Advanced Computer Networking (ACN)

IN2097 – WiSe 2023–2024

Prof. Dr.-Ing. Georg Carle

Sebastian Gallenmüller, Max Helm, Benedikt Jaeger,
Marcel Kempf, Patrick Sattler, Johannes Zirngibl

Chair of Network Architectures and Services
School of Computation, Information, and Technology
Technical University of Munich

Link-Layer Protocols

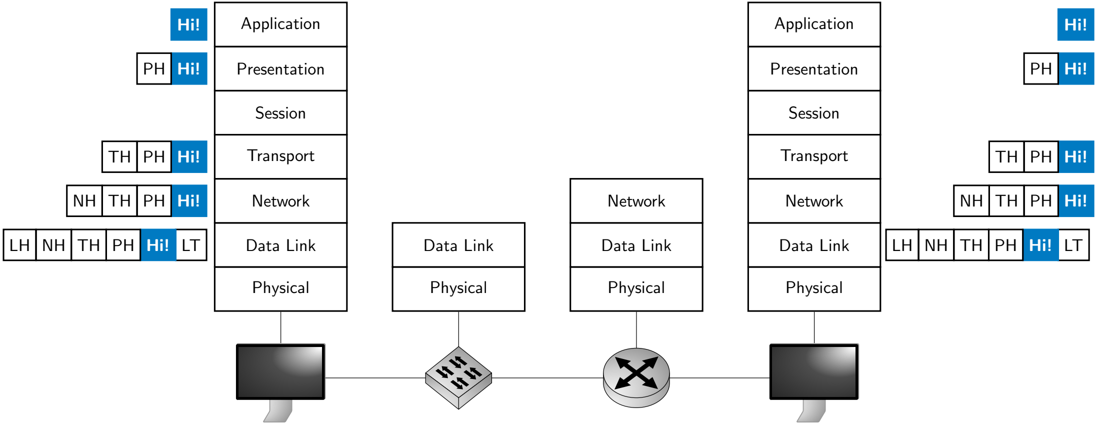
Link-Layer Protocols

Protocol mechanisms
Contents

All or some of the following:

- Addressing/naming: manage identifiers
- Fragmentation: divide large message into smaller chunks to fit lower layer
- Re-sequencing: reorder out-of-sequence protocol data units (PDUs)
- Error control: detection and correction of errors and losses
  - retransmission; forward error correction
- Flow control: avoid flooding/overwhelming of receiver
- Congestion control: avoid flooding of slower network nodes/links
- Resource allocation: administer bandwidth, buffers, CPU among contenders
- Multiplexing: combine several higher-layer sessions into one “channel”
- Compression: reduce data rate by encoding
- Privacy, authentication: security policy (against listening/exploitation)

Protocol mechanisms
Protocol layering



Send side (layer N)

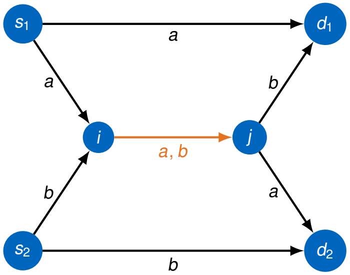
1. input: header + payload of layer N+1
2. extend input with header of layer N
3. output: pass extended data to layer N-1

Receive side (layer N)

1. input: payload of layer N–1
2. process data and remove header of layer N from input
3. output: pass payload of layer N to layer N+1

Protocol mechanisms
Forwarding/routing vs. network coding

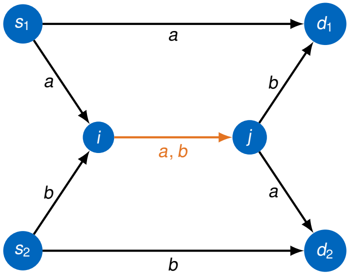
Nodes d_1 and d_2 should receive messages a, b



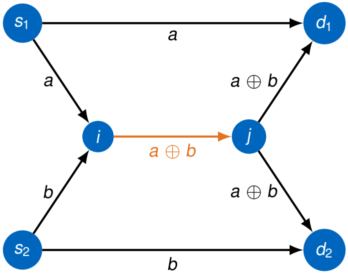
- Forwarding and routing
- Only one packet can be transmitted via a single link at the same time
- Bottleneck at link between i and j

Protocol mechanisms
Forwarding/routing vs. network coding

Nodes d_1 and d_2 should receive messages a, b



- Forwarding and routing
- Only one packet can be transmitted via a single link at the same time
- Bottleneck at link between i and j



- Network coding
- Transmits a single, modified packet $a \oplus b$ between i and j (no bottleneck!)
- d_1 and d_2 can reconstruct original packets from the two received packets respectively

Protocol mechanisms
Forwarding/routing vs. network coding

Advanced protocol mechanisms

- **Network Coding**
  - A different type of routing
  - Nodes in a network combine packets possibly from different sources and generate groups of encoded packets
  - Network coding allows to achieve maximum possible information flow in a network
  - Covered in specific lecture Network Coding (IN2315)
  - Outgoing packets are arbitrary combinations of previously received packets
  - Coding, i.e. combining packets, may happen on any node in the network (in contrast to FEC)
- **Traditional routing and forwarding**
  - Routing determines best paths from source to destination
  - Packets are forwarded by switches and routers along one of these paths
  - Packet payloads remain unaltered

Protocol mechanisms
Protocol layering

Observation

- Certain protocol mechanisms of one layer also used in other layer
- Examples:
  - layer 4 mechanism (e.g., TCP ACKs & retransmissions) as also used in layer 2 (e.g., WLAN retransmissions)
  - routing in layer 3, but with certain technologies (ATM, MPLS) also below

Protocol mechanisms
Protocol layering

Observation

- Certain protocol mechanisms of one layer also used in other layer
- Examples:
  - layer 4 mechanism (e.g., TCP ACKs & retransmissions) as also used in layer 2 (e.g., WLAN retransmissions)
  - routing in layer 3, but with certain technologies (ATM, MPLS) also below

True definition of a layer n protocol (by Radia Perlman)

- Anything designed by a committee whose charter is to design a layer n protocol

Protocol mechanisms
Layering considered harmful?

Benefits of layering

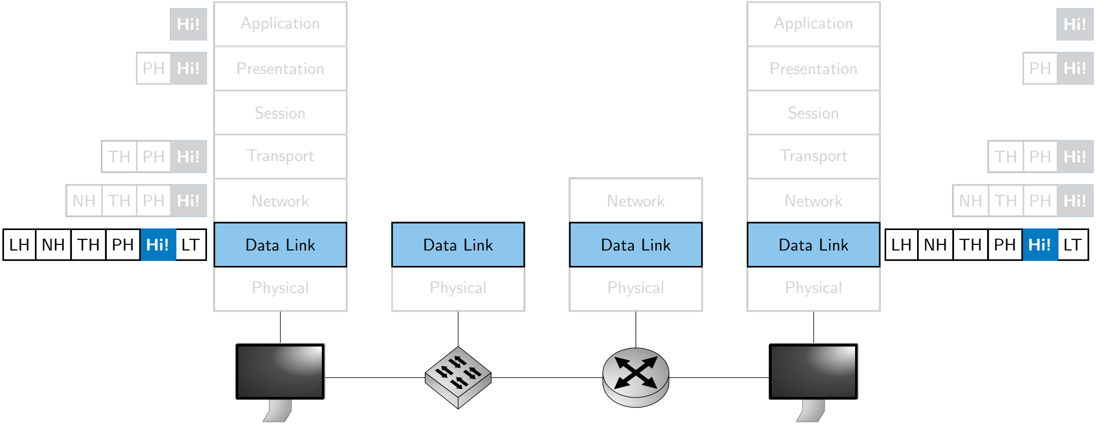
- Need layers to manage complexity
  - don’t want to reinvent Ethernet-specific protocol for each application
- Common functionality
  - “ideal” network

but:

- Layer N may duplicate lower layer functionality (e.g. error recovery)
- Different layers may need same information
- Layer N may need to peek into layer N+x

Link-Layer Protocols

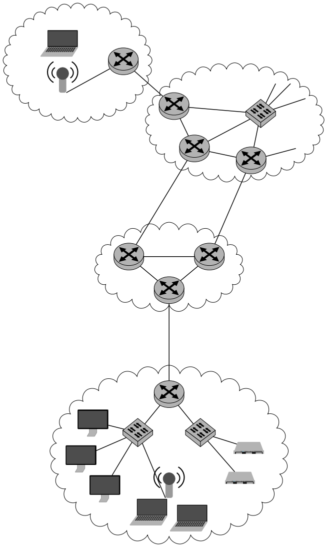
Link Layer


Hi!
PH Hi!
TH PH Hi!
NH TH PH Hi!
LH NH TH PH Hi! LT
Application
Presentation
Session
Transport
Network
Data Link
Physical
Data Link
Physical
Network
Data Link
Physical
Application
Presentation
Session
Transport
Network
Data Link
Physical
Hi!
PH Hi!
TH PH Hi!
NH TH PH Hi!
LH NH TH PH Hi! LT

Link Layer
Link layer terminology

- Hosts and routers are nodes
- Communication channels that connect adjacent nodes along communication path are links
  - wired links
  - wireless links
  - LANs
- Layer-2 packet is called frame
- Layer-3 packet often called packet, sometimes also datagram

The data link layer has the responsibility of transferring a datagram from one node to an adjacent node over a link.



Link Layer
Services

Framing, link access

- Encapsulate datagram into frame, adding header, trailer
- Channel access if shared medium
- “MAC” addresses used in frame headers to identify source and destination node
  - different from IP address!
  - Question: *Why are there different addresses at L2 and L3?*

Reliable delivery between adjacent nodes

- Rarely used on low bit-error rate links (fiber, some twisted pair)
- Wireless links: high error rates
  ► L2 retransmission scheme, e.g., in wireless LAN (IEEE 802.11)
  - Question: *Why both link-level and end-to-end reliability?*

Link Layer
Services Continued

Flow control

- Pacing between adjacent sending and receiving nodes

Error detection

- Errors caused by signal attenuation, noise
- Receiver detects presence of errors:
  - signals sender for retransmission or drops frame

Error correction

- Receiver identifies and corrects error(s)
  - Error correcting codes: correcting bit errors without retransmission
  - Terminology “error correction” may include retransmissions

Half-duplex and full-duplex

- With half duplex, nodes at both ends of link can transmit, but not at same time

Link Layer
Two types of "links"

Point-to-point

- point-to-point link between Ethernet switch and host
- PPP for dial-up access

Broadcast (shared wire or medium)

- old-fashioned Ethernet
- upstream HFC (Hybrid Fiber Coax)
- 802.11 wireless LAN

Multiple access protocols

Situation

- Single shared broadcast channel
- Two or more simultaneous transmissions by nodes: interference
  - **Collision** if node receives two or more signals at the same time

Definition of a **Multiple access protocol**:

- Distributed algorithm that determines how nodes share channel, i.e., determine when node can transmit
- Communication about channel sharing uses channel itself, i.e., no out-of-band channel for coordination

MAC Protocols: A Taxonomy (Three broad classes)

Channel Partitioning

- Divide channel into smaller “pieces” (time slots, frequency, code)
- Allocate piece to node for exclusive use

Random Access

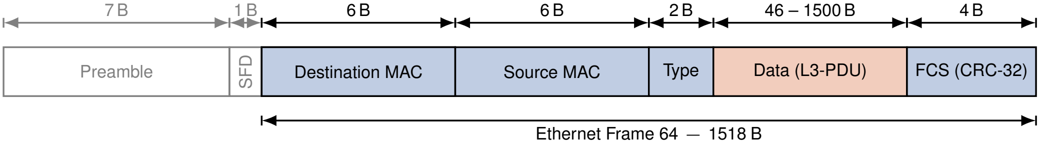
- Channel not divided, allow collisions, “recover” from collisions
- Examples of random access MAC protocols:
  - ALOHA, slotted ALOHA
  - CSMA, CSMA/CD, CSMA/CA

“Taking turns”

- Nodes take turns, nodes with more to send can take longer turns
- Polling from central site, token passing
- Bluetooth, FDDI, IBM Token Ring

Ethernet frame structure

Sending adapter encapsulates IP datagram (or other network layer protocol packet) in **Ethernet frames**

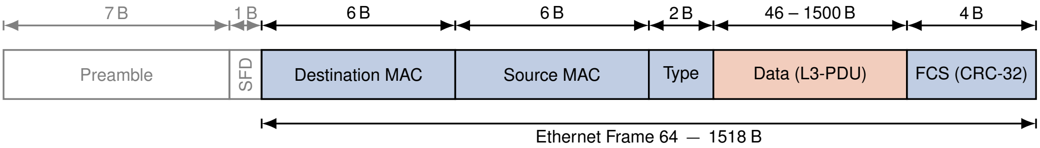


- Ethernet packet (physical layer):

IPG	Inter packet gap, minimum idle period between two packets
Preamble	Preamble (7 byte: `1010101010...`)
SFD	Start-of-frame delimiter (`10101011`)

Ethernet frame structure

Sending adapter encapsulates IP datagram (or other network layer protocol packet) in **Ethernet frames**



- Ethernet frame (data link layer):

Dst MAC	Destination Address
Src MAC	Source Address
Type/*Length*	Ethernet II frame format: Protocol type of payload (e.g. IP, ARP, ...) *Ethernet I and IEEE 802.3 frame format (rarely used today):* *Length of payload in byte*
Data	Data
PAD	Padding (if data length is less than 46 byte)
FCS	Frame Check Sequence: CRC-32

For comparison: IPv4 datagram [1]

Offset	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0 B	Version				IHL				TOS								Total Length															
4 B	Identification																Flags			Fragment Offset												
8 B	TTL								Protocol								Header Checksum															
12 B	Source Address																															
16 B	Destination Address																															
20 B	Options / Padding (optional)																															

Link Layer
MAC addresses

32 bit IPv4 address

- Network layer address
- used to get datagram to destination IP subnet

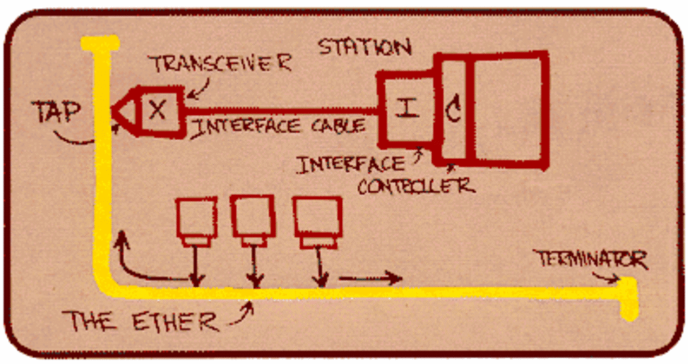
MAC / LAN / physical / Ethernet address

- Function: transmit frame from one interface to another physically-connected interface (same network)
- 48 bit MAC address (for most LANs)
  - burned in network adapter ROM or configurable in software

Link-Layer Protocols

Ethernet
Overview

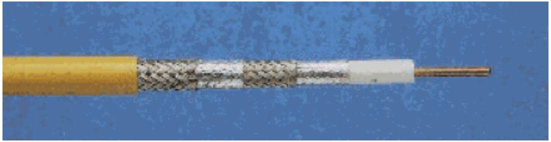
- Most common wired LAN technology
- Cheap network cards (NICs)
- First widely used LAN technology
- Simpler and cheaper than Token ring / ATM / MPLS
- Kept up with speed race: 10 Mbps - 400 Gbps

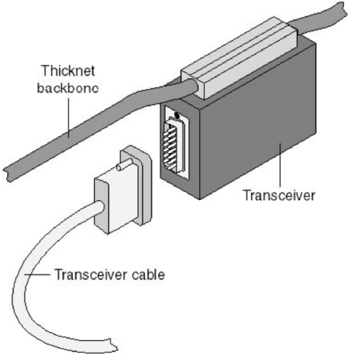


Metcalfe's Ethernet sketch (1976)

Ethernet
10Base5 - Thick Ethernet (IEEE 802.3, standardized 1983)

- Single bus system of thick coax cable (yellow)
- 10Base5: 10 Mbit/s
- Segments of 500 m, can be coupled with repeaters (max. 5 segments)
- Transceiver (transmitter & receiver) MAU (medium attachement unit) with carrier sensing function
- Transceiver cable max. 50 m





Ethernet
10Base2 - Thin Ethernet (IEEE 802.3a, standardized 1985)

- Single bus system of thinner coax cables (cheaper and more flexible)
- 10Base2: 10 Mbit/s
- Segments of max 185 m (max. 5 segments)
- Transceiver can be part of Ethernet adapter

Figure 1: T-piece

Figure 2: BNC terminator

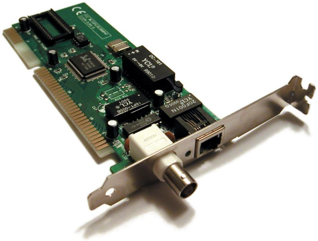


Figure 3: NIC with BNC connector

Ethernet

Bus vs. Star

Logical bus topology (10Base5, 10Base2):

- All nodes are part of a common collision domain
- Defect bus wire splits network in two parts

Star topology (newer standards):

- Active switch in center
- Each “spoke” runs a (separate) Ethernet protocol, therefore a defect wire disconnects only one host

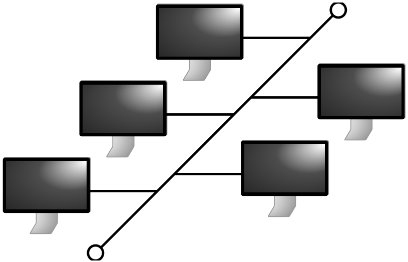


Figure 4: Bus topology

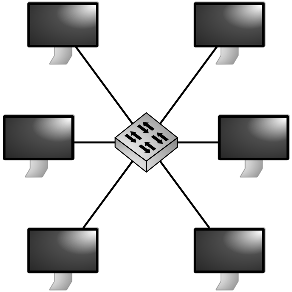


Figure 5: Star topology

10Base-T - Twisted Pair (IEEE 802.3i, standardized 1990)

- Uses star topology (hubs or switches) to connect devices
- CAT-3 or CAT-5 cables (uses two pairs of twisted wires)
- Reuses standardized connectors and wiring of telephone networks
- 10Base-T: 10 Mbit/s
- Segments of max 100 m (max. 5 segments)

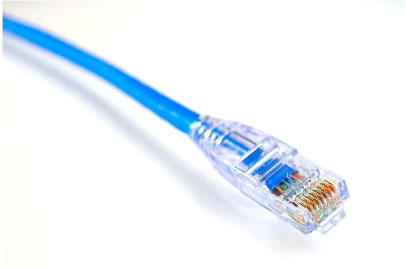


Figure 6: 8P8C connector (also know as RJ45)

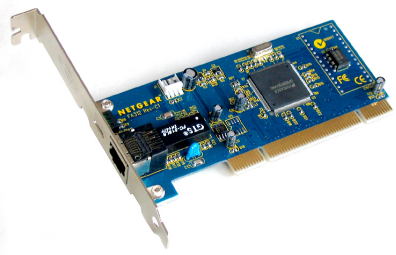


Figure 7: NIC with RJ45 connector

Ethernet
RJ45-based Ethernet Standards

100Base-TX - Fast Ethernet (IEEE 802.3u, standardized 1995)

- CAT-5 cables or better (uses two pairs of twisted wires)
- 100Base-TX: 100 Mbit/s

1000Base-T - Gigabit Ethernet (IEEE 802.3ab, standardized 1999)

- CAT-5 cables or better (uses four pairs of twisted wires)

10Gbase-T - 10 Gigabit Ethernet (IEEE 802.3an, standardized 2006)

- standardized in 2006
- CAT-6a cables or better

2.5Gbase-T / 5Gbase-T (IEEE 802.3bz, standardized 2016)

- works fine on most CAT-5 installations

Ethernet
RJ45-based Ethernet

Advantages

- robustness
- cheap, existing wiring

Disadvantages

- short cable lengths
- high energy consumption (for 10G)

NIC	Offload	Media	Idle Power (W) 3.3v	12v	Total
Intel(Base-T)	No	Base-T	6.0	15.2	21.2
Solarflare(Base-T)	No	Base-T	1.0	17.0	18.0
Broadcom(Fibre)	Yes	Fibre	5.9	7.2	13.1
Solarflare(Fibre)	No	Fibre	2.6	3.1	5.7

(a) 10G Ethernet [2]

NIC	Media	Throughput (Gbps) Theoretical	Actual	Active Power (W)
Intel 1G	Base-T	2	1.7	1.9
Broadcom Multiport(2x1G)	Base-T	4	3.3	7.0
Intel Multiport(2x1G)	Base-T	4	3.3	3.6
Intel Multiport(4x1G)	Base-T	8	5.7	12.5

(b) 1G Ethernet [2]

[2] R. Sohan, A. Rice, W. M. Andrew, et al., “Characterizing 10 gbps network interface energy consumption,” in IEEE Local Computer Network Conference, IEEE, 2010, pp. 268–271

Ethernet
Other Ethernet standards

Many different Ethernet standards

- Sharing a common MAC protocol and frame format
- Different bandwidths: 10M, 100M, 1G, 2.5G, 5G, 10G, 25G, 40G, 100G, 200G/400G (standardized in 2018)
- Different physical layer media, such as:
  - twisted pair (xBase-T)
  - twinaxial cabling (twinax)
  - unshielded twisted-pair (xBase-T1)
  - multimode optical fibre (short range)
  - singlemode optical fibre (long range)
  - backplane
  - chip-to-chip interfaces on NIC

Ethernet
Supporting different physical media

Pluggable transceiver module

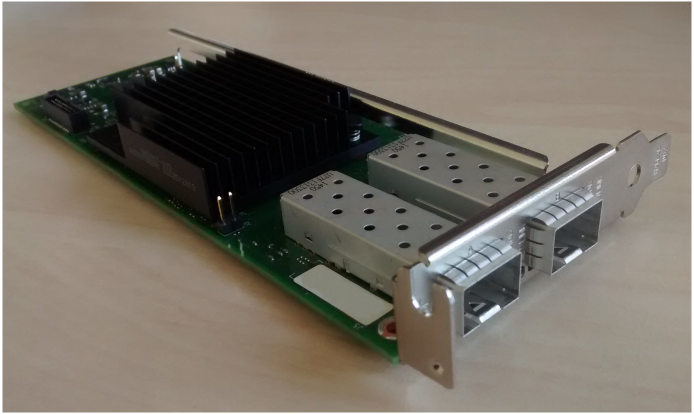


Figure 9: NIC with two slots for pluggable transceivers

Modern transceiver modules

- SFP (small form-factor pluggable) modules
- Most common standard for switchable transceivers
- Different generations (SFP for 1 GbE, SFP+ for 10 GbE, . . .)
- SFP modules are very common for professional equipment

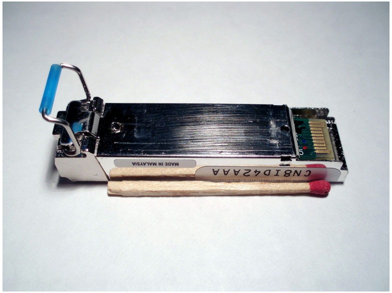


Figure 10: SFP module

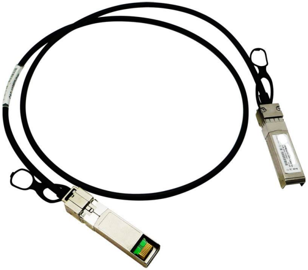


Figure 11: Direct-Attach Copper (DAC) twinaxial cable with integrated SFP modules (cheap, used for low range connections $\lesssim 15\,\mathrm{m}$)

Ethernet
Limitations of layer 2

Could Ethernet scale up to a very large (global) network?

Ethernet
Limitations of layer 2

Could Ethernet scale up to a very large (global) network?

Scalability problems:

- Flat addresses
- No hop count (so loops may lead to disaster)
- Missing additional protocols (such as ICMP)
- Perhaps missing features:
  - Fragmentation
  - Error messages
  - Congestion feedback

Link-Layer Protocols

Protocol mechanisms

Link Layer

Ethernet

MAC addresses

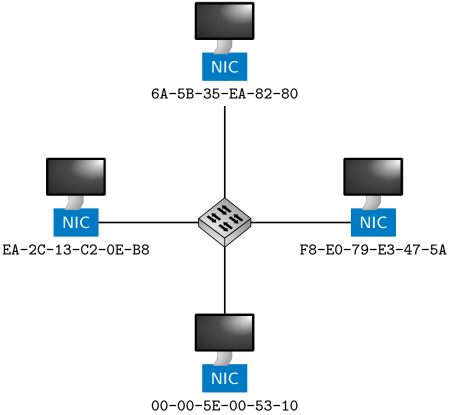
Layer 2 switching

Spanning tree

Bibliography

MAC addresses
Example Network

Each adapter on a LAN has a unique MAC address



MAC addresses
MAC address layout

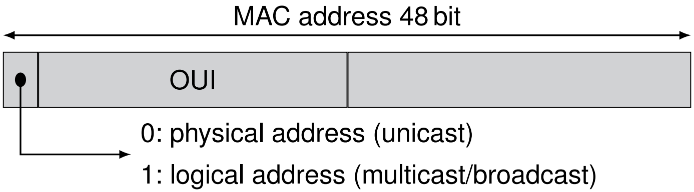
- Human-friendly notation for MAC addresses
  - six groups of two hex digits, separated by “-” or “:”, in transmission order, e.g., 0C-C4-11-6F-E3-98
- Multicast and broadcast
  - Broadcast address: FF-FF-FF-FF-FF-FF
  - Multicast address: least-significant bit of first byte has value “1”
- Organisation Unique Identifier (OUI): company id
  - manufacturer purchases portion of MAC address space from IEEE Registration Authority (assuring uniqueness)
  - OUI: First 3 byte of address in transmission order
  - OUI enforced: 2nd least significant bit of first byte has value “0”,
  - otherwise: locally administered MAC address
- Locally administered MAC addresses:
  - Similar to private address blocks on layer 3
  - E.g. used for VMs
- MAC address: flat address portability (+ implication on privacy)
  - can move LAN card from one LAN to another
- IP address: hierarchical address NOT portable
  - address depends on IP subnet to which node is attached

MAC addresses
Bit-reversed representation of MAC address

- Corresponds to convention of transmitting least-significant-bit of each byte first in serial data communications (transmission of LAN addresses over the wire)
- Also known as “canonical form”, “LSB format” or “Ethernet format“ (LSB: Least Significant Bit):
  - First bit of each byte on the wire maps to least significant (i.e., right-most) bit of each byte in memory (cf. RFC 2469)
- Token Ring (IEEE 802.5) and FDDI (IEEE 802.6) do **not** use canonical form, but instead: most-significant bit first

MAC addresses
MAC addressing modes

- General address types (L2 and L3): Unicast, Multicast, Broadcast, Anycast
- Terminology to distinguish destination MAC addresses
  - Physical addresses: identify specific MAC adapters
  - Logical addresses: identify logical group of MAC destinations



- LAN broadcast address: FF-FF-FF-FF-FF-FF
- Transmission of multicast frames
  - sender transmits frame with multicast destination address
- Reception of multicast frames
  - NICs can be configured to capture frames whose destination address is:
    - their unicast address, **or**
    - one of a set of multicast addresses

MAC addresses
Addresses and naming

Addresses are defined across three layers

1./2. Physical / link level
   - Medium Access Control (MAC)

3. Network/IP level
   - IP addresses
     $\leftrightarrow$ mapping to domain names

4. Transport/application level
   - Ports
     $\leftrightarrow$ mapping to services
     - Standardized, well-known ports
     - Dynamic mapping

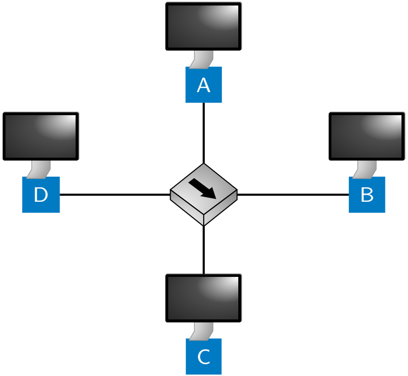
Link-Layer Protocols

Layer 2 switching

Hub

Physical-layer ("dumb") repeaters:

- Bits arriving on one link go out on all other links at same rate
- Frames from all nodes connected to hub can collide with each other
- No frame buffering
- No collision detection at hub: host NICs detect collisions

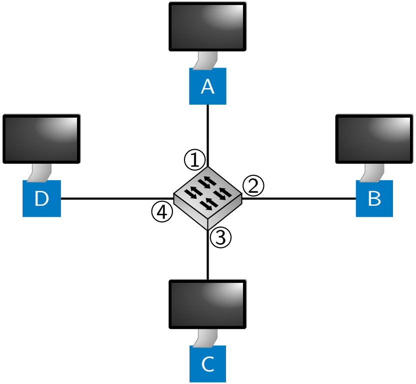


Layer 2 switching
Switch

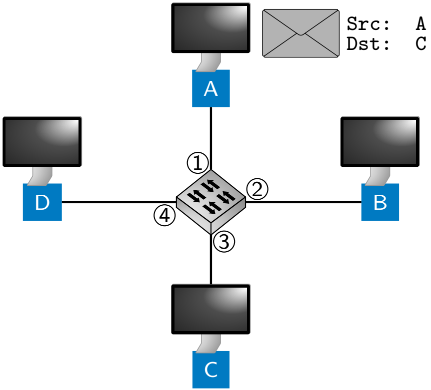
- Link-layer devices: smarter than hubs, take active role
  - Store & forward of Ethernet frames or cut-through-switching
  - Examine incoming frame's MAC address, **selectively** forward frames to one or more outgoing links
- Transparent
  - Hosts are unaware of presence of switches
- Plug-and-play, self-learning
  - Switches do not need to be configured

Layer 2 switching
Switch: simultanous transmission

- Hosts have dedicated, direct connection to switch
- Switches buffer packets
- Ethernet protocol used on each incoming link, but no collisions; full duplex
    - each link is its own collision domain
- Switching: A-to-C and B-to-D simultaneously, without collisions
    - not possible with dumb hub



Layer 2 switching
Switch: self-learning



Switches learn which hosts can be reached through which interfaces

- When a frame is received, a switch “learns” location of sender: incoming LAN segment
- Records sender/location pair in switch table
- Expiry time: soft state mechanism

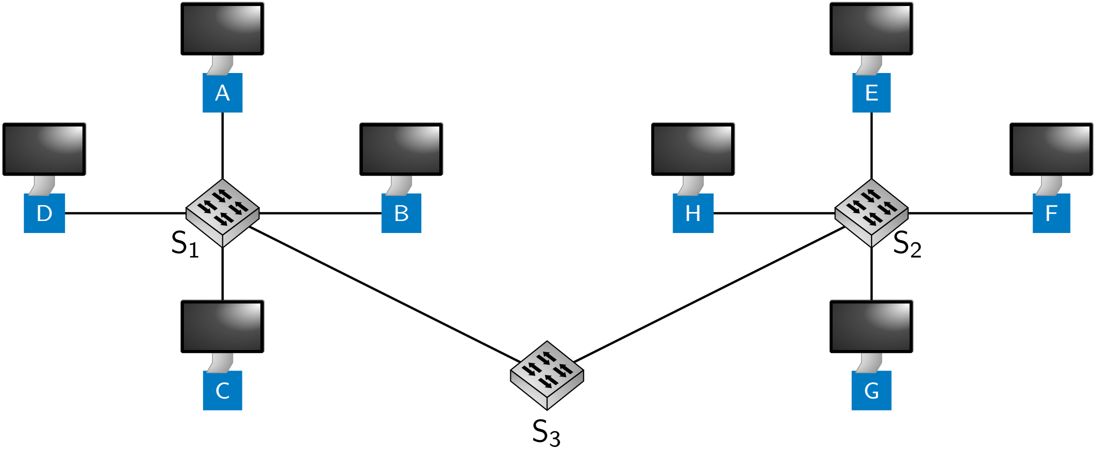
MAC address	interface	time
A	1	60

Table 1: Switch table *(after learning location of A)*

Layer 2 switching
Switch: frame filtering/forwarding

1. record link associated with sending host
2. index switch table using MAC destination address
3. if entry found for destination:
   if destination on segment from which frame arrived:
      drop the frame
   else:
      forward the frame on interface indicated
   else:
   flood *(forward on all interfaces except the interface on which frame arrived)*

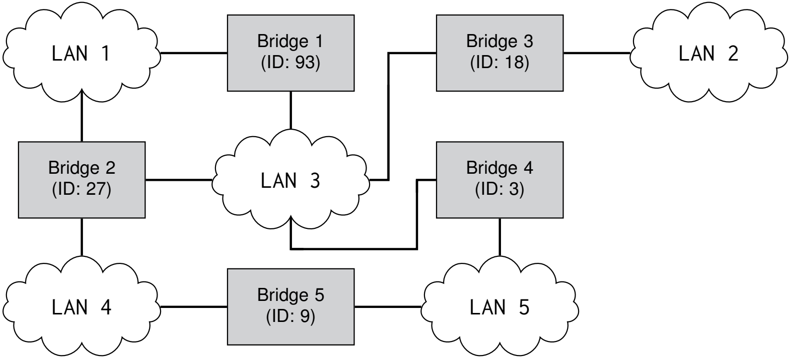
Interconnecting switches



Q: Sending from A to G - how does S_1 know to forward frame destined to G via S_3 and S_2?

A: Self-learning! (works exactly the same as in single-switch case!)

Spanning tree
Preventing loops



Spanning tree protocol

- Bridges gossip among themselves
- Compute loop-free subset
- Forward data on the spanning tree
- Other links are backups

Spanning tree
Spanning Tree Protocol

- Spanning Tree Protocol (STP): standardized as IEEE 802.1D
- Algorithm by Radia Perlman
- Algorithm:
  - Uses bridge_ID (concatenation of 16 bit bridge_priority and MAC_addr)
  - Step 1: select root bridge, i.e. bridge with lowest bridge_ID
  - Step 2: determine least cost paths to root bridge
    - each bridge determines cost of each possible path to root
    - each bridge picks least-cost path
    - port connecting to that path becomes root port (RP)
    - bridges on network segment determine bridge port with least-cost-path to root, i.e. the designated port (DP)
  - Step 3: disable all other root paths
- Bridge Protocol Data Units (BPDUs) are sent regularly (default: 2 s) to STP multicast address

Spanning tree
Spanning Tree Protocol

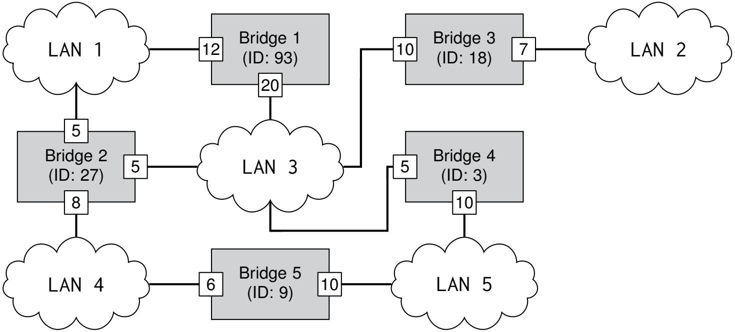
Bridge Protocol Data Units (BPDUs)

- Configuration BPDUs transmit bridge_IDs and root path costs
- Topology Change Notification (TCN) BPDU announce changes in network topology
- Topology Change Notification Acknowledgment (TCA)

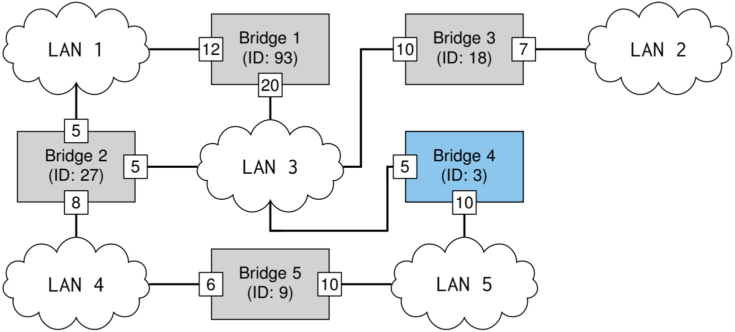
STP switch port states

- Blocking
- Listening
- Learning
- Forwarding
- Disabled

Spanning Tree Protocol

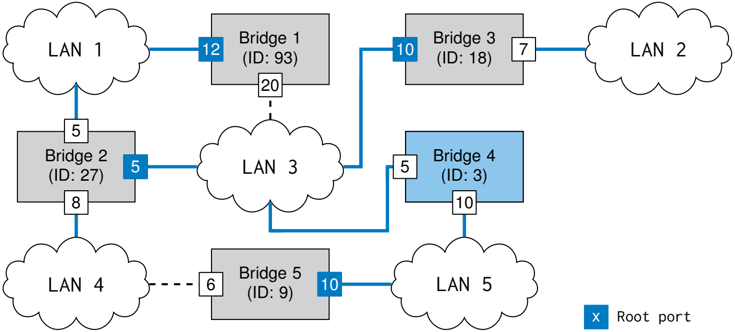


- Select root bridge



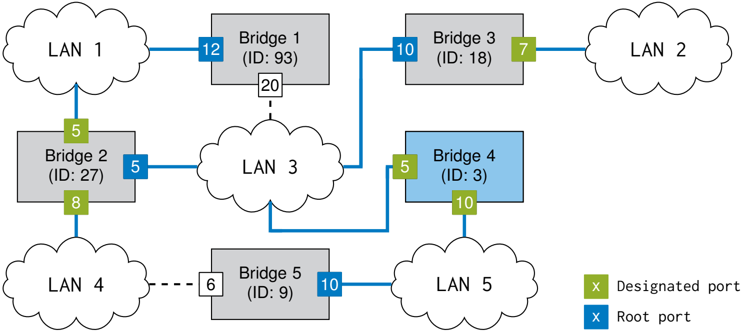
- Find shortest paths to root bridge

Spanning tree
Spanning Tree Protocol



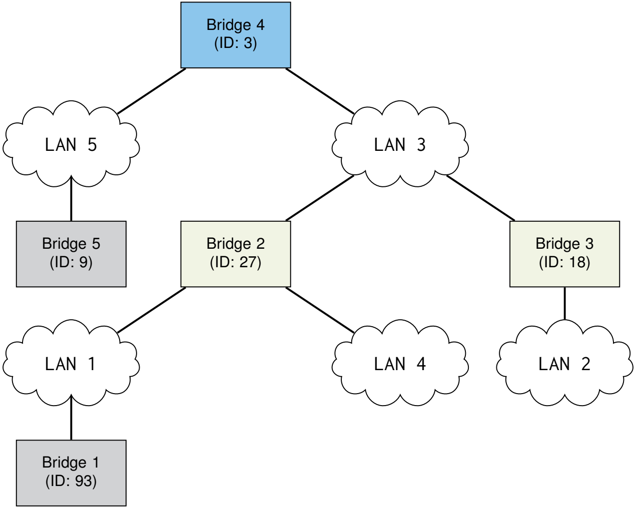
Bridge	Costs of paths to root bridge
B3	10 (via LAN 3)
B1	17 = 12 + 5 (via LAN 1 & LAN 3)
	20 (via LAN 3)
	30 = 12 + 8 + 10 (via LAN 1 & LAN 4 & LAN 5)
B2	5 (via LAN 3)
	18 = 8 + 10 (via LAN 4 & LAN 5)
	25 = 5 + 20 (via LAN 1 & LAN 3)
B5	10 (via LAN 5)
	11 = 6 + 5 (via LAN 4 & LAN 3)
	31 = 6 + 5 + 20 (via LAN 4 & LAN 1 & LAN 3)

Spanning tree
Spanning Tree Protocol



- Designated port: provides connectivity for LAN
  - e.g., Bridge 2 becomes designated bridge for LAN 1 and LAN 4

Spanning tree
Resulting spanning tree



Spanning tree
Acknowledgements

- Jim Kurose, University of Massachusetts, Amherst
- Keith Ross, Polytechnic Institute of NYC
- Olivier Bonaventure, University of Liege
- Srinivasan Keshav, University of Waterloo

[1] DARPA, Internet Protocol, https://tools.ietf.org/html/rfc791, 1981.

[2] R. Sohan, A. Rice, W. M. Andrew, and K. Mansley, “Characterizing 10 gbps network interface energy consumption,” in IEEE Local Computer Network Conference, IEEE, 2010, pp. 268–271.