Communication Networks

Prof. Laurent Vanbever

Communication Networks

Spring 2022

Laurent Vanbever

nsg.ee.ethz.ch

ETH Zürich (D-ITET)

14 March 2022

Materials inspired from Scott Shenker & Jennifer Rexford

Last week on

Communication Networks

Communication Networks

Part 2: Concepts

routing

reliable delivery

How do you guide IP packets from a source to destination?

Forwarding *vs* Routing

summary

	forwarding	routing
goal	directing packet to an outgoing link	computing the paths packets will follow
scope	local	network-wide
implem.	hardware usually	software usually
timescale	nanoseconds	milliseconds (hopefully)

---

sufficient and necessary condition

Theorem: a global forwarding state is valid **if and only if**

- there are no dead ends
  no outgoing port defined in the table
- there are no loops
  packets going around the same set of nodes

Producing valid routing state is harder

but doable

prevent dead ends
easy

prevent loops
hard

This is the question you should focus on

Essentially,

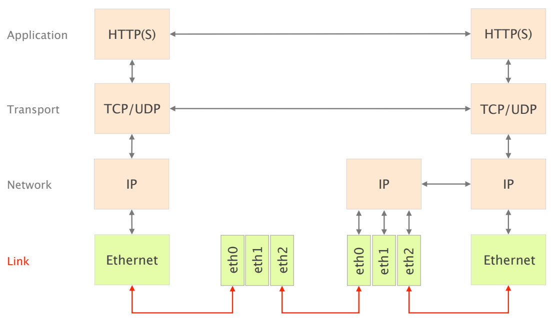
there are **three ways to compute** valid routing state

	Intuition	Example
#1	Use tree-like topologies	Spanning-tree
#2	Rely on a global network view	Link-State SDN
#3	Rely on distributed computation	Distance-Vector BGP

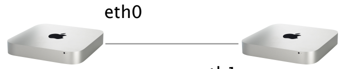
This week on
Communication Networks

This week we'll start speaking about
How the Internet actually works

We'll do that layer-by-layer, bottom-up,
starting with the Link layer



How do local computers communicate?



Communication Networks
Part 2: The Link Layer

#1	What is a link?
#2	How do we identify link adapters?
#3	How do we share a network medium?
#4	What is Ethernet?
#5	How do we interconnect segments at the link layer?

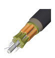
Communication Networks
Part 2: The Link Layer

#1	What is a link?
	How do we identify link adapters?
	How do we share a network medium?
	What is Ethernet?
	How do we interconnect segments at the link layer?

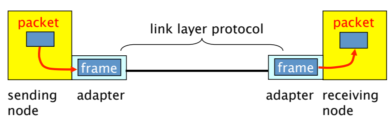
Link — Communication medium and Network adapter

 Wifi 

 Ethernet 

 Fiber 

Network adapters communicate
together through the medium



Network adapters communicate together through the medium

sender

encapsulate packets in a frame

add error checking bits, flow control, …

receiver

look for errors, flow control, …

extract packet and passes it to the network layer

The Link Layer provides a best-effort delivery service to the Network layer

L3	Network	global best-effort delivery
L2	Link	local best-effort delivery
L1	Physical	physical transfer of bits

The Link Layer provides a best-effort delivery service to the Network layer, composed of 5 sub-services

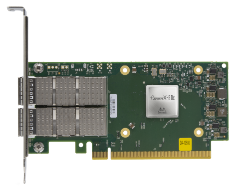
encoding	represents the 0s and the 1s
framing	encapsulate packet into a frame adding header and trailer
error detection	detects errors with checksum
error correction	optionally correct errors
flow control	pace sending and receiving node

As of March 2021, state-of-the-art Ethernet adapters clock at 200 Gbps

215 million pkt/sec

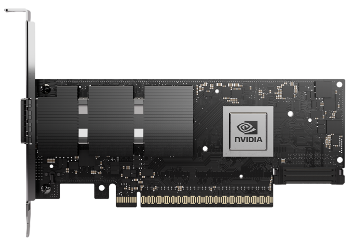
sub 0.8 usec latency

PCIe Gen 4.0



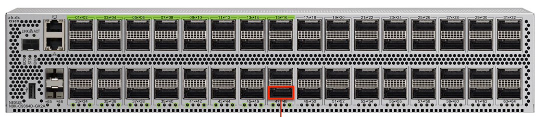
source: [Mellanox ConnectX-6]

400 Gbps adapters are around the corner



source: [NVIDIA ConnectX-7 (InfiniBand)]

400 Gbps Ethernet switches are already on the market



Cisco Nexus 9364D-GX2A

source: [cisco]

64x400 GbE ports (QSFP-DD)

25.6 Tbps backplane capacity

Communication Networks

Part 2: The Link Layer

What is a link?

#2 How do we identify link adapters?

How do we share a network medium?

What is Ethernet?

How do we interconnect segments at the link layer?

Medium Access Control addresses

MAC addresses…

MAC addresses…

identify the sender & receiver adapters
used on a link

are uniquely assigned
hard-coded into the adapter when built

use a flat space of 48 bits
allocated hierarchically

MAC addresses are hierarchically allocated

34:36:3b:d2:8a:86

The first 24 bits blocks are assigned
to network adapter vendor by the IEEE

34:36:3b:d2:8a:86

Apple, Inc.
1 Infinite Loop
Cupertino CA 95014
US

see http://standards-oui.ieee.org/oui/oui.txt

The second 24 bits block is assigned
by the vendor to each network adapter

34:36:3b:d2:8a:86

assigned by Apple
to my adapter

The address with all bits set to 1
identifies the broadcast address

ff:ff:ff:ff:ff:ff

enables to send a frame to
all adapters on the link

By default, adapters only decapsulates frames
addressed to the local MAC or the broadcast address

The promiscuous mode enables to decapsulate
everything, independently of the destination MAC

Why don't we simply use IP addresses?

Links can support any protocol (not just IP)
different addresses on different kind of links

Adapters may move to different locations
cannot assign static IP address, it has to change

Adapters must be identified during bootstrap
need to talk to an adapter to give it an IP address

Adapters must be identified during bootstrap
need to talk to an adapter to give it an IP address

We need to solve two problems when we bootstrap an adapter

Who am I?
MAC-to-IP binding

How do I acquire an IP address?

Who are you?
IP-to-MAC binding

Given an IP address reachable on a link,
How do I find out what MAC to use?

Who am I?
MAC-to-IP binding

How do I acquire an IP address?

Dynamic Host Configuration Protocol

Who are you?
IP-to-MAC binding

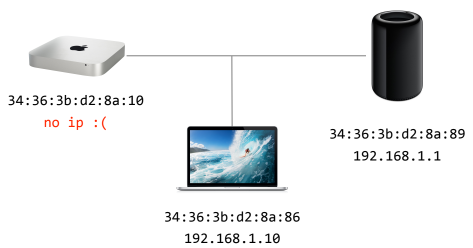
Given an IP address reachable on a link,
How do I find out what MAC to use?

Address Resolution Protocol

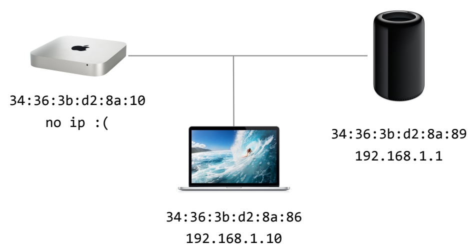
Network adapters traditionally acquire an IP address using the Dynamic Host Configuration Protocol (DHCP)

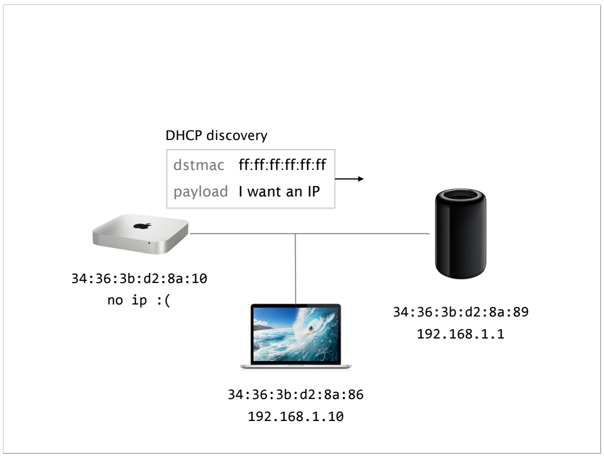
Every connected device needs an IP address…

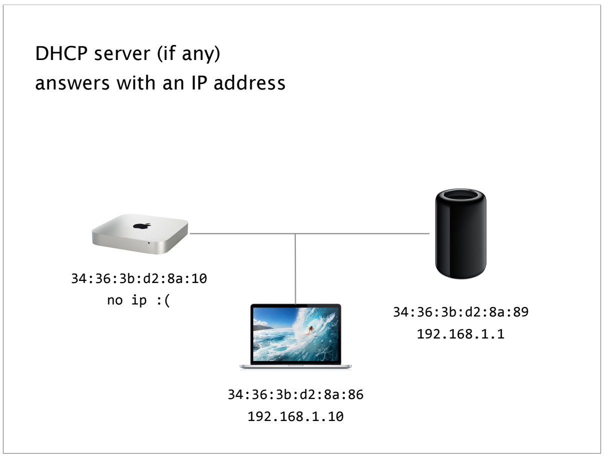
Newark Airport… source: http://i.imgur.com/m1SQa6W.jpg

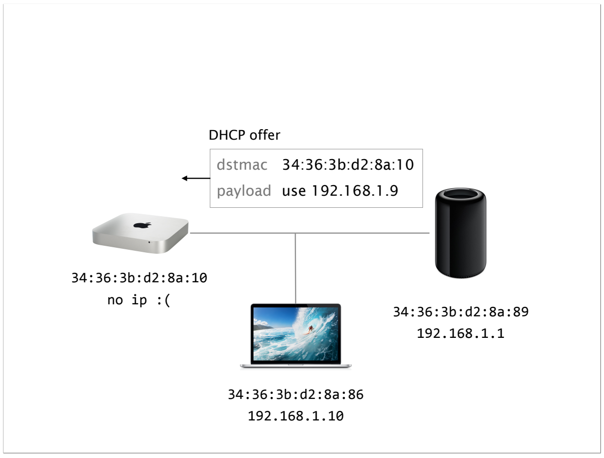


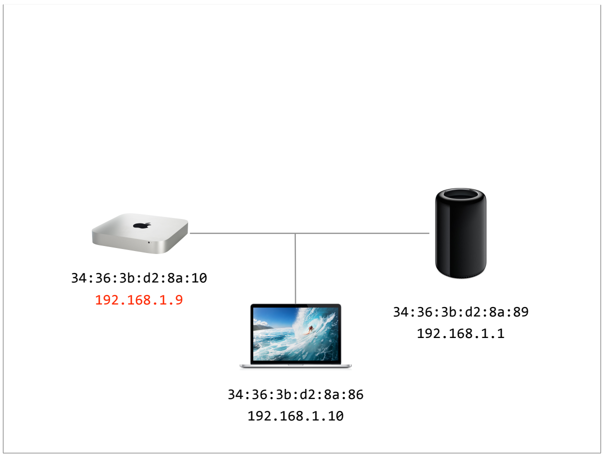
Host sends an "IP request" to everyone on the link using the broadcast address



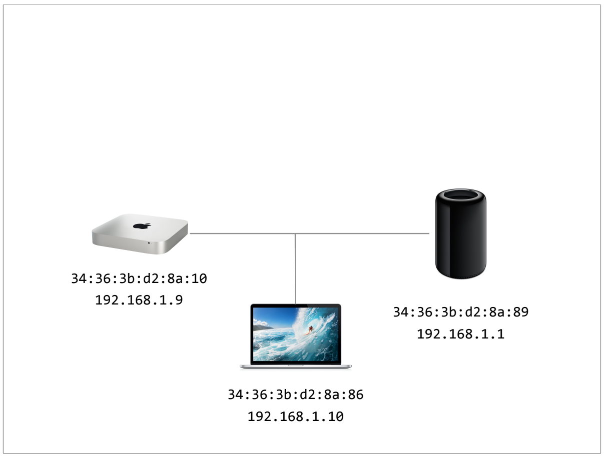


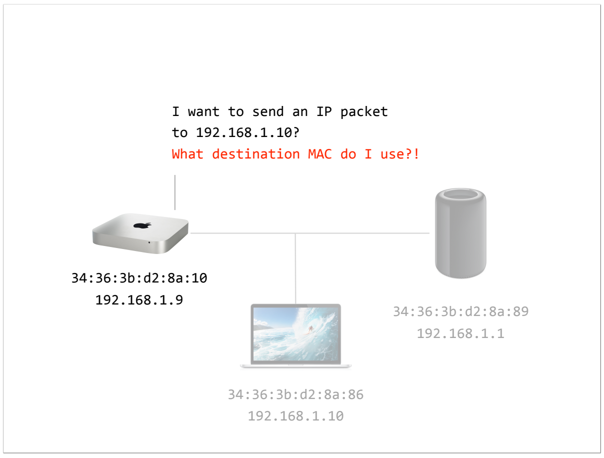


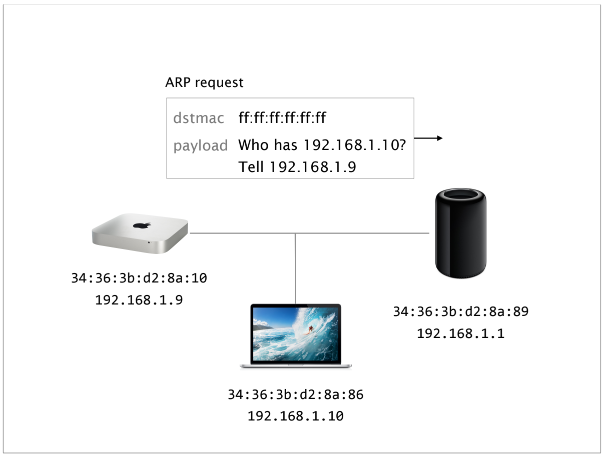


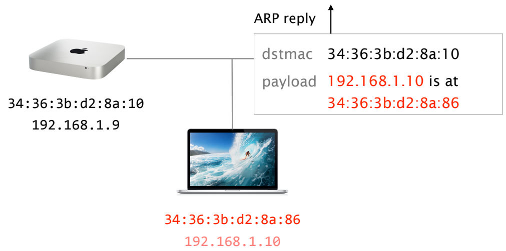


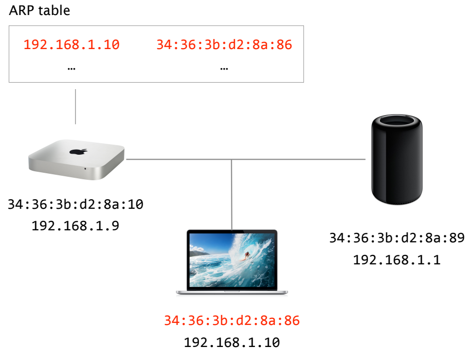
The Address Resolution Protocol (ARP) enables a host to discover the MAC associated to an IP











Communication Networks

Part 2: The Link Layer

What is a link?

How do we identify link adapters?

#3 How do we share a network medium?

What is Ethernet?

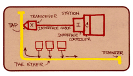
How do we interconnect segments at the link layer?

Some medium are multi-access:
>1 host can communicate at the same time

Some medium are multi-access:
>1 host can communicate at the same time

Wireless networks

Satellite networks



original Ethernet networks

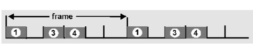
Cellular networks

Some medium are multi-access:
>1 host can communicate at the same time

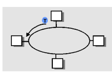
Problem	Solution
collisions lead to garbled data	distributed algorithm for sharing the channel
	When can each node transmit?

Essentially, there are three techniques
to deal with Multiple Access Control (MAC)

Divide the channel into pieces
either in time or in frequency



Take turns
pass a token for the right to transmit



Random access
allow collisions, detect them and then recover

Communication Networks

Part 2: The Link Layer

What is a link?

How do we identify link adapters?

How do we share a network medium?

#4 What is Ethernet?

How do we interconnect segments at the link layer?

Ethernet…

was a broadcast technology
each packet was received by all attached hosts

is *the* dominant wired LAN technology
by far the most widely used

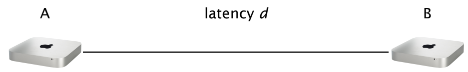
kept up with the speed race
from 10 Mbps to 400 Gbps (next: 800 Gbps and 1.6 Tbps!)

Ethernet offers an unreliable, connectionless service

unreliable	Receiving adapter does not acknowledge anything
	Packets passed to the network layer can have gaps which can be filled by the transport protocol (TCP)
connectionless	No handshaking between the send and receive adapter

"Traditional" Ethernet relies on CSMA/CD

CSMA/CD imposes limits on the network length



Suppose A sends a packet at time *t*

B sees an idle line just before t+d and sends a packet

Effect
B would detect a collision and sends a jamming signal
A can detect the collision only after t+2d

For this reason, Ethernet imposes a minimum packet size (512 bits)

This imposes restriction on the length of the network

Network length [m] $= \frac{\text{min\_frame\_size * speed of light}}{\text{2 * bandwidth}}$

= 768 meters for 100 Mbps

What about for 1 Gbps, 10 Gbps, 100 Gbps?

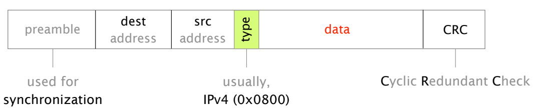
Modern Ethernet links interconnects *exactly* two hosts, in full-duplex, rendering collisions impossible!

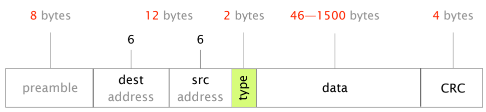
CSMA/CD is *only* needed for half-duplex communications
10 Gbps Ethernet does not even allow half-duplex anymore

This means the 64 bytes restriction is not strictly needed
but IEEE chose to keep it

Multiple Access Protocols are still important for Wireless
important concepts to know in practice

The Ethernet header is simple, composed of 6 fields only





Ethernet efficiency (payload/tot. frame size): ~97.5%
Maximum throughput for 100 Mbps: ~97.50 Mbps

Communication Networks

Part 2: The Link Layer

What is a link?

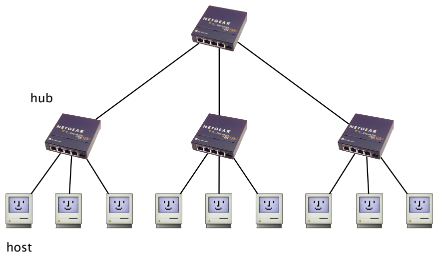
How do we identify link adapters?

How do we share a network medium?

What is Ethernet?

#5 How do we interconnect segments at the link layer?

Historically, people connected Ethernet segments together at the physical level using Ethernet hubs



Hubs work by repeating bits from one port to all the other ones

---

Hubs are now OBSOLETE

advantages	disadvantages
simple, cheap	inefficient, each bit is sent everywhere limits the aggregates throughput
	limited to one LAN technology can't interconnect different rates/formats
	limited number of nodes and distances cannot go beyond 2500m on Ethernet

---

Local Area Networks are now almost exclusively composed of Ethernet switches

---

Switches connect two or more LANs together at the Link layer, acting as L2 gateways

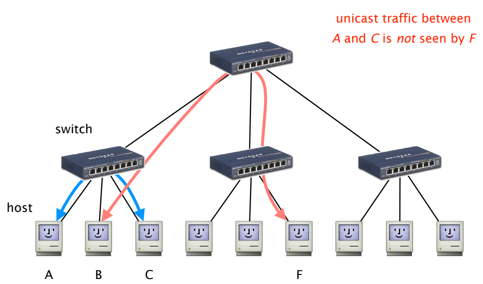
Switches are "store-and-forward" devices, they

- extract the destination MAC from the frame
- look up the MAC in a table (using exact match)
- forward the frame on the appropriate interface

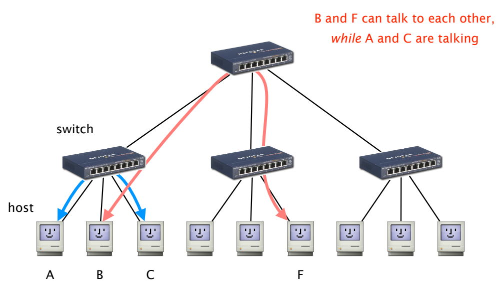
Switches are similar to IP routers, except that they operate one layer below

---

Unlike with hubs, switches enable each LAN segment to carry its own traffic



Unlike with hubs, switches supports concurrent communication



The advantages of switches are numerous

advantages

only forward frames where needed
avoids unnecessary load on segments

join segment using different technologies

improved privacy
host can just snoop traffic traversing their segment

wider-geographic span
separates segments allow longer distance

Switches are plug-and-play devices, they build their forwarding table on their own

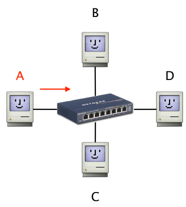
Switches are "store-and-forward" devices, they

- extract the destination MAC from the frame
- look up the MAC in a table (using exact match)
- forward the frame on the appropriate interface

Switches are plug-and-play devices, they build their forwarding table on their own

When a frame arrives:

- inspect the source MAC address
- associate the address with the port
- store the mapping in the switch table
- launch a timer to eventually forget the mapping



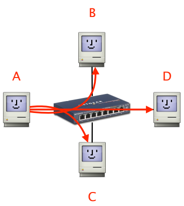
switch learns how to reach A

In cases of misses, switches simply floods the frames

When a frame arrives with an unknown destination

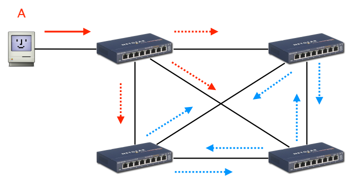
- forward the frame out of all interfaces
  except for the one where the frame arrived

Hopefully, this is an unlikely event



when in doubt, **shout!**

While flooding enables automatic discovery of hosts, it also creates problems when the networks has loops



Each frame leads to the creation of *at least two new frames!*
exponential increase, with no TTL to remove looping frames…

While loops create major problems, networks *need* redundancy for tolerating failures!

solution

Reduce the network
to one logical spanning tree

Upon failure,
automatically rebuild a spanning tree

In practice, switches run a *distributed* Spanning-Tree Protocol (STP)

Algorhyme

I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.

A tree that 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 from root are traced.
In the tree, these paths are placed.
A mesh is made by folks like me,
Then bridges find a spanning tree.

— *Radia Perlman*

A tree that 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 from root are traced.
In the tree, these paths are placed.
A mesh is made by folks like me,
Then bridges find a spanning tree.

Constructing a Spanning Tree in a nutshell

Switches…

elect a root switch
the one with the smallest identifier

determine if each interface is
on the shortest-path from the root
disable it if not

For this switches exchange Bridge Protocol Data Unit (BDPU) messages

Each switch *X* iteratively sends

BPDU (Y, d, X) to each neighboring switch

Y: the switch ID it considers as root

d: the # hops to reach it

initially

Each switch proposes itself as root
sends (X,0,X) on all its interfaces

Upon receiving (Y, d, X), checks if Y is a better root
if so, considers Y as the new root, flood updated message

Switches compute their distance to the root, for each port
simply add 1 to the distance received, if shorter, flood

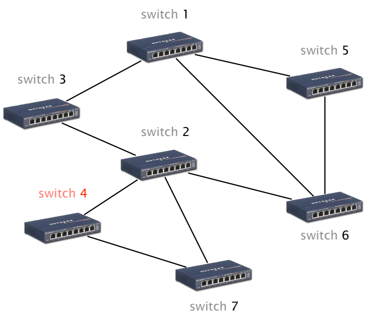
Switches disable interfaces not on shortest-path

tie-breaking

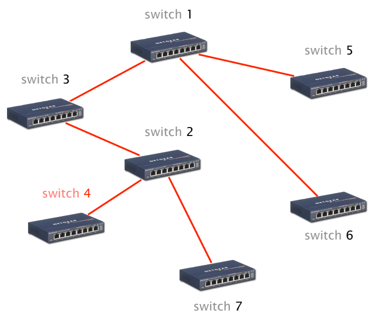
Upon receiving ≠ BPDUs from ≠ switches with = cost
Pick the BPDU with the lower switch sender ID

Upon receiving ≠ BPDUs from a neighboring switch
Pick the BPDU with the lowest port ID (e.g. port 2 < port 3)

Apply the algorithm starting with switch 4



Apply the algorithm starting with switch 4



To be robust,
STP must react to failures

Any switch, link or port can fail
including the root switch

Root switch continuously sends messages
announcing itself as the root (1,0,1), others forward it

Failures is detected through timeout (soft state)
if no word from root in *X*, times out and claims to be the root

The Local Area Networks we have considered so far define single broadcast domains

If one user broadcast a frame,
every other user receives it

As the network scales,
network operators like to segment their LANs

Why?

Improves security
smaller attack surface (visibility & injection)

Improves performance
limit the overhead of broadcast traffic (e.g. ARP)

Improves logistics
separates traffic by role (e.g. staff, students, visitors)

Organizational changes are too frequent to segment networks purely physically—rewiring is a major pain

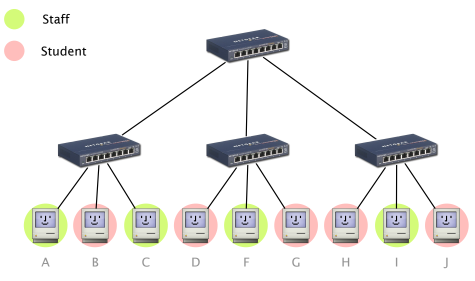
What about doing this in software though?

Enters “Virtual Local Area Networks” (VLANs)

Definition

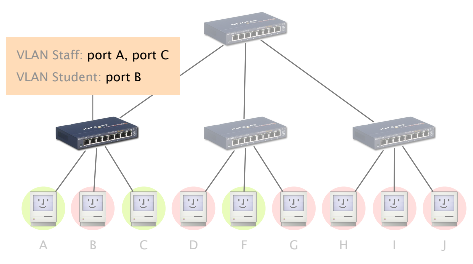
A VLAN logically identifies
a set of ports attached to
one (or more) Ethernet switches,
forming one broadcast domain

A VLAN identifies a set of ports attached to
one or more Ethernet switches

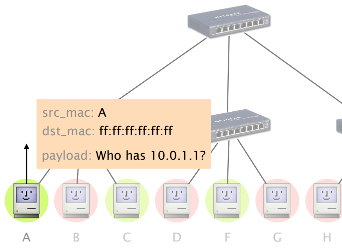


Switches need configuration tables telling them
which VLANs are accessible via which interfaces

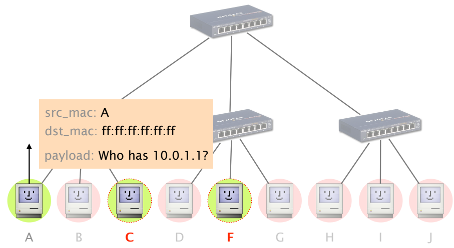
Switches need configuration tables telling them
which VLANs are accessible via which interfaces

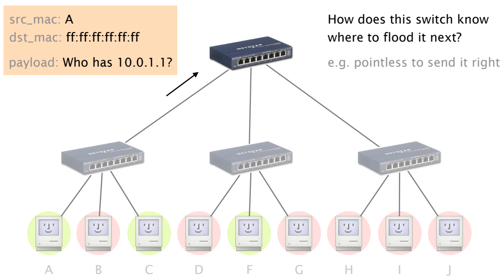


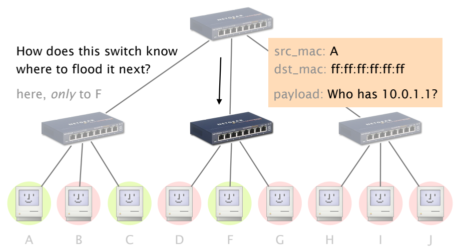
Consider that A sends a broadcast frame
say, an ARP request



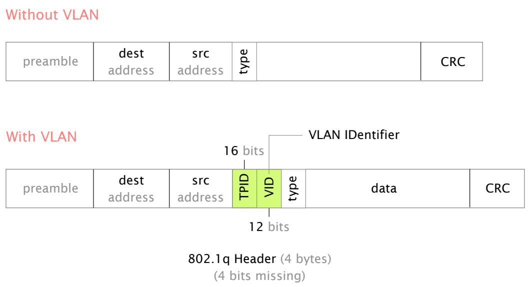
That frame should be received by all staff members:
i.e. C and F, and *only* them



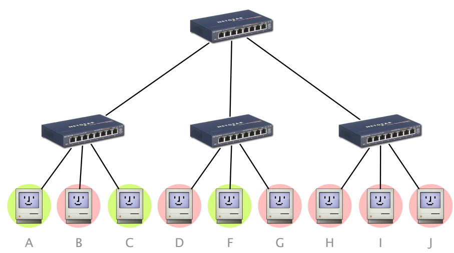




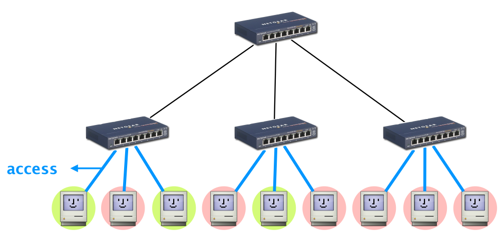
To identify VLAN, switches add new header
when forwarding traffic to another switch



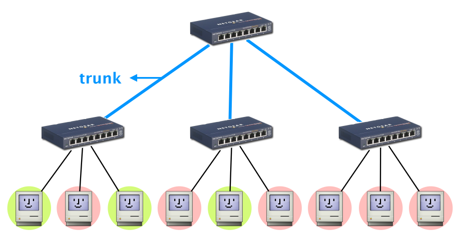
With VLANs, Ethernet links are divided in two sets:
access and trunks (inter switches) links



Access links belong to one VLAN
they do not carry 802.1q headers



Trunk links carry traffic for more than one VLAN
and as such carry 801.1q tagged frames



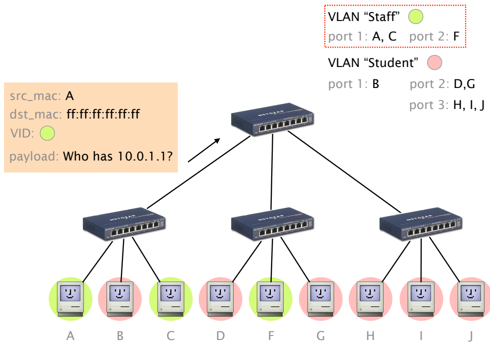
Each switch runs
one MAC learning algorithm for each VLAN

When a switch receives a frame with
an unknown or a broadcast destination,

it forwards it over all the ports that
belong to the same VLAN

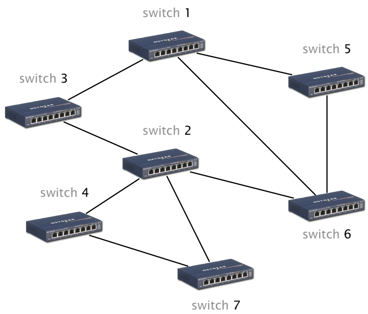
When a switch learns a source address on a port

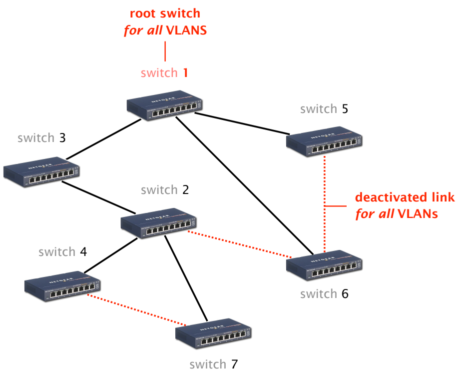
it associates it to the VLAN of this port and
only uses it when forwarding frames on this VLAN

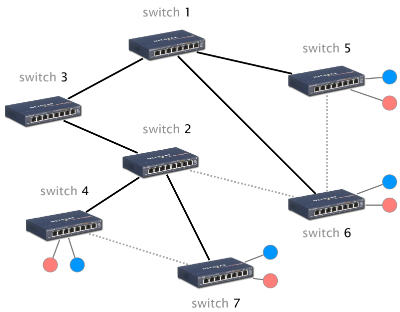


Switches can also compute per-VLAN spanning-tree
allowing a distinct SPT for each VLAN

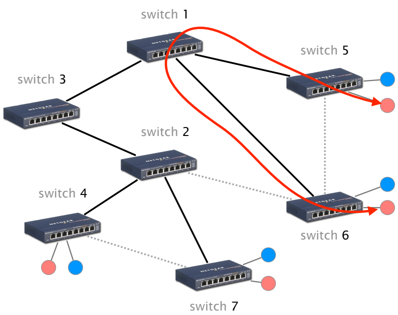
allow the operators to use more of their links

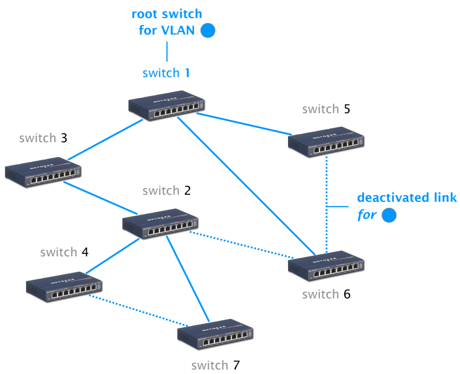


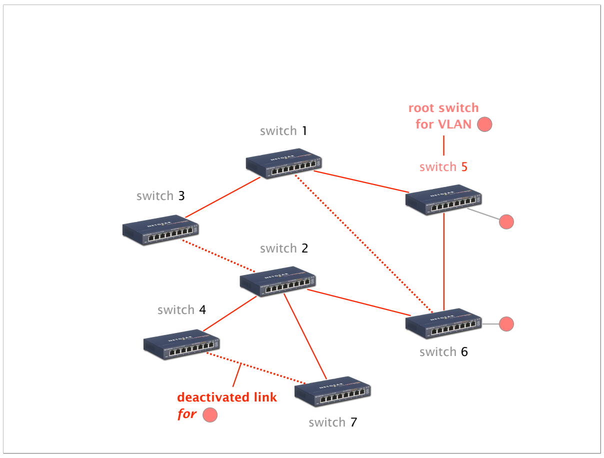


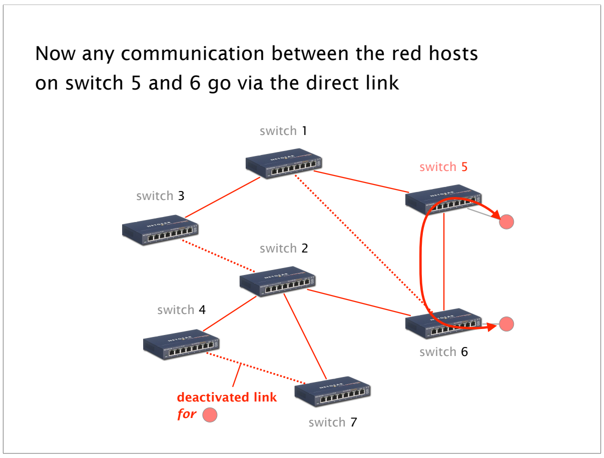


Any communication between the red hosts
on switch 5 and 6 need to go via switch 1…






root switch
for VLAN
switch 1
switch 5
switch 3
switch 2
switch 4
switch 6
deactivated link
for
switch 7


Now any communication between the red hosts
on switch 5 and 6 go via the direct link
switch 1
switch 5
switch 3
switch 2
switch 4
switch 6
deactivated link
for
switch 7


Communication Networks
Spring 2022
Laurent Vanbever
nsg.ee.ethz.ch
ETH Zürich (D-ITET)
14 March 2022