Rechnernetze – Computer Networks

Lecture 7: Packet Switching

Prof. Dr.-Ing. Markus Fidler

Institute of Communications Technology
Leibniz Universität Hannover

May 31, 2024

Outline

Ethernet

Switches and Bridges

- Logical Link Control
- Intermediate Systems
- Spanning Tree Protocol
- Virtual LANs

IEEE 802.3 Ethernet

History of Ethernet

- 1976: invented by Metcalfe and Boggs at Xerox
  cable version inspired by Abramson's (wireless) ALOHA
- 1980: industrial standard by Digital, Intel, and Xerox (DIX)
- 1985: IEEE 802.3 standard initially with minor differences that did not prevail

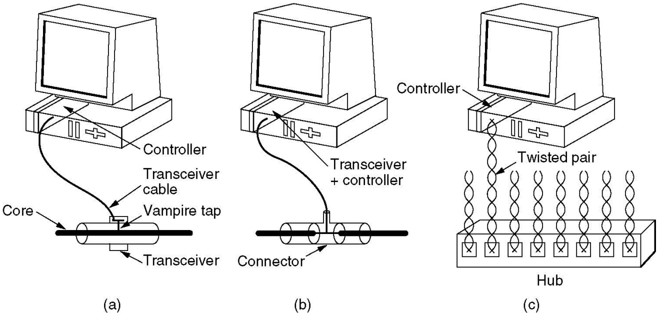
Main characteristics of classical (10 Mb/s) Ethernet

- medium access control: 1-persistent CSMA/CD
- physical layer: Manchester encoding with $\pm$ 0.85 V

Evolved versions may differ

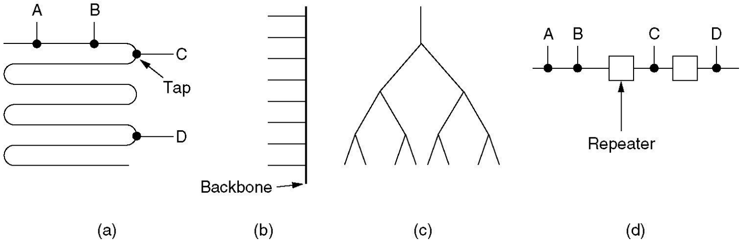
- 1995: fast Ethernet 100 Mb/s, IEEE 802.3u
- 1998: gigabit Ethernet 1 Gb/s, IEEE 802.3z
- 2002: 10-gigabit Ethernet 10 Gb/s, IEEE 802.3ae ...

Connectors



[Source: Tanenbaum, Computer Networks]

- (a) thick coax with vampire tap (error-prone)
- (b) thin coax with BNC (bayonet) connectors (T junction)
- (c) twisted pair with hub (star topology, also broadcast)



[Source: Tanenbaum, Computer Networks]

Repeaters and hubs

- physical layer devices
- receive, amplify, and retransmit/broadcast signals
- two respectively more than two interfaces

From the point of view of the data link layer a series of connected cable segments looks as if it were a single segment.

MAC frame format

Bytes	8	6	6	2	0-1500	0-46	4
(a)	Preamble	Destination address	Source address	Type	Data	Pad	Check-sum
(b)	Preamble | SOF	Destination address	Source address	Length	Data	Pad	Check-sum

[Source: Tanenbaum, Computer Networks]

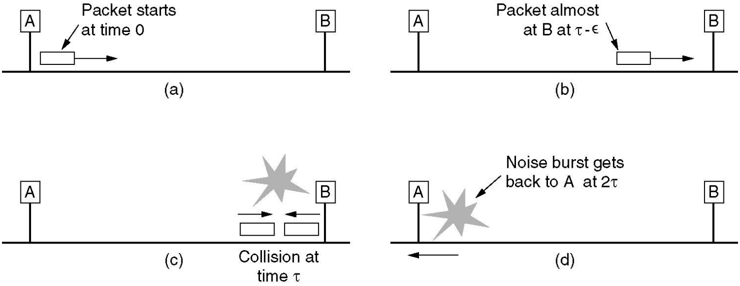
Frame format according to DIX (a), respectively, IEEE 802.3 (b)

- preamble: 0101010101... to synchronize the receiver
- start of frame (SoF): 10101011
- destination and source address: 6 byte
  - unicast (msb=0), multicast (msb=1), and broadcast (all 1s)
  - bit 46 distinguishes local from global addresses
- type/length: specifies network layer entity resp. length of data
- padding to ensure minimum frame length of 64 Byte
- checksum: 32 bit cyclic redundancy check (CRC)

Padding to ensure minimum frame size

Transmitters seek to detect collisions

- stop transmitting and retransmit after random backoff
- need to be able to detect collisions; extreme case: minimal frame size and stations at maximal distance
- need: transmission time $T_t > 2$ propagation delay T_p



[Source: Tanenbaum, Computer Networks]

Minimum frame size vs. maximum segment length

Classical Ethernet

- 10 Mb/s
- minimal frame size 64 Byte respectively 512 bit
- minimal transmission time $T_t = 51.2\ \mu$s
- maximal segment length 500 meters, at most 4 repeaters resulting in 2500 meters distance maximum
- maximal propagation delay $T_p = 12.5\ \mu$s
- hence, $T_t > 2T_p$ holds with some safety margin

Minimum frame size vs. maximum segment length

Classical Ethernet

- 10 Mb/s
- minimal frame size 64 Byte respectively 512 bit
- minimal transmission time $T_t = 51.2\ \mu$s
- maximal segment length 500 meters, at most 4 repeaters resulting in 2500 meters distance maximum
- maximal propagation delay $T_p = 12.5\ \mu$s
- hence, $T_t > 2T_p$ holds with some safety margin

Gigabit Ethernet

- 2500 meters distance requires minimal frame size of 6400 Byte
- or 25 meters distance with a minimal frame size of 64 Byte
- instead the compromise is 200 meters versus 512 Byte

The problem increases further at higher speeds!

Binary exponential backoff

Ethernet uses 1-persistent CSMA: if a station has a frame to send it starts transmitting as soon as the channel is idle

- small access delay
- high risk of collisions

Need backoff to avoid repeatedly colliding retransmissions.

Ethernet uses binary exponential backoff

- the backoff time is chosen randomly, it is uniform in $[0, w-1]$ slot times
- slot time is 51.2 μs (512 bit minimum frame size) $> 2T_p$
- the initial distribution has a width of $w_{\min} = 2$
- after each subsequent collision w is doubled up to $w_{\max} = 1024$
- after a successful transmission w is reset to $w_{\min}$

Efficiency of CSMA/CD

The efficiency η of CSMA/CD is bounded by (without derivation)

$$\eta = \frac{1}{1 + 2e\frac{Cd}{lv_l}}$$

where

- l frame length
- C capacity
- d distance
- v_l speed of light

In case of typical frame size (1500 Byte) CSMA/CD is inefficient if the capacity of the channel is high and/or if the distance is large.

Further, the capacity C is shared by all stations.

Fast, Gigabit, and 10-Gigabit Ethernet

- 1995, IEEE 802.3u: Fast Ethernet, 100 Mb/s
  - backwards compatible addendum to IEEE 802.3
  - bit time reduced from 100 ns to 10 ns
  - specified only for twisted pair cabling (e.g. Cat. 5 UTP) with hub/switch tree topology
  - the max. distance is reduced by a factor of ten, i.e. 250 m
- 1998, IEEE 802.3z: Gigabit Ethernet, 1 Gb/s
  - uses only point-to-point links with hubs or switches
  - bit duration 1 ns, distance of 25 m is unacceptable
  - min. frame size of 512 Bytes achieves max. distance of 200 m
  - frame bursting allows a station to send several frames in a row
- 2002, IEEE 802.3ae: 10-Gigabit Ethernet
  - preserves MAC frame format
  - no CSMA/CD support

Outline

Ethernet star topology

- hub
  - physical layer device
  - outputs incoming signals on all ports
  - broadcast topology (built of point-to-point links)
  - only one transmission is possible at the same time
  - one single collision domain (CSMA/CD)

Switched Ethernet

Institut für Kommunikations-Technik

Ethernet star topology

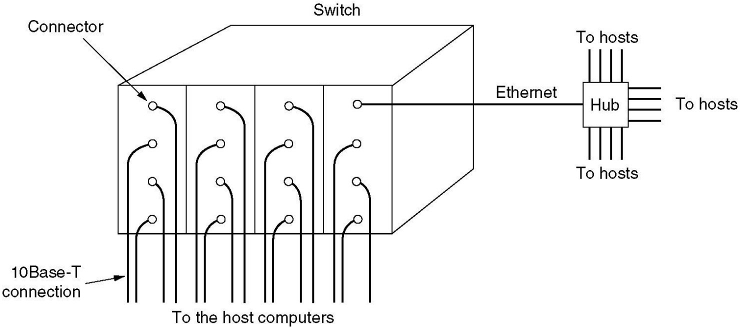
- hub
  - physical layer device
  - outputs incoming signals on all ports
  - broadcast topology (built of point-to-point links)
  - only one transmission is possible at the same time
  - one single collision domain (CSMA/CD)
- limits
  - distance, i.e., diameter of the network for collision detection
  - number of stations, since all have to share the capacity
  - CSMA/CD performance

Switched Ethernet

Ethernet star topology

- hub
  - physical layer device
  - outputs incoming signals on all ports
  - broadcast topology (built of point-to-point links)
  - only one transmission is possible at the same time
  - one single collision domain (CSMA/CD)
- switch
  - data link layer device
  - switch knows (learns) mapping of MAC addresses to ports
  - incoming frames are switched only to the right outgoing port
  - using two twisted pairs per station achieves full duplex
  - backplane with several line cards each with a number of ports
    - need buffers that temporarily store packets to avoid collisions
    - line cards can be individual collision domains
    - if ports are buffered each port is an individual collision domain

Switched Ethernet



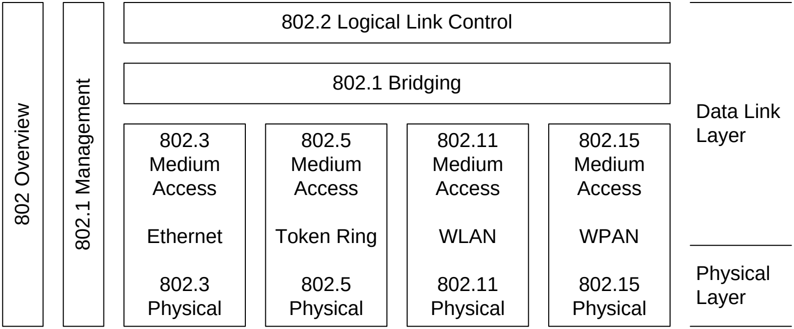
[Source: Tanenbaum, Computer Networks]

Collision domain

- hub: all connected stations
- switch: either line card or single ports

IEEE 802 LAN protocol family

The Institute of Electrical and Electronics Engineers (IEEE) published numerous standards for local area networks (LANs), of which the most important are shown.



IEEE 802.2 logical link control (LLC)

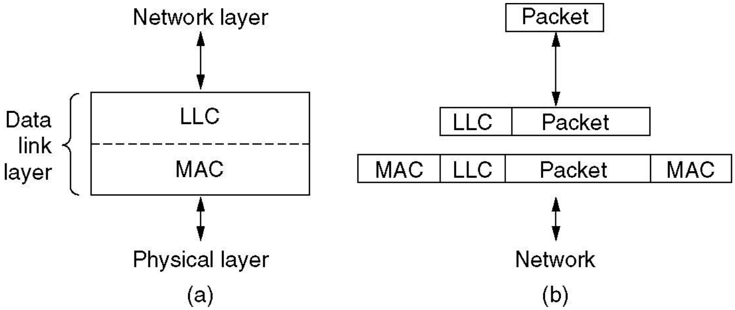
Institut für Kommunikations-Technik

Functionality

- subset of high level data link control (HDLC)
- provides a common interface to all IEEE 802 LANs

Services

- unacknowledged connectionless (unreliable datagram service)
  - error control, flow control, and sequence have to be ensured by upper layers
- acknowledged connectionless (reliable datagram service)
  - each datagram is acknowledged
- acknowledged connection oriented
  - connect, disconnect
  - error control
  - flow control
  - in sequence delivery

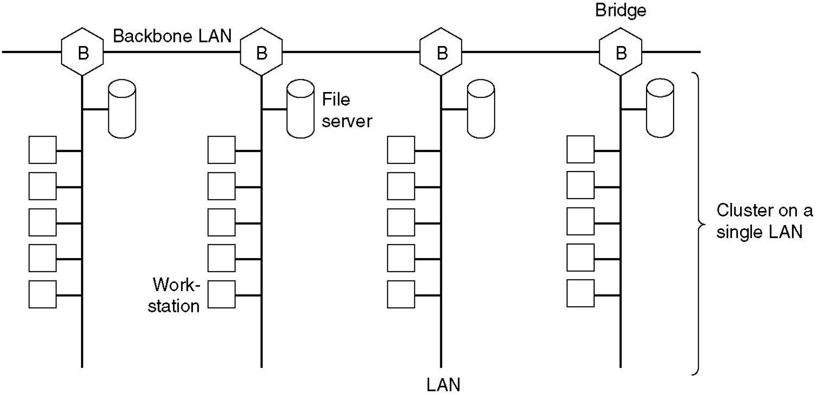


[Source: Tanenbaum, Computer Networks]

The data link layer comprises two sub-layers

- LLC: logical link control
- MAC: medium access control

Bridging



[Source: Tanenbaum, Computer Networks]

Advantages

Separate LANs connected by bridges

- individual LANs are autonomous
- individual LANs may be geographically separated
  - backbone may use e.g. optical fiber to overcome large distances
- individual LANs may use different standards (Ethernet, Wifi)
- a LAN can be split up into several LANs to handle the load
  - accommodate local traffic, e.g. file, mail, print server
- reliability
  - potentially malfunctioning stations are isolated behind bridges
- security
  - bridges forward only selected frames, avoids possibilities for sniffing (LAN interface cards have a promiscuous mode where they read all frames)

Classification: bridges, switches, hubs, and repeaters

Intermediate systems at

- data link layer
  - LLC bridges
  - MAC bridges, switches
- physical layer
  - repeater
  - hub

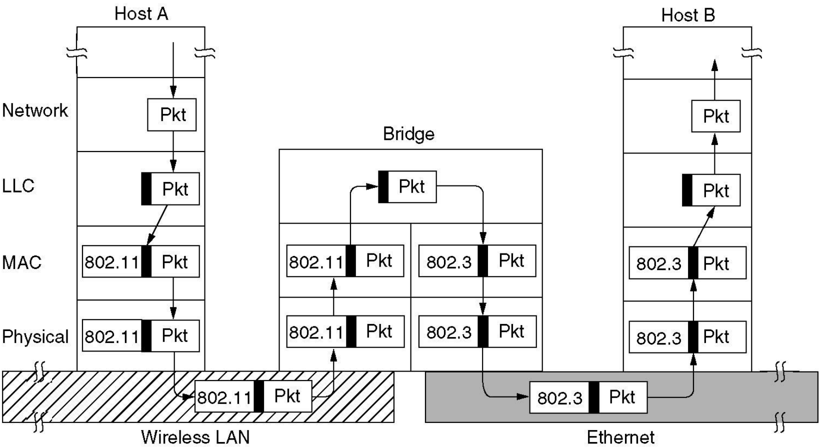
Bridges can be

- transparent
  - self-learning
  - port to MAC address mapping
  - broadcast if mapping unknown
- source routing

Application layer	Application gateway
Transport layer	Transport gateway
Network layer	Router
Data link layer	Bridge, switch
Physical layer	Repeater, hub

[Source: Tanenbaum, Computer Networks]

Example: LLC bridge



[Source: Tanenbaum, Computer Networks]

Difficulties

Bridging between different LANs is not trivial

- different MAC frame formats
  - certain MAC header fields may not have a meaning in different technologies
- different data rates
  - if connected LANs use different speeds the bridge may have to buffer frames temporarily
  - can become a serious problem if congestion becomes persistent
- different maximum frame sizes
  - different technologies use different maximum frame sizes
  - no mechanism for fragmentation of frames is specified
  - too large frames can only be discarded
- different security mechanisms
  - e.g. IEEE 802.11 Wifi uses data link layer encryption, IEEE 802.3 does not

Filtering and forwarding

Switches, respectively, bridges perform

- filtering: decide whether a frame should be forwarded or just dropped
- forwarding: determine the interface to which a frame is directed

The filtering/forwarding decision is based on the switch table. The switch indexes the table by the destination's MAC address

- if there is no entry in the table the switch copies the frame to all interfaces except for the incoming interface, i.e. broadcast
- if there is an entry it contains the outgoing interface
  - if the target outgoing interface is the same as the incoming interface the switch just discards the frame
  - if the target outgoing interface is different than the incoming interface the switch forwards the frame

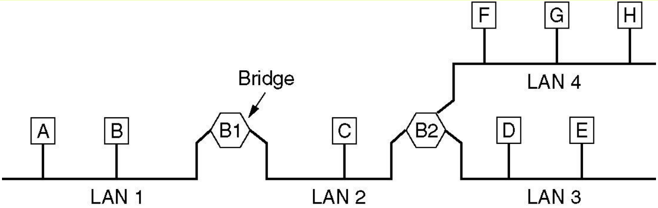
Self-learning

Transparent switches, respectively, bridges build their switching table autonomously

- initially the switch table is empty
- whenever the switch receives a frame it stores in its table
  - the MAC source address field
  - the incoming interface
  - the time of reception
- table entries are deleted if no frame from the respective source address is received for some time (aging time = few minutes)

So-called backward learning $\Rightarrow$ switches are plug and play devices!

Example

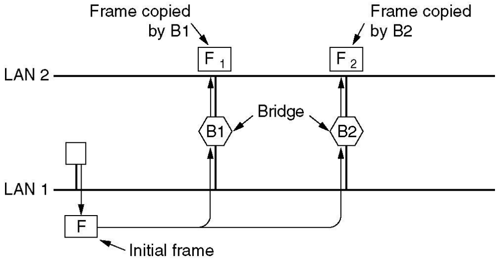


[Source: Tanenbaum, Computer Networks]

Example: all bridges start with empty switch tables

- A sends to B
  - B1 stores (A,LAN1) and sends the frame on LAN2
  - B2 stores (A,LAN2) and sends the frame on LAN3 and LAN4
- B sends to A
  - B1 stores (B,LAN1) and discards the frame
- E sends to A
  - B2 stores (E,LAN3) and sends the frame on LAN2
  - B1 stores (E,LAN2) and sends the frame on LAN1

Redundant bridges



[Source: Tanenbaum, Computer Networks]

Redundant bridges increase reliability but introduce loops

- destination of frame F is not known to either of the bridges
- both bridges copy frame F to LAN2 resulting in F1 and F2
- B1 sees F2 and B2 sees F1, both with unknown destination
- both bridges copy F2 and F1, respectively, to LAN1
- ... ⇒ broadcast storm

Solution: spanning tree topology

Before starting to forward frames the bridges organize as a tree.

- the spanning tree reaches every LAN
- some potential connections between LANs (bridges) are ignored/disabled
- the spanning tree topology is loop-free

The result is a topology where there exists only one unique path for any source destination pair.

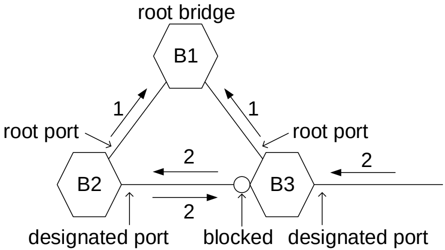
The tree is constructed by the spanning tree protocol (STP)

- bridges maintain tables about the topology
- bridges disable certain ports to avoid loops
- table entries expire after a certain timeout if not refreshed
- STP runs continuously to update the tree (in case of failures)

Algorithm for construction of the spanning tree

- a **root bridge** is selected to form the root of the tree
  - each bridge has a unique ID
  - the bridge with the smallest ID is selected as the root
- each bridge determines shortest paths to the root bridge
  - bridges identify the port that is on the shortest path to the root as their **root port**
- on each LAN the bridge that offers the shortest path to the root is selected
  - the respective port of the bridge is called the **designated port**
- ports that are not root nor designated ports are blocked from data transmission
- tie-breaking rule: whenever two paths have the same length, the bridge that has the smaller ID wins

Example



- B1 has the smallest index and becomes root
- B2 and B3 select their root port
- using a tiebreaking rule B2 is chosen as designated bridge, marks designated port for the LAN connected to B2 and B3
- B3 is designated bridge for the LAN connected only to B3

Bridges do not see the topology of the entire network nor the bridges' IDs $\rightarrow$ need to exchange configuration messages

- configuration messages contain
  - ID for what the bridge believes to be the root bridge
  - distance in hops from the sending bridge to the root bridge
  - ID of the bridge that sends the message
- bridges record the best configuration message; a configuration message is better if
  - it identifies a root with a smaller ID
  - it identifies a root with an equal ID but with a shorter distance
  - it identifies a root with an equal ID and an equal distance but the sending bridge has a smaller ID

Spanning tree protocol

Exchange of configuration messages:

- initially, bridges have no information
  - each bridge thinks it is the root bridge
  - root bridges periodically send corresponding configuration messages on all of their ports
- whenever a bridge receives a configuration message that is better than the currently recorded information, the bridge
  - adds one to the distance field
  - stores the new information (and discards the old information)

Spanning tree protocol

Exchange of configuration messages continued:

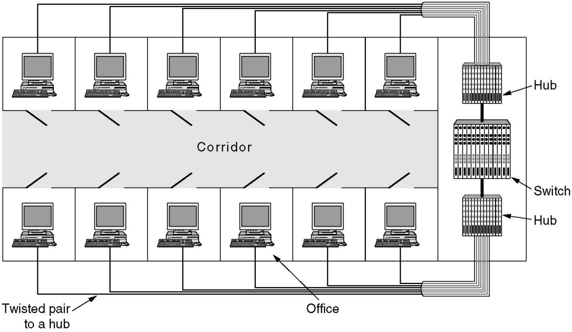
- if a bridge receives a configuration message that indicates that it is not the root bridge, it
  - stops generating configuration messages
  - only forwards configuration messages from other bridges, after adding one to the distance field
- if a bridge receives a configuration message that indicates that it is not the designated bridge for a LAN, it
  - stops sending configuration messages over that port
- if a bridge does not receive configuration messages for a specified period it assumes a failure and starts generating configuration messages again

Separating LANs

Requirement: place stations on separate but interconnected LANs

- increase security and privacy
- perform load balancing

Need separate hubs for each LAN, wire stations accordingly



[Source: Tanenbaum, Computer Networks]

Need to group users on LANs

- decouple logical from physical topology (organizational structures instead of physical layout of the building)
- organizational changes appear frequently, reconfiguring the logical topology means, however, rewiring the network

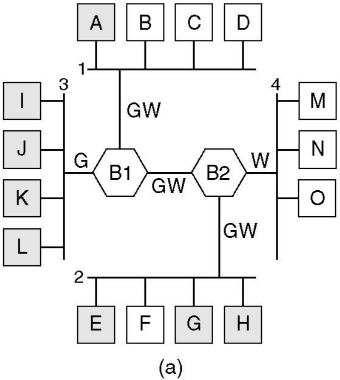
Virtual LANs

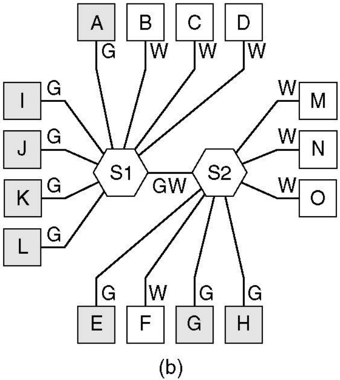
- map several logical networks onto one physical network
- perform the 'rewiring' in software

Mark different VLANs by different colors, e.g. gray and white.
VLAN enabled bridges/switches

- map e.g. MAC addresses or incoming ports to VLAN colors
- map VLAN colors to outgoing ports
- broadcast frames only on 'correctly colored' ports

Example

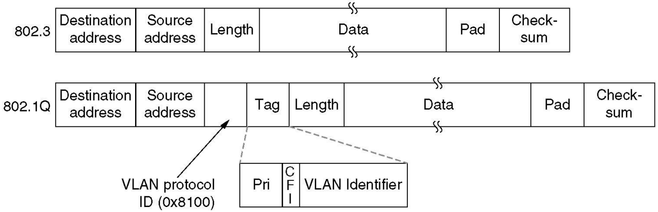




[Source: Tanenbaum, Computer Networks]

Virtual LANs are marked by colors: gray (G) and white (W)

Ethernet VLAN frame format

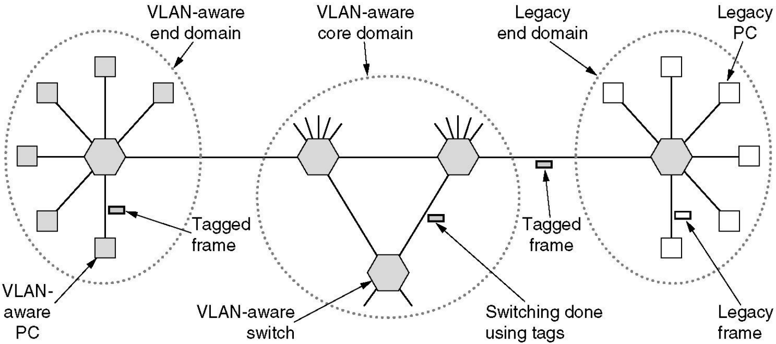


[Source: Tanenbaum, Computer Networks]

IEEE 802.1Q defines a new frame format

- 0x8100 in the length field indicates an IEEE 802.1Q frame
- non IEEE 802.1Q capable devices discard IEEE 802.1Q frames since 0x8100 is not permitted, the maximum length is 1500
- IEEE 802.1Q frames have an additional tag
  - 3 bit priority for quality of service
  - 1 bit canonical format indicator (for encapsulation of 802.5)
  - 12 bit VLAN identifier (explicit tagging)

VLAN domains



[Source: Tanenbaum, Computer Networks]

VLAN aware bridges can also be used as plug and play devices

- learn which VLANs are mapped to which ports autonomously
- received VLAN frames provide the necessary information