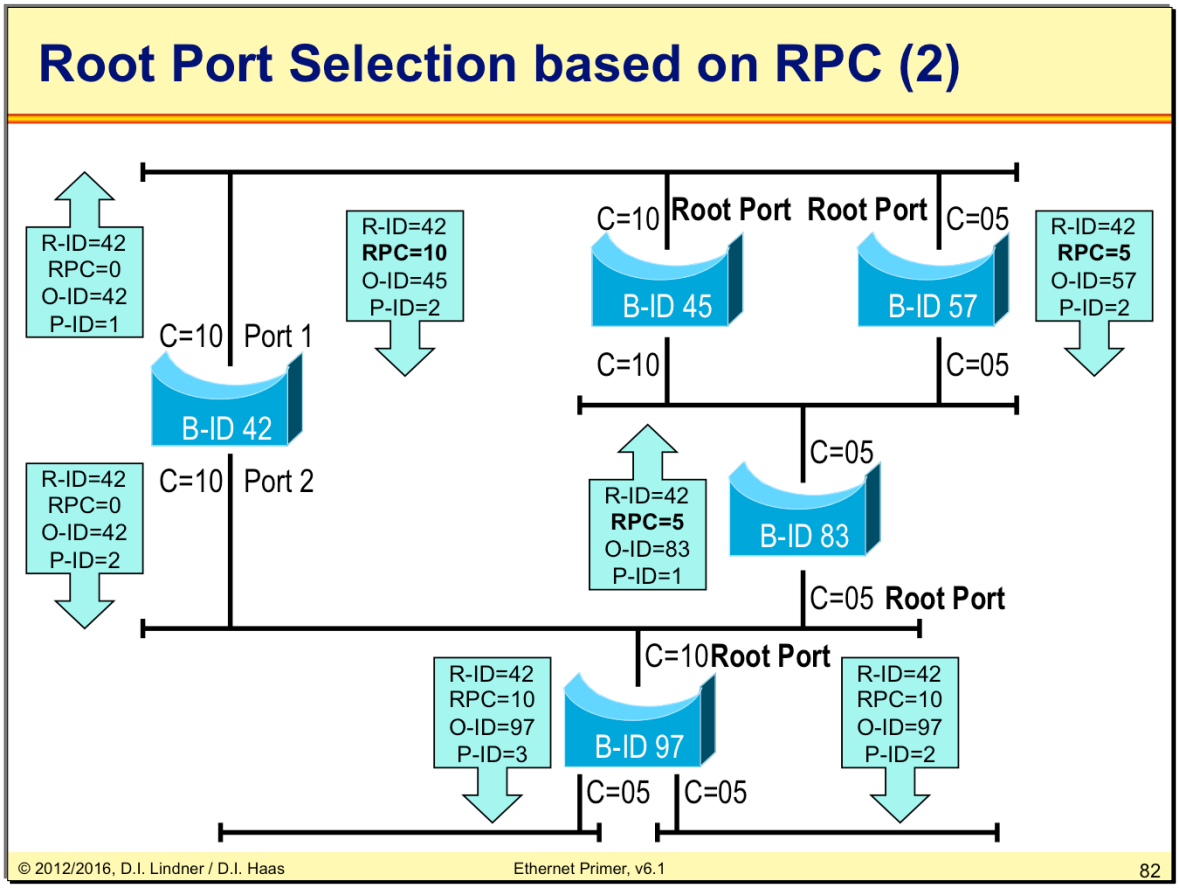
If bridge receives Configuration BPDU with *higher* Root Bridge ID as own Bridge ID the bridge continues sending Configuration BPDUs with own Bridge ID as proposed Root Bridge ID on all ports, the other bridges should give up.



Now, every bridge determines which of its ports has the lowest Root Path Cost. Root Path Cost = sum of all port costs from this bridge to the RB, including port costs of all intermediate bridges. This port becomes the Root Port. In case of equal costs the port ID decides (lower means better).

The principle calculation method: Root Path Cost received in BPDU + port cost of the local port receiving that BPDU.

Similar to Root Bridge selection, a Designated Bridge (DB) is selected for each LAN-segment which is the bridge with the lowest Root Path Cost on its Root Port. In case of equal costs the bridge with the lowest Bridge ID wins again.

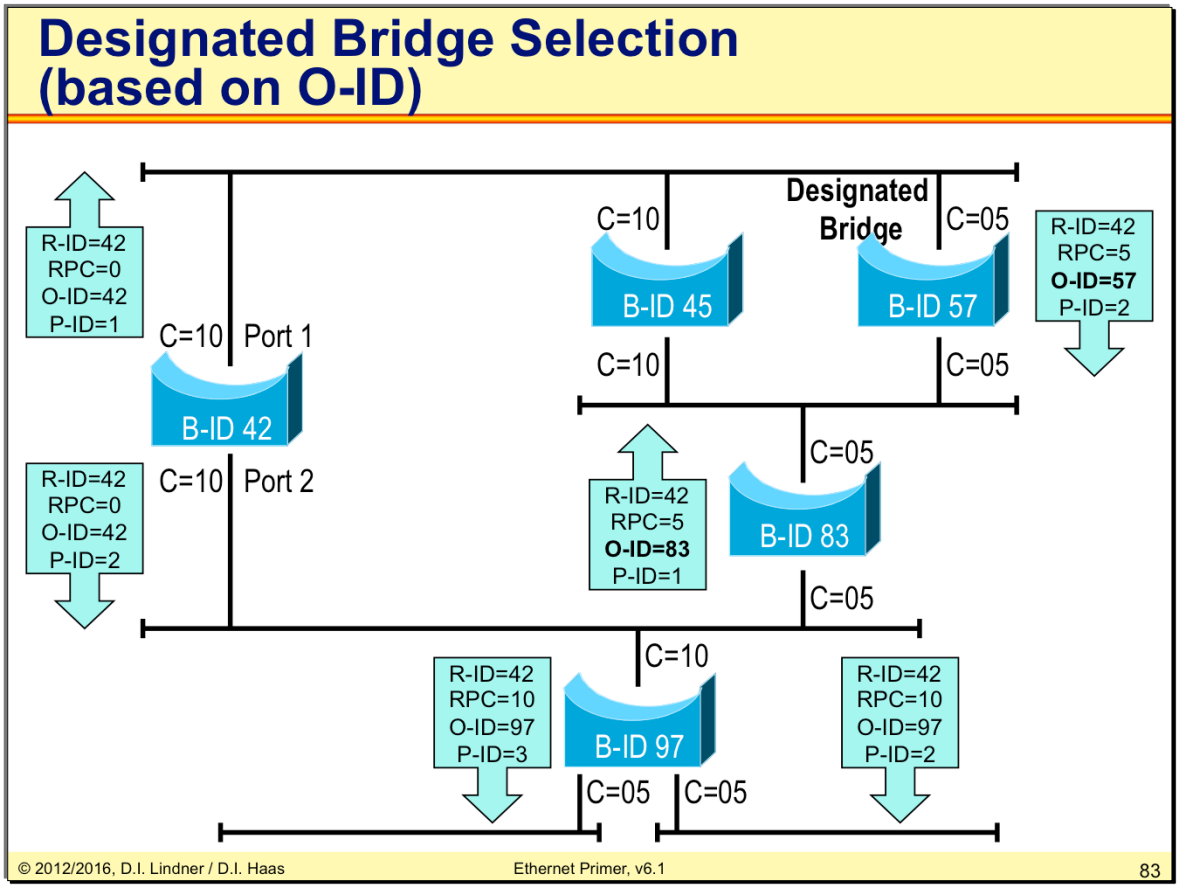


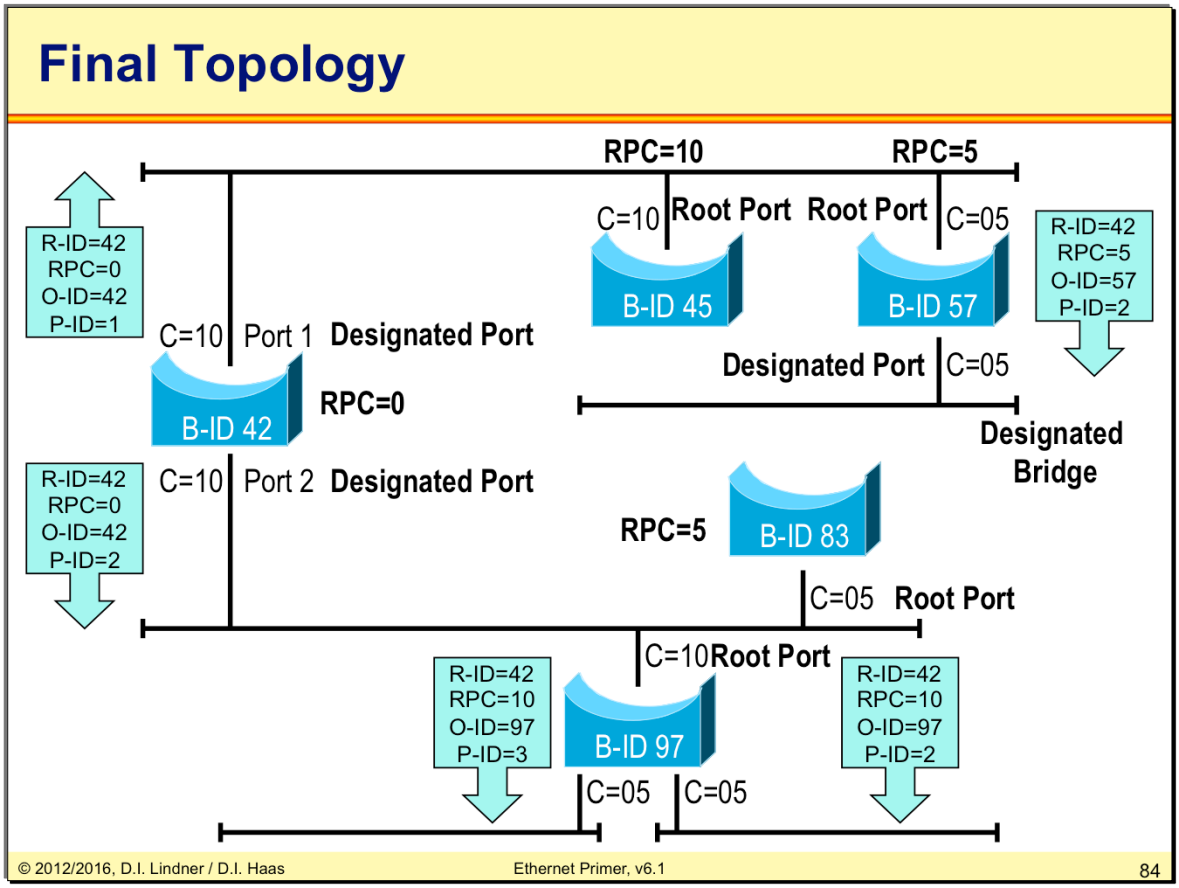
Using the Root Path Cost field in the Configuration BPDU, a bridge indicates its distance to the RB.

Strategy for decision:

If a bridge receives a Configuration BPDU from a bridge which is closer to the RB, the receiving bridge adds its own port costs to the Configuration BPDU and forwards this message to all other ports.

If a bridge receives a Configuration BPDU from a bridge which is more distant to the RB, the receiving bridge drops the message and sends its own Configuration BPDU on this port containing its own Root Path Cost.


Designated Bridge Selection (based on O-ID)
R-ID=42
RPC=0
O-ID=42
P-ID=1
C=10
Port 1
B-ID 42
R-ID=42
RPC=0
O-ID=42
P-ID=2
C=10
Port 2
C=10
B-ID 45
Designated Bridge
C=05
B-ID 57
R-ID=42
RPC=5
O-ID=57
P-ID=2
C=10
C=05
R-ID=42
RPC=5
O-ID=83
P-ID=1
C=05
B-ID 83
C=05
C=10
R-ID=42
RPC=10
O-ID=97
P-ID=3
B-ID 97
R-ID=42
RPC=10
O-ID=97
P-ID=2
C=05
C=05
© 2012/2016, D.I. Lindner / D.I. Haas
Ethernet Primer, v6.1
83



Procedure Parameters Summary:

Root Bridge -> lowest Bridge ID.

Root Ports via Root Path Costs -> which sum of costs contained in the Configuration BPDU and the receiving interface Port Costs.

Designated Bridge -> lowest Root Path Costs for a given LAN segment.

Root switch has only Designated Ports, all of them are in forwarding state.

Other switches have exactly one Root Port (RP) upstream, zero or more Designated Ports (DP) downstream and zero or more Nondesignated Ports (blocked).

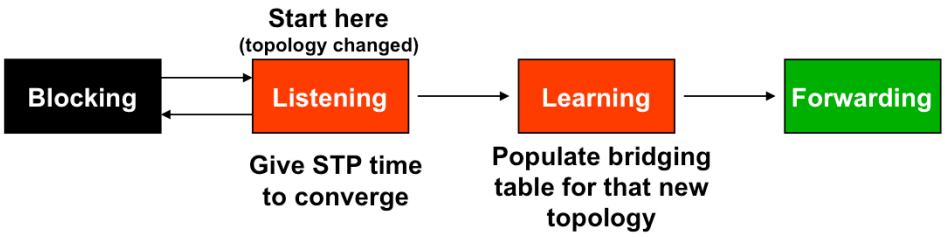
Now every designated bridge declares its ports as designated ports and puts them (together with the Root Port) in the Forwarding State.

All other bridges keep their non-RP and non-DP ports in the Blocking State.

From this moment on, the normal network operation is possible and there is only one path between any two arbitrary end systems.

Redundant links remain in active stand-by mode. If root port fails, other root port becomes active. Still it is reasonable to establish parallel paths in a switched network in order to utilize this redundancy in an event of failure. The STP automatically activates redundant paths if the active path is broken. Note that BPDUs are always sent or received on blocking ports. Note that (very-) low price switches might not support the STP and should not be used in high performance and redundant configurations.

Port States



- **At each time, a port is in one of the following states:**
  - Blocking, Listening, Learning, Forwarding, or Disabled
- **Only Blocking or Forwarding are final states (for enabled ports)**
- **Transition states**
  - 15 s Listening state is used to converge STP
  - 15 s Learning state is used to learn MAC addresses for the new topology
- **Therefore it lasts 30 seconds until a port is placed in forwarding state**

Agenda

- **Ethernet Origins**
- **Transparent Bridge**
- **Spanning Tree**
- **Ethernet Switch**
- **VLAN**
- **Spanning Tree Details**
  - Old STP
  - Convergence
  - Rapid Spanning Tree Protocol (RSTP)
- **High Speed Ethernet**

STP Error Detection

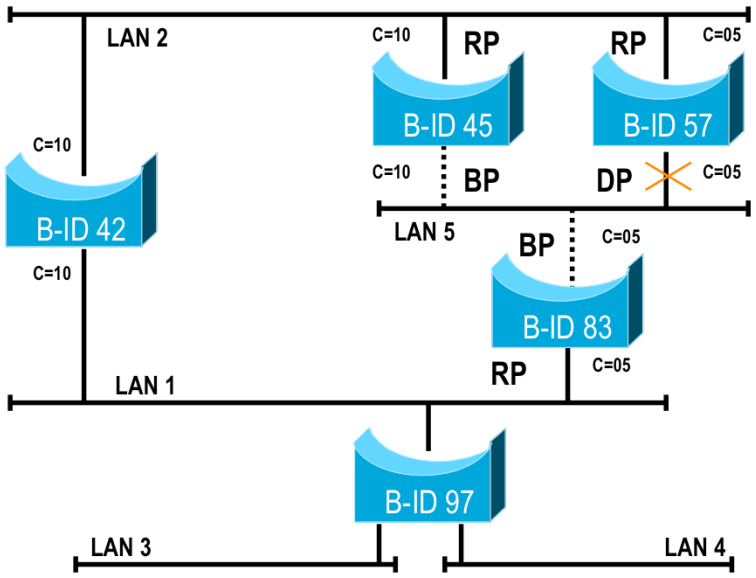
- **The root bridge generates (triggers)**
  - Every 1-10 seconds (hello time interval) a Configuration BPDU to be received on the root port of every other bridge and carried on through the designated ports
  - Bridges which are not designated are still listening to such messages on blocked ports
- **If triggering ages out two scenarios are possible**
  - Root bridge failure
    - A new root bridge will be selected based on the lowest Bridge-ID and the whole spanning tree may be modified
  - Designated bridge failure
    - If there is an other bridge which can support a LAN segment this bridge will become the new designated bridge

 Ethernet Primer, v6.1 87

Under normal conditions the root bridge generates every hello-time period a "´Heartbeat"-BPDU. All other bridges expected to hear the heartbeat and they have to pass it on in case it is received. If the heartbeat disappears – for whatever reason – however a new STP will be built. During the time of convergence (between 30 and 50 seconds for the old STP, about up to 3-5 seconds for the RSTP) any-to-any connectivity ind the LAN will be disturbed or prevented, hence we have an outage time in the network.

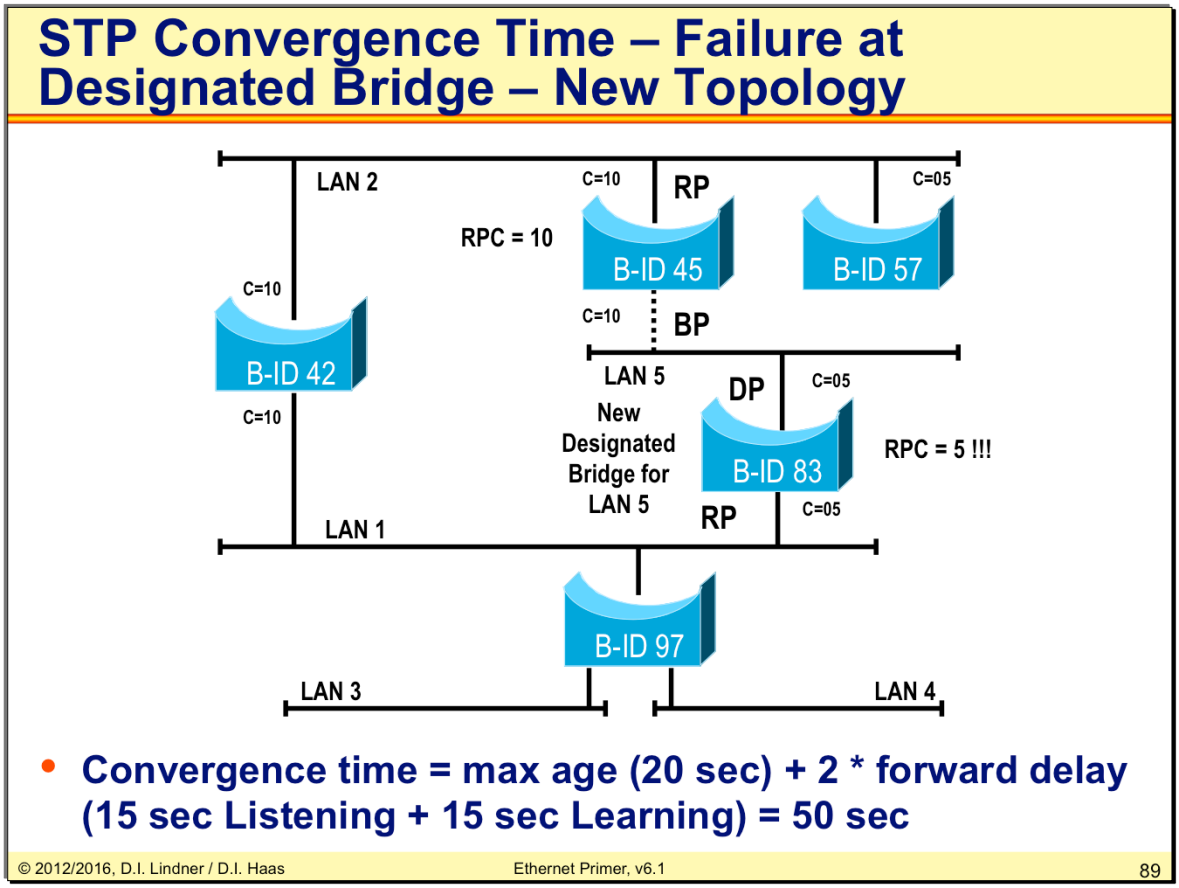
Old STP which is covered in this section is described in the IEEE 802.1D-1998 standard.

STP Convergence Time – Failure at Designated Bridge



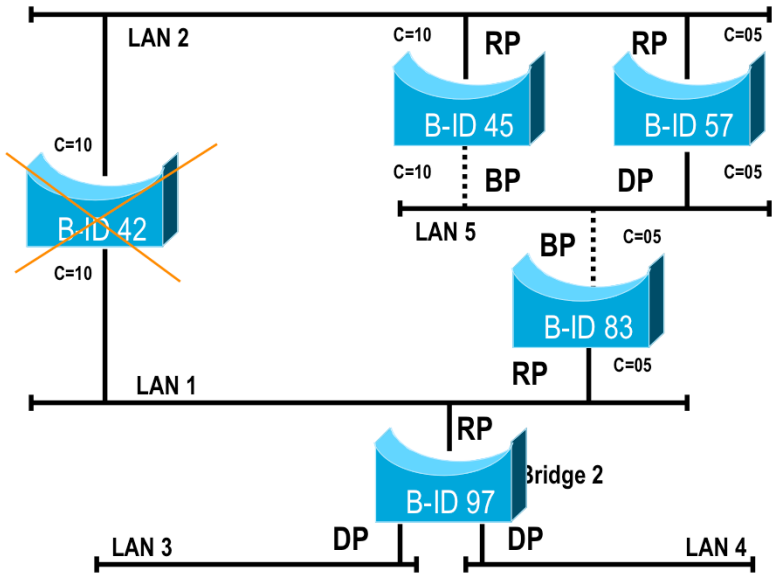
- **Time = max age (20 sec) to be waited until new STP is triggered**

Scenario 1: Designated port (DP) of Bridge 57 fails. Bridge 45 and bridge 83 do not receive the heartbeat on their blocked ports (BP) anymore although heartbeat is seen on their root ports (RP). After max-age time (20 seconds) a new STP is triggered by bridge 45 and bridge 83.



Scenario 1: Here you see the new topology. Bridge 83 became the designated bridge for LAN5.

STP Convergence Time – Failure of Root Bridge

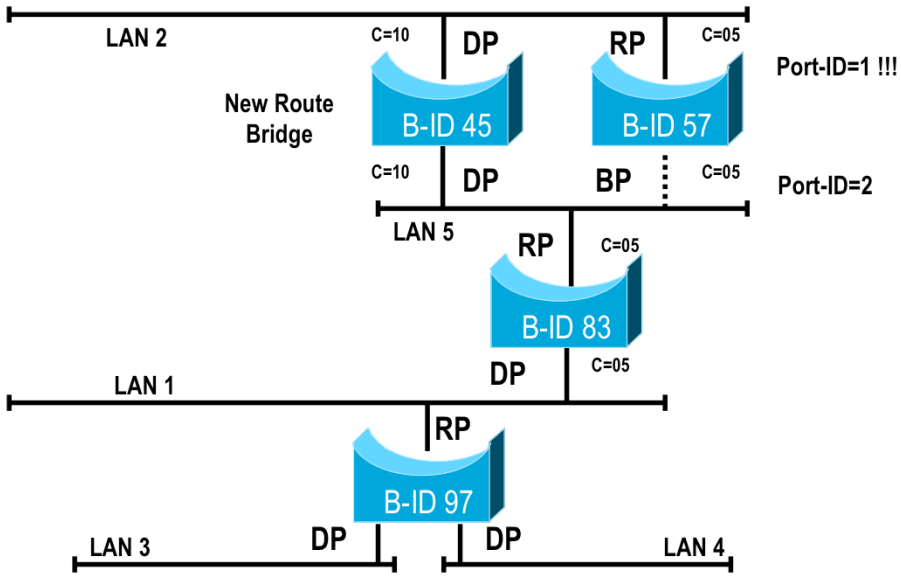


- **Time = max age (20 sec) + 2*forward delay (15 sec Listening + 15 sec Learning) = 50 sec**

 Ethernet Primer, v6.1 90

Scenario 2: Root bridge 42 fails. All other bridges do not receive the heartbeat neither on their root ports nor on their blocked ports (BP). After max-age time (20 seconds) a new STP is triggered by all remaining bridges 45.

STP Convergence Time – Failure of Root Bridge – New Topology

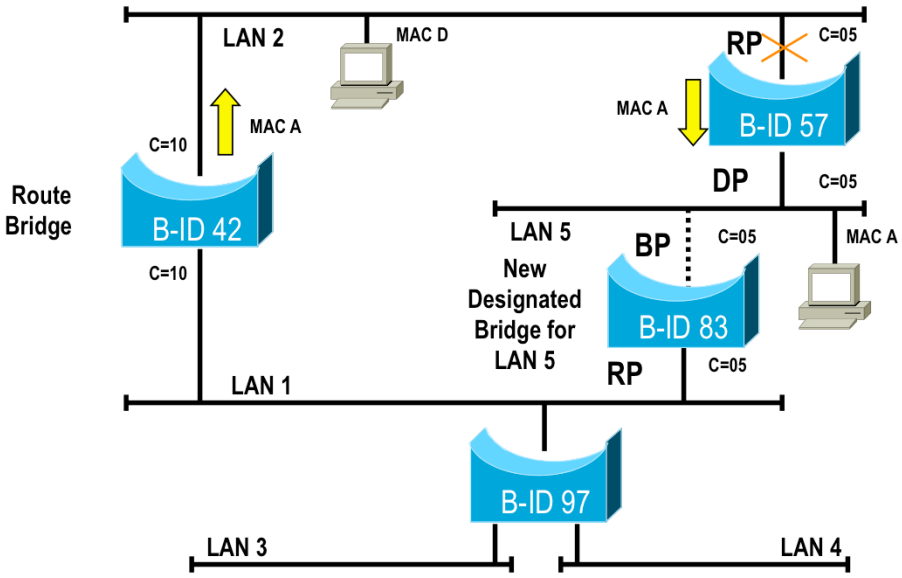


- **Time = max age (20 sec) + 2*forward delay (15 sec Listening + 15 sec Learning) = 50 sec**

 Ethernet Primer, v6.1 91

Scenario 2: Here you see the new topology. Bridge 45 became the new root bridge. Bridge 57 has equal RPC on both ports hence the port-id decides which is RP and which is BP.

STP Convergence Time – Failure of Root Port



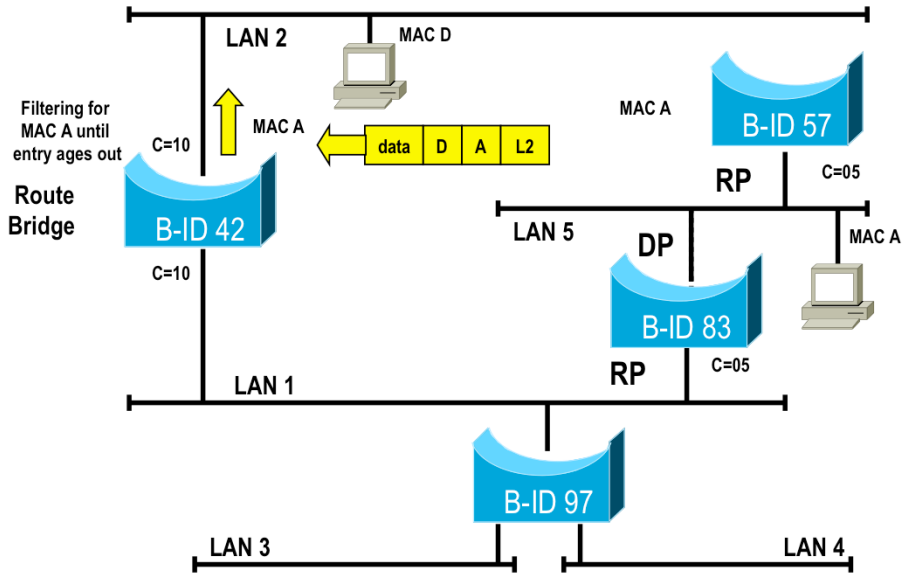
- **Time = max age (20 sec) has not to be waited until new STP is triggered**

 Ethernet Primer, v6.1 92

Scenario 3: RP of Bridge 57 fails. In that case bridge 57 has not to wait for max-age period before triggering the new STP. Reason: Bridge is designated bridge but RP fails and there is no other connectivity to the root bridge possible

Yellow arrows show the signposts in the bridging table to reach MAC address A before the failure.

STP Convergence Time – Failure of Root Port - Interruption of Connectivity D->A



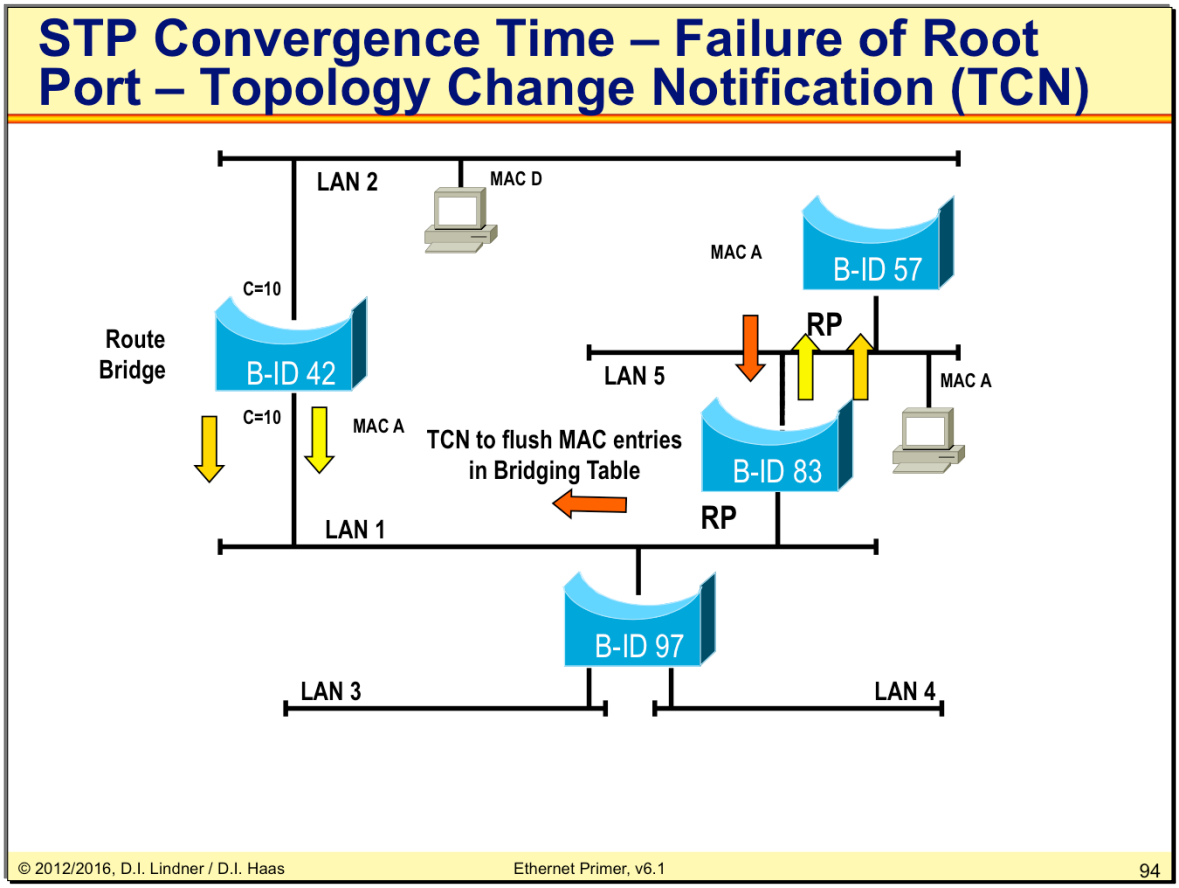
- **Convergence Time = 2*forward delay (15 sec Listening + 15 sec Learning) = 30 sec**

 Ethernet Primer, v6.1 93

Scenario 3: Here you see the new topology. Bridge 83 became the new designated bridge for LAN5.

Recognize what happens if station D sends a frame to station A. The pointer in bridge 42 still points in the wrong direction and the frame will be filtered by bridge 42 until the entry times out after 5 minutes. Of course if A would send a broadcast frame the table would immediately be repaired but what if not.

Hence bridges should install an additional procedure to overcome such situations without interaction of end-system functionality like the mentioned broadcast of A. This procedure is called topology notification.



Bridge 57 and 83 send TCN BPDUs out on their Root Ports (red arrows: TC bit set). After such a message is received by an upstream bridge it will be locally acknowledged by the upstream bridge in the reverse direction (yellow arrows: TCA bit set). If that finally appears to the root bridge, the root will sent a Conf BPDU with both flags set (orange arrows: TC and TCA bit set) for 35 seconds which has to be passed on downstream by the other bridges. All switches receiving TC+TCA=1 will age out (flush) their bridging tables in 15 seconds instead of waiting for 3 minutes.

STP Disadvantages

- **Active paths are always calculated from the root, but the actual information flow of the network may use other paths**
  - Note: network-manager can control this via Bridge Priority, Path Costs und Port Priority to achieve a certain topology under normal operation
  - Hence STP should be designed to overcome plug and play behavior resulted by default values
- **Redundant paths cannot be used for load balancing**
  - Redundant bridges may be never used if there is no failure of the currently active components
  - For remote bridging via WAN the same is true for redundant WAN links
- **Convergence time between 30 and 50 seconds**
  - Note: in order to improve convergence time Rapid Spanning Tree Protocol has been developed (802.1D version 2004)

 Ethernet Primer, v6.1 95

Note: Old STP which is covered in this section is described in the IEEE 802.1D-1998 standard.

Agenda

- **Ethernet Origins**
- **Transparent Bridge**
- **Spanning Tree**
- **Ethernet Switch**
- **VLAN**
- **Spanning Tree Details**
  - Old STP
  - Convergence
  - Rapid Spanning Tree Protocol (RSTP)
- **High Speed Ethernet**

Introduction

- **RSTP is part of the IEEE 802.1D-2004 standard**
  - Originally defined in IEEE 802.1w
  - Old STP IEEE 802.1D-1998 is now superseded by RSTP
- **Computation of the Spanning Tree is identical between STP and RSTP**
  - Conf-BPDU and TCN-BPDU still remain
  - New BPDU type "RSTP" has been added
    - Version=2, type=2
- **RSTP BPDUs can be used to negotiate port roles on a particular link**
  - Only done if neighbor bridge supports RSTP (otherwise only Conf-BPDUs are sent
  - Using a Proposal/Agreement handshake
- **Designed to be compatible and interoperable with the traditional STP – without additional management requirements**

 Ethernet Primer, v6.1 97

RSTP is designed to be compatible and interoperable with the traditional STP (IEEE 802.1D version 1998) – without additional management requirements. If an RSTP-enabled bridge is connected to an STP bridge, only Configuration-BPDUs and Topology-Change BPDUs are sent but no port role negotiation is supported.

An RSTP Bridge Port automatically adjusts to provide interoperability, if it is attached to the same LAN as an STP Bridge. Protocol operation on other ports is unchanged. Configuration and Topology Change Notification BPDUs are transmitted instead of RST BPDUs which are not recognized by STP Bridges. Port state transition timer values are increased to ensure that temporary loops are not created through the STP Bridge. Topology changes are propagated for longer to support the different FilteringDatabase flushing paradigm used by STP. It is possible that RSTP's rapid state transitions will increase rates of frame duplication and misordering.

BPDUs convey Configuration and Topology Change Notification (TCN) Messages. A Configuration Message can be encoded and transmitted as a Configuration BPDU or as an RST BPDU. A TCN Message can be encoded as a TCN BPDU or as an RST BPDU with the TC flag set. The Port Protocol Migration state machine determines the BPDU types used.

In most cases, RSTP performs better than Cisco's proprietary extensions (Port-Fast, Uplink-Fast, Backbone-Fast) without any additional configuration. 802.1w is also capable of reverting back to 802.1d in order to interoperate with legacy bridges (thus dropping the benefits it introduces) on a per-port basis.

Major Features

- **BPDUs are no longer triggered by root bridge**
  - Instead, each bridge can generate BPDUs independently and immediately (on-demand)
- **Much faster convergence**
  - Few seconds (typically within 1 – 5 seconds)
- **Better scalability**
  - No network diameter limit
- **New port roles and port states**
  - Non-Designated Port role split in Alternate and Backup
  - Root Port and Designated Port role still remain the same
  - Port state discarding instead of disabled, learning and blocking

 Ethernet Primer, v6.1 98

Remember:

Root Port Role: Receives the best BPDU (so it is closest to the root bridge).

Designated Port Role: A port is designated if it can send the best BDPU on the segment to which it is connected. On a given segment, there can be only one path towards the root-bridge.

Port States Comparison

STP (802.1d) Port State	RSTP (802.1w) Port State	Is Port included in active Topology?	Is Port learning MAC addresses?
disabled	discarding	No	No
blocking	discarding	No	No
listening	discarding	Yes	No
learning	learning	Yes	Yes
forwarding	forwarding	Yes	Yes

 Ethernet Primer, v6.1 99

There are only 3 port states left in RSTP, corresponding to the 3 possible operational states. The 802.1d states disabled, blocking and listening have been merged into a unique 802.1w discarding state.

There is no difference between a port in blocking state and a port in listening state; they both discard frames and do not learn MAC addresses. The real difference lies in the role the spanning tree assigns to the port. It can safely be assumed that a listening port will be either a designated or root and is on its way to the forwarding state. Unfortunately, once in forwarding state, there is no way to detect from the port state whether the port is root or designated, which contributes to demonstrating the failure of this state-based terminology. RSTP addresses this by decoupling the role and the state of a port.

The role is now a variable assigned to a given port. The root port and designated port roles remain, while the blocking port role is now split into the backup and alternate port roles.

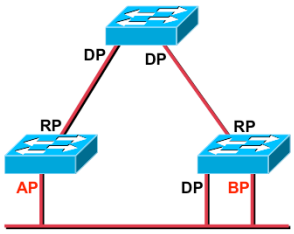
A non-designated port is a blocked port that receives a more useful BPDU than the one it would send out on its segment. The "more useful BPDU" can be received from the same switch (on another port on the same LAN segment) or from another switch (also on the same LAN segment). The first is called a **backup** port, the latter an **alternate** port.

Note: To make the confusion even worse -> The name *blocking* is used for the *discarding state* in Cisco implementations!!!

Backup and Alternate Ports

- **If a port is neither Root Port nor Designated Port**
  - It is a Backup Port – if this bridge is a Designated Bridge for that LAN
  - Or an Alternate Port otherwise

Backup and Alternate Ports:

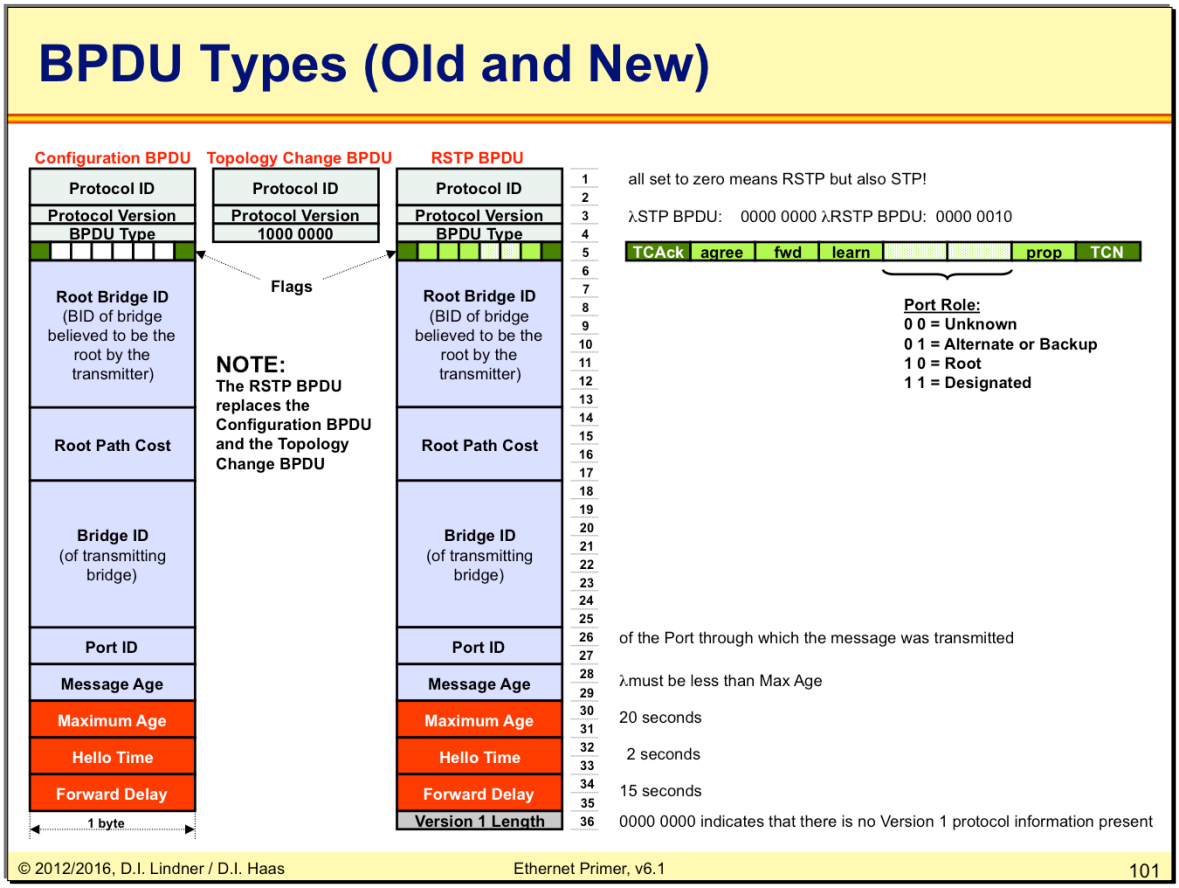


 Ethernet Primer, v6.1 100

AP alternate port, BP is now backup port.

Alternate Port: A port blocked by receiving better BPDUs from a different bridge. It provides an alternate path to the root bridge

Backup Port: A port blocked by receiving better BPDUs from the same bridge. Provides a redundant connectivity to the same segment.



Note1: A Configuration BPDU has same structure than a RSTP BPDU with the following exceptions:

1) A Configuration BPDU is only 35 byte long, that is, there is no "Version 1 length" field

2) A Configuration BPDU only uses two flags, that is, TCAck (bit 7) and TCN (bit 0)

3) BPDU type differentiate between CONF BPD and TCN BPDU

Note2: If the Unknown value of the Port Role parameter is received, the state machines will effectively treat the RST

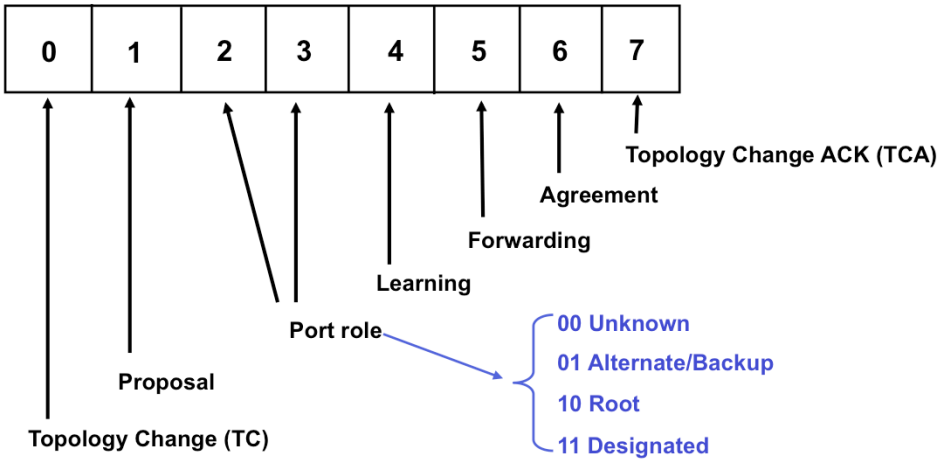
BPDU as if it were a Configuration BPDU.

Flags:

- TCN (bit 1)
- Proposal (bit 2)
- Port Role (bits 3, 4)
- Learning (bit 5)
- Forwarding (bit 6)
- Agreement (bit 7)
- Topology Change Acknowledgment (bit 8)

BPDU Flag Field – New Values

- TC and TCA were already introduced by old STP
- Other bits were unused by old STP
- RSTP also uses the 6 remaining bits

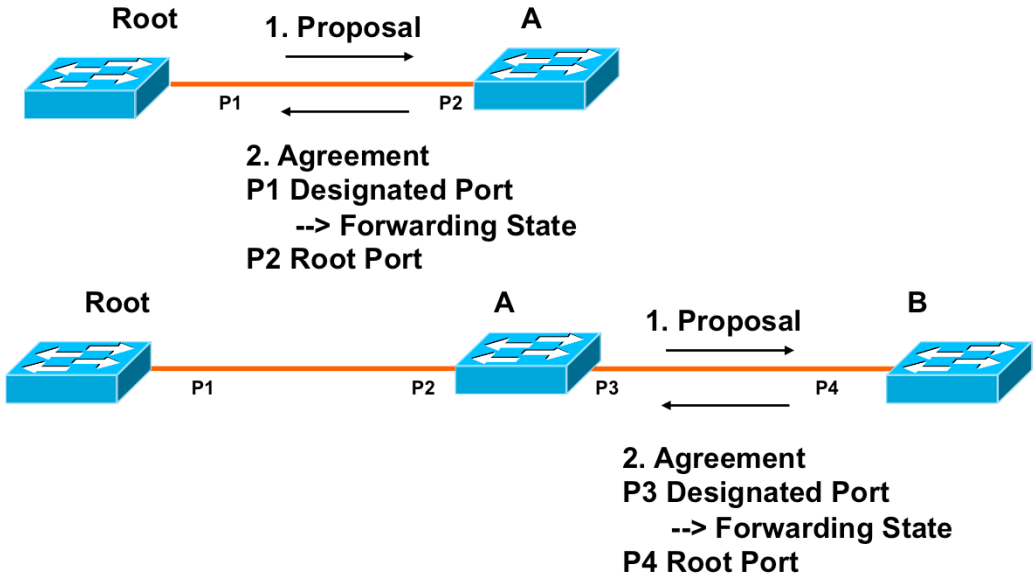


 Ethernet Primer, v6.1 102

The new bits encode the role state of the port originating the BPDU and handle the proposal/ agreement mechanism.

Proposal/Agreement Sequence

- **Suppose a new link is created between the root and switch A and a new switch B is inserted**



 Ethernet Primer, v6.1 103

There is an explicit handshake between bridges upon link up event. The bridge sends a proposal to become designated for that segment. The remote bridge responses with an agreement if the port on which it received the proposal is the root port of the remote bridge. As soon as receiving an agreement, the bridge moves the port to the forwarding state. If the remote bridge has a better role like it is nearer to the root bridge or is the root bridge itself, it will not accept the proposal but will send an own proposal. Whatever is that case, the role and state of the ports is settled within exchange of 2 or 4 messages.

NEW BPDU Handling

- **Faster Failure Detection**
  - BPDUs acting now as keepalives messages
    - Different to the 802.1D STP a bridge now sends a BPDU with its current information every <hello-time> seconds (2 by default), even if it does not receive any BPDU from the root bridge
  - If hellos are not received for 3 consecutive times, port information is invalidated
    - Because BPDU's are now used as keepalive mechanism between bridges
    - If a bridge fails to receive BPDUs from a neighbor, the connection has been lost
  - Max age not used anymore
    - For listening and waiting for STP to converge

 Ethernet Primer, v6.1 104

Rapid Transition to Forwarding State is the most important feature in 802.1w. The legacy STP was passively waiting for the network to converge before turning a port into the forwarding state New RSTP is able to actively confirm that a port can safely transition to forwarding. It is a real feedback mechanism, that takes place between RSTP-compliant bridges through proposal / agreement sequence.

Algorithm Overview

- **Designated Ports transmit Configuration BPDUs periodically to detect and repair failures**
  - Blocking (aka Discarding) ports send Conf-BPDUs only upon topology change
- **Every Bridge accepts "better" BPDUs**
  - From any Bridge on a LAN or revised information from the prior Designated Bridge for that LAN
- **To ensure that old information does not endlessly circulate through redundant paths in the network and prevent propagation of new information**
  - Each Configuration Message includes a message age and a maximum age
- **Transitions to Forwarding is now confirmed by downstream bridge**
  - Therefore no Forward-Delay is necessary!

 Ethernet Primer, v6.1 105

On a given port, if hellos are not received three consecutive times, protocol information can be immediately aged out (or if max-age expires). Because of the previously mentioned protocol modification, BPDUs are now used as a keepalive mechanism between bridges. A bridge considers that it loses connectivity to its direct neighbor root or designated bridge if it misses three BPDUs in a row. This fast aging of the information allows quick failure detection. If a bridge fails to receive BPDUs from a neighbor, it is certain that the connection to that neighbor is lost. This is opposed to 802.1D where the problem might have been anywhere on the path to the root.

Rapid transition is the most important feature introduced by 802.1w. The legacy STP passively waited for the network to converge before it turned a port into the forwarding state. The achievement of faster convergence was a matter of tuning the conservative default parameters (forward delay and max-age timers) and often put the stability of the network at stake. The new rapid STP is able to actively confirm that a port can safely transition to the forwarding state without having to rely on any timer configuration. There is now a real feedback mechanism that takes place between RSTP-compliant bridges. In order to achieve fast convergence on a port, the protocol relies upon two new variables: edge ports and link type.

Link Types and Edge Port

- **Shared Link (Half Duplex !!!)**
  - Are not supported by RSTP (ambiguous negotiations could result)
  - Instead slow standard STP is used here
- **Point-to-point Link (Full Duplex !!!)**
  - Supports proposal-agreement process
- **Edge Port**
  - Hosts reside here
  - Transitions directly to the Forwarding Port State, since there is no possibility of it participating in a loop
  - May change their role as soon as a BPDU is seen
- **RSTP fast transition**
  - Only possible on edge ports or point-to-point links

 Ethernet Primer, v6.1 106

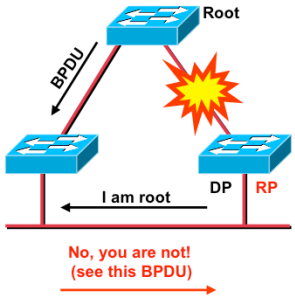
RSTP can only achieve rapid transition to forwarding: on edge ports (either full-duplex or half-duplex) or on point-to-point links (trunks between L2 switches using full-duplex), but not on shared links.

Edge ports, which are directly connected to end stations, cannot create bridging loops in the network and can thus directly perform on link setup transition to forwarding, skipping the listening and learning states of old STP.

Link type shared or point-to-point is automatically derived from the physical duplex mode of a port: A port operating in full-duplex will be assumed to be point-to-point, a port operating in half-duplex will be assumed to be a shared port.

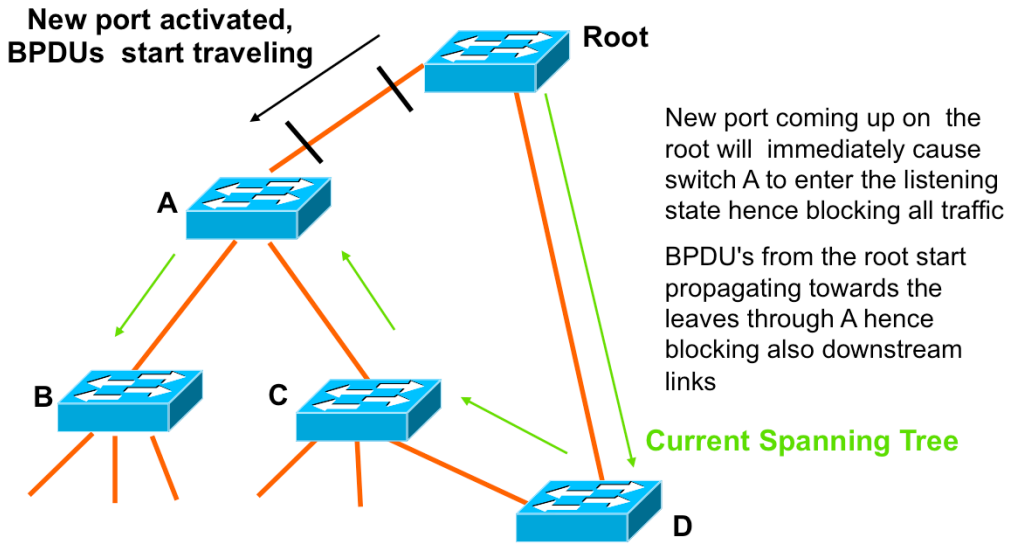
Main Differences to STP

- **BPDUs are sent every hello-time, and not simply relayed anymore**
  - Immediate aging if three consecutive BPDUs are missing
- **When a bridge receives better information ("I am root") from its DB, it immediately accepts it and replaces the one previously stored**
  - But if the RB is still alive, this bridge will notify the other via BPDUs



Slow Convergence with Legacy STP 1

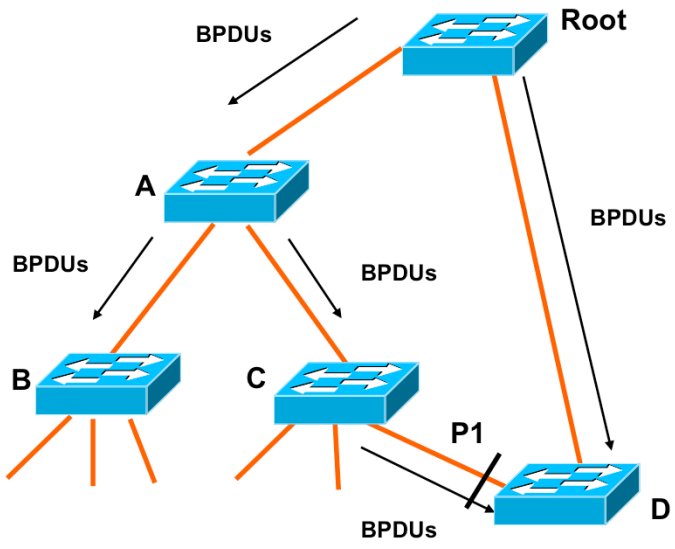
A new link between A and Root is being added to the bridged network



New port coming up on the root will immediately cause switch A to enter the listening state hence blocking all traffic

BPDU's from the root start propagating towards the leaves through A hence blocking also downstream links

Slow Convergence with Legacy STP 2

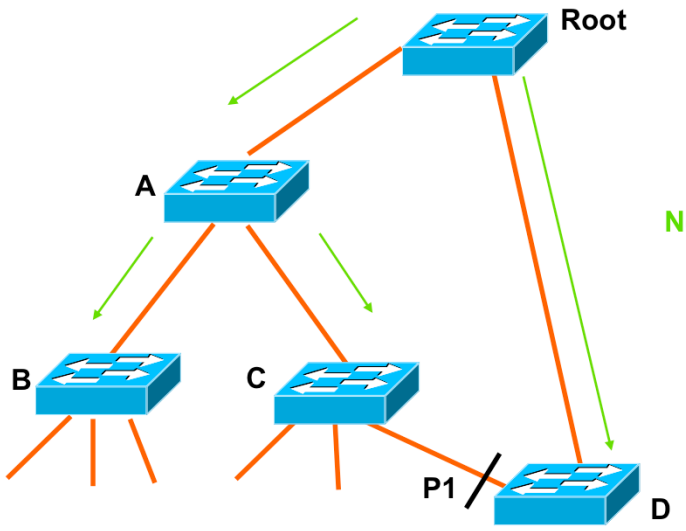


Very quickly, the BPDUs from the root bridge reach D that immediately blocks its port P1.

The topology has now converged, but the network is disrupted for twice forward delay because all switches needs time for listening (STP convergence time) and learning

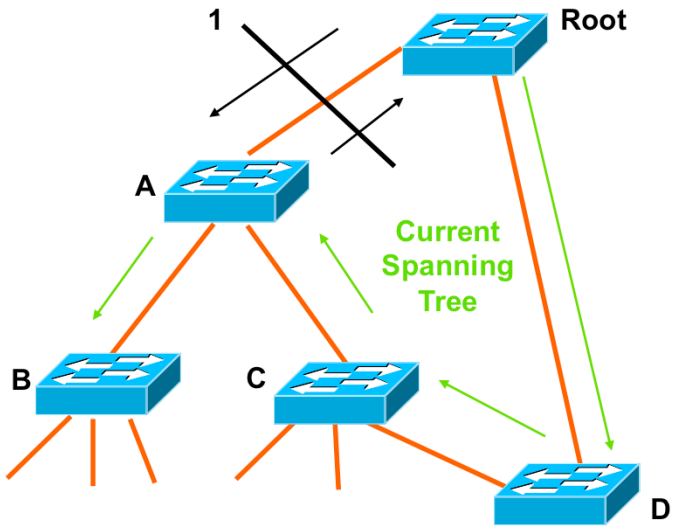
30 seconds no network connectivity !!!

Slow Convergence with Legacy STP 3



Fast Convergence with RSTP 1

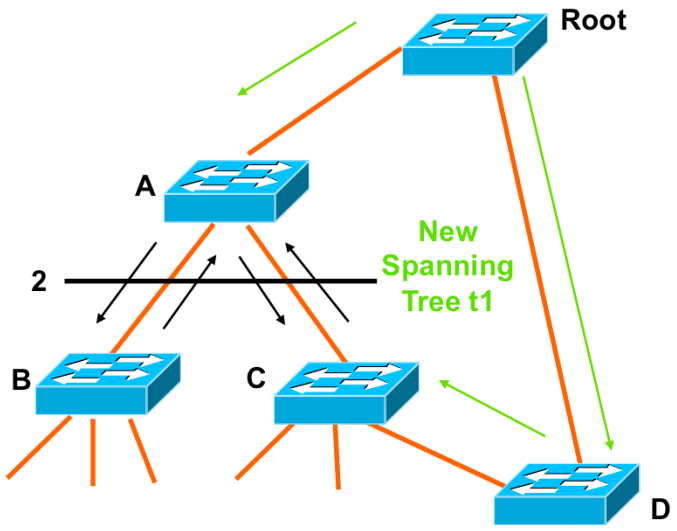
A new link between A and Root is being added to the bridged network



Both ports on link between A and the root are put in so called designated blocking as soon as they come up.

As soon as A receives the roots BPDU, it blocks its non-edge designated ports until synchronization is achieved. Through the agreement A explicitly authorizes the root bridge to put its port in forwarding

Fast Convergence with RSTP 2

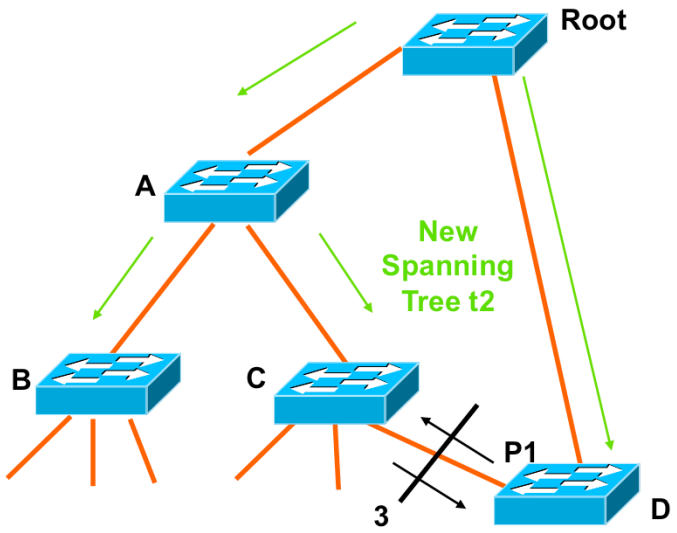


Now the link between switch A and the root is put in forwarding state.

The network below switch A is still blocking until port roles are negotiated at the next stage between switch A and switch B or A and C.

Switch B and C will enter the new spanning tree and A will put its ports in the forwarding state and the negotiations will proceed between C and D

Fast Convergence with RSTP 3



Switch C blocks its port to D because it root path costs of D are better than the root path costs of C

We have reached the final topology, which means that port P1 on D ends up blocking. It's the same final topology as for the STP example.

But we got this topology just time necessary for the new BPDU's to travel down the tree. No timer has been involved in this quick convergence.

Convergence Time < 1 second

Agenda

- **Ethernet Origins**
- **Transparent Bridge**
- **Spanning Tree**
- **Ethernet Switch**
- **VLAN**
- **Spanning Tree Details**
- **High Speed Ethernet**
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

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

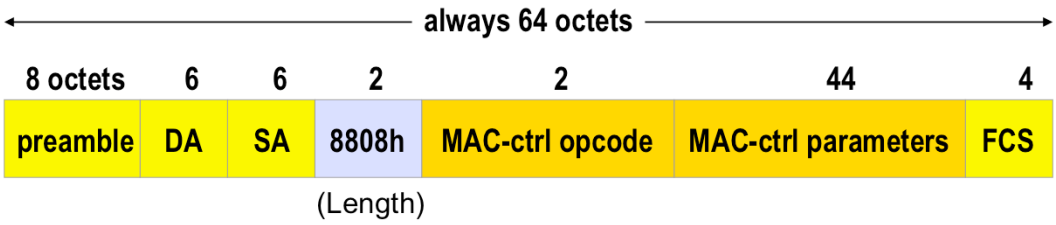
Note: Full-duplex mode is possible on point-to-point links between two elements in the network (end-system to switch, switch to switch, end-system to end-system) if there are two physical communication paths available (2 fiber optic links or 4 copper wires used for symmetrical transmission). Now CSMA/CD 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.

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**



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)**

 Ethernet Primer, v6.1 118

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

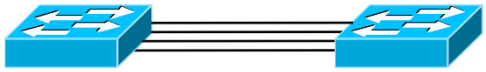
- **Destination address is either**
  - Address of destination station or
  - Broadcast address or
  - Special multicast address 01-80-C2-00-00-01
- **The special multicast address prevents bridges to transfer associated pause-frames to not concerned network segments**
  - Hence flow-control (with pause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

Ethernet Primer (v6.1)

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**"

1 Gbit/s or even 10 Gbit/s long reach connection !!!



Ether Channel

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

 Ethernet Primer, v6.1 121

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.

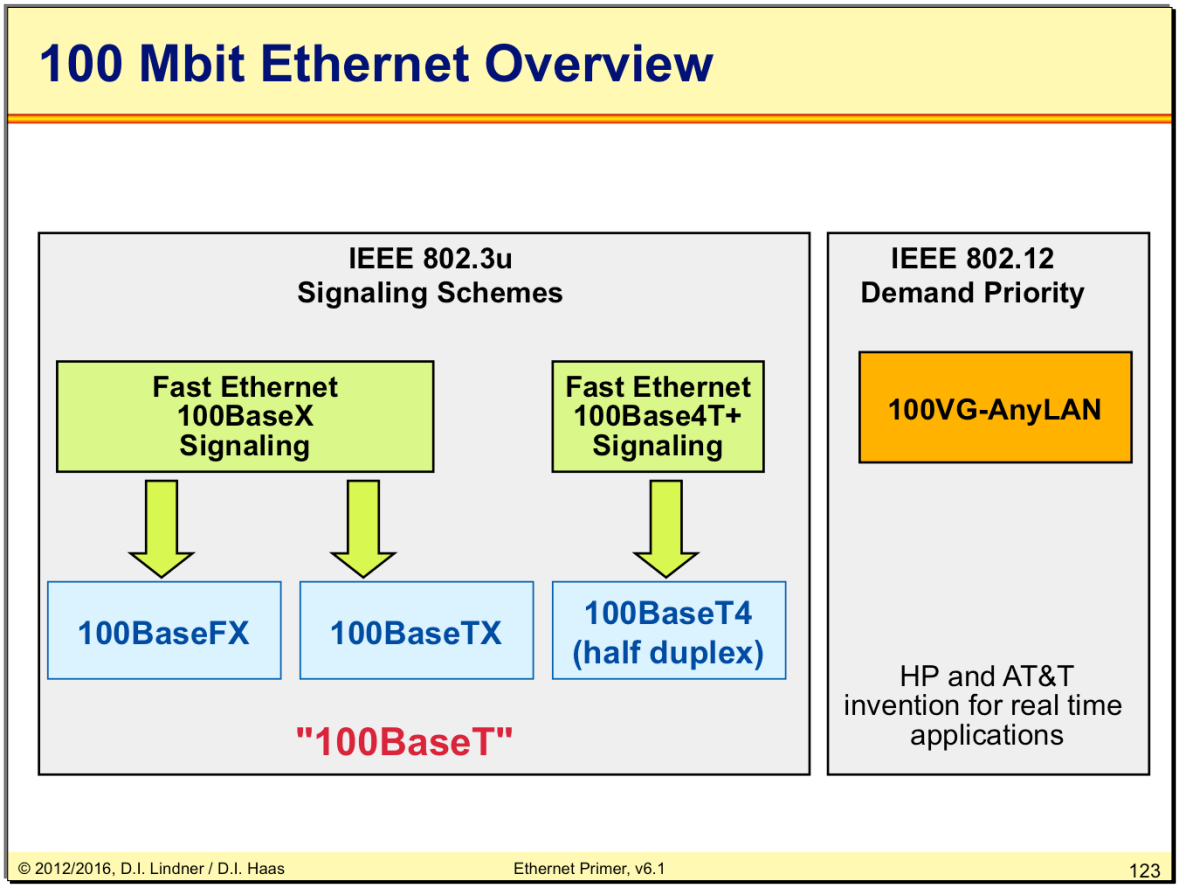
Recent experiments demonstrated the interconnection of two Ethernet Switches over a span of more than 100 km! Thus Ethernet became a WAN technology! Today, many carriers use Ethernet instead of ATM/SONET/SDH or other rather expensive technologies. GE and 10GE is relatively cheap and much simpler to deploy. 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 Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed Ethernet
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet



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.

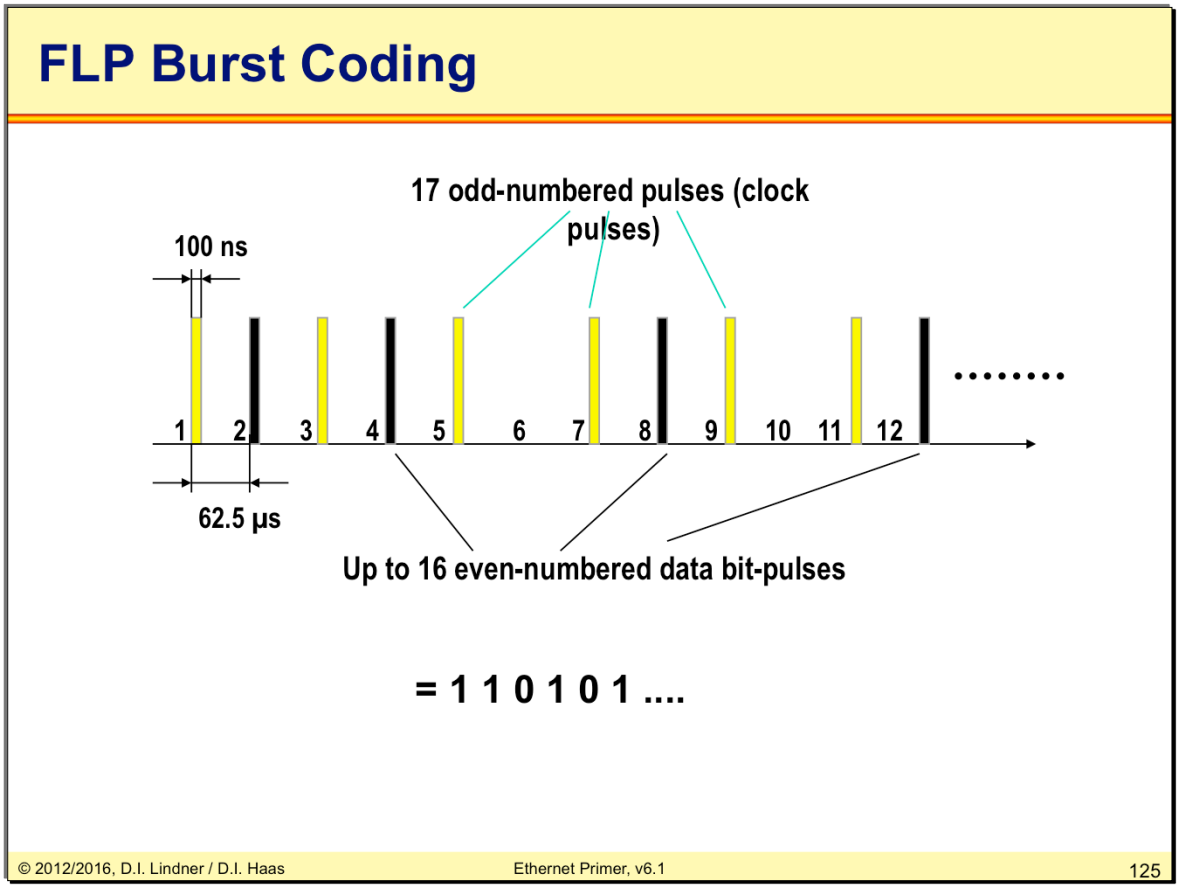
Autonegotiation (1)

- **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.



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".

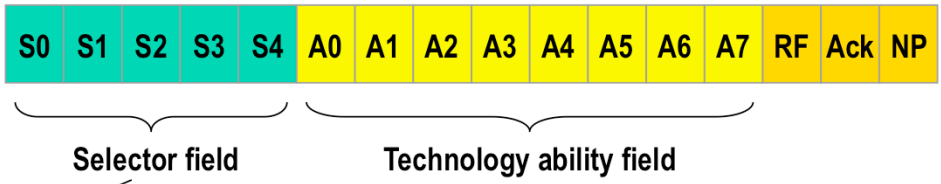
Autonegotiation (2)

- 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



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

 Ethernet Primer, v6.1 128

Remote Fault (RF)

Signals that the remote station has recognized an error

Next Page (NP)

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

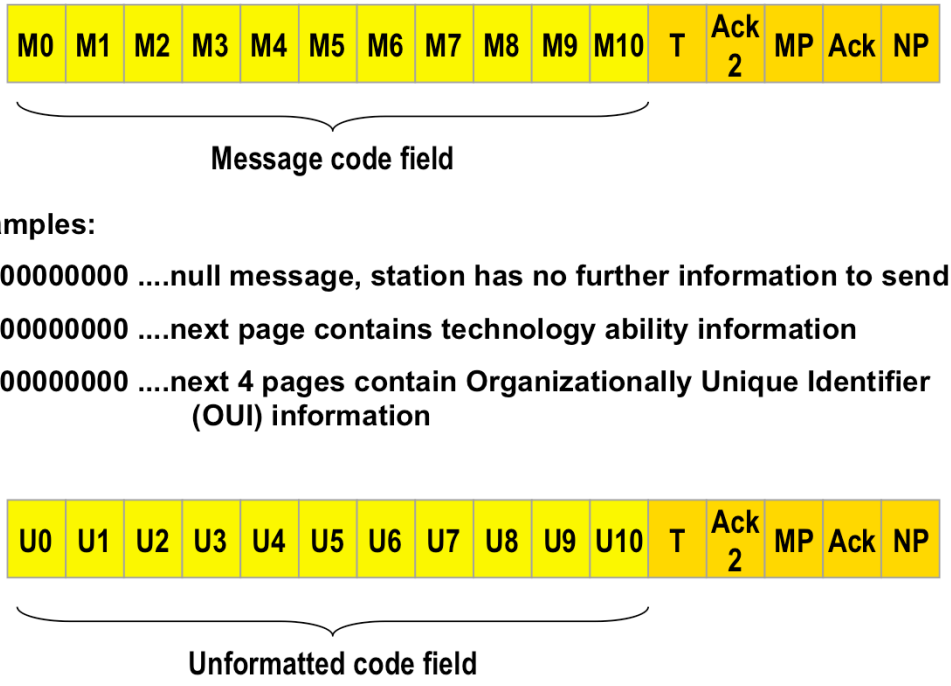
Acknowledge (Ack)

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



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)

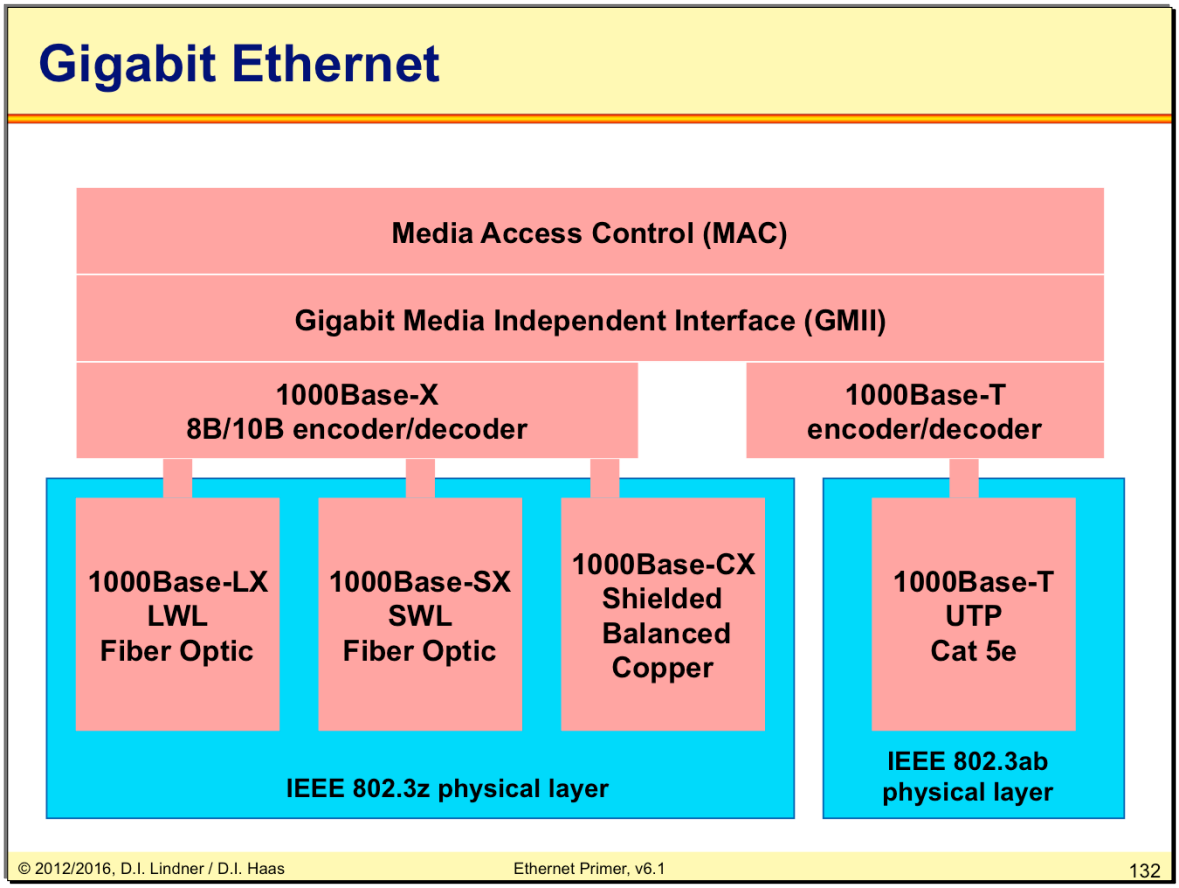
Provides synchronization during exchange of next-pages information
T-bit is always set to the inverted value of the 11th bit of the last received link-codeword

Agenda

- Ethernet Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- High Speed Ethernet
  - Introduction
  - Fast Ethernet
  - Gigabit Ethernet
  - 10 Gigabit Ethernet

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 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.

CSMA/CD Restrictions (Half Duplex Mode)

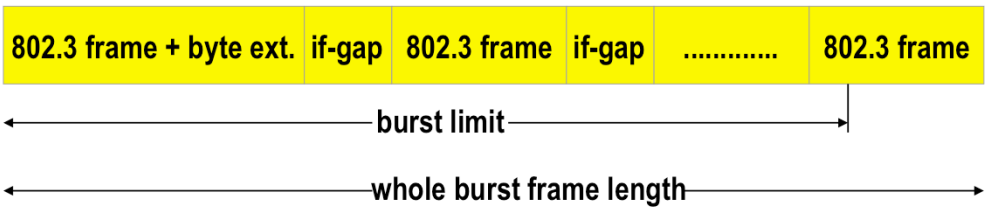
- **The conventional collision detection mechanism CSMA/CD**
  - Limits the net diameter to 20m!
- **Solutions to increase the maximal net expansion:**
  - Carrier Extension:
    - Extension bytes appended to (and removed from) the Ethernet frame by the physical layer
    - Frame exists a longer period 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 Primer, v6.1 133

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



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)

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

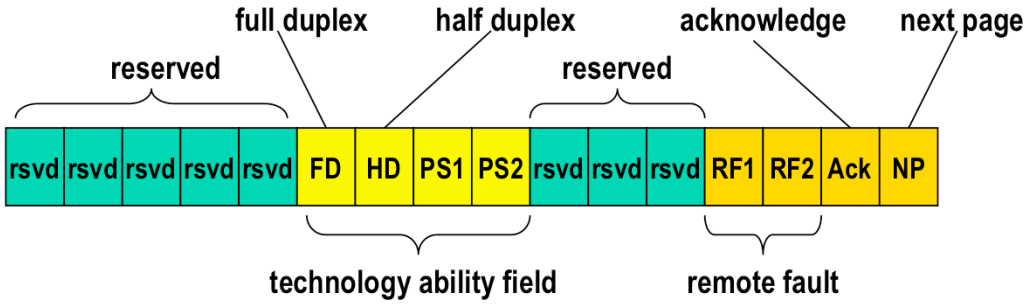
1000BaseX Autonegotiation

- **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

 Ethernet Primer, v6.1 136

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

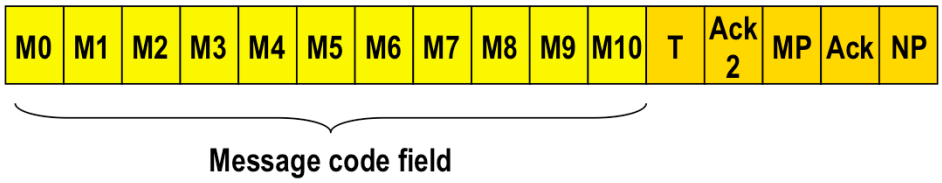


PS1	PS2	description
0	0	no pause
0	1	asymmetrical pause
1	0	symmetrical pause
1	1	symmetrical and asymmetrical pause

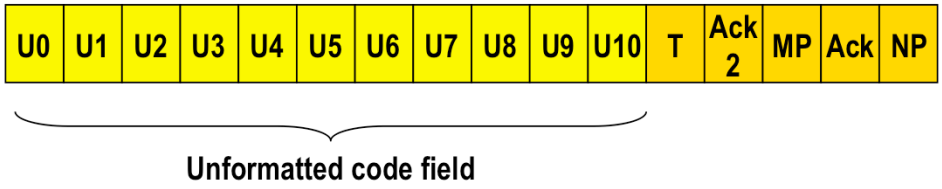
RF1	RF2	description
0	0	no error
0	1	offline
1	0	connection error
1	1	autonegotiation error (no common capabilities)

Next-Pages

Normal message-page (predefined codes)



Vendor specific page (non predefined codes)



 Ethernet Primer, v6.1 138

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 Origins
- Transparent Bridge
- Spanning Tree
- Ethernet Switch
- VLAN
- Spanning Tree Details
- 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 Primer, v6.1 141

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 Primer, v6.1 145

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!