The link layer – Switched LANs and More

To do ...

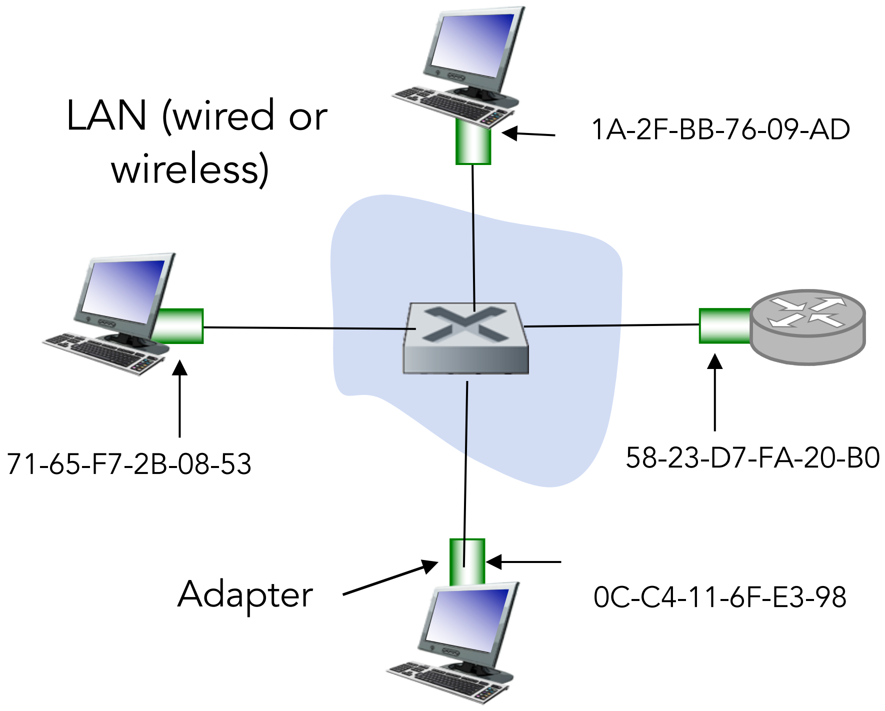
- LANs and virtual LANs
- MPLS
- Data center networks

Slides partially based on J. Kurose and K. Ross, "Computer Networking: A Top Down Approach", 7Ed., 2016

Northwestern

Forwarding within a switched local area network

- A set of machines within a small unit, like a lab or a department, are interconnected by a layer-2 switch
  - Working at the link layer, don't recognize network addresses (IPs)
  - They switch link-layer frames rather than datagrams
  - And don't use RIP or OSPF for routing
- How do they forward frames then?
  - Host and routers, or rather their adapters (network i/f) have a link-layer address
  - a.k.a LAN, physical or **MAC address**
  - Link-layer switches' interfaces don't have a MAC, they are there to connect hosts and routers transparently



MAC addresses

- Network layer address or IP address
  - 32 bits for IPv4
  - Hierarchical, if the machine moves networks the address changes
- MAC address
  - Flat structure, used "locally" to get frame from one interface to another physically-connected interface (same network, in an IP-addressing sense)
  - 6 bytes burned in NIC ROM, sometimes software settable
  - e.g.: 1A-23-F9-CD-0C-9B

hexadecimal (base 16) notation
(each "numeral" represents 4 bits)

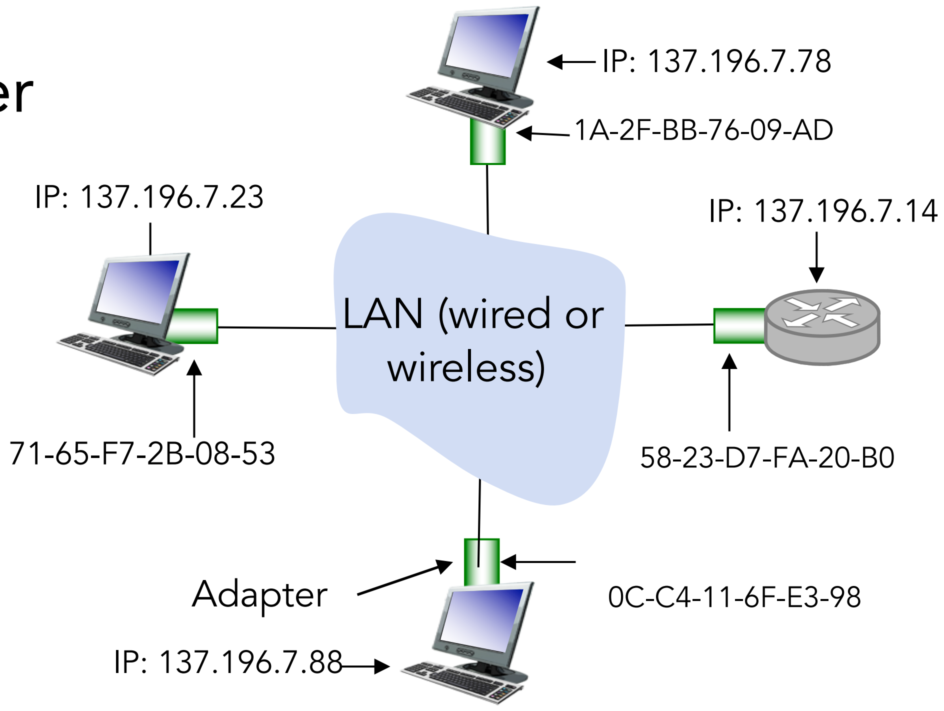
MAC address and ARP [RFC 826]

- To ensure no two adapters have the same, allocation administered by IEEE
  - Manufacturer buys portion (24b) of MAC address space
- Analogy
  - MAC address ~ Passport number – same one wherever you are
  - IP address ~ postal address – depends on IP subnet to which node is attached
- Since we have both, we need a protocol to translate between them – Address Resolution Protocol
  - Like DNS, which translates host names to IP addresses, but within the same LAN

ARP: address resolution protocol

- Each host and router has an ARP table in memory
  - IP/MAC address mappings for some LAN nodes:
    < IP address; MAC address; TTL>
  - TTL (Time To Live): time after which address mapping will be forgotten (typically 20 min)
  - Not necessarily includes every host and router on the subnet

IP	MAC	TTL
137.196.7.78	1A-2F-BB-76-09-AD	12:45:00
137.196.7.14	58-23-D7-FA-20-89	17:52:00



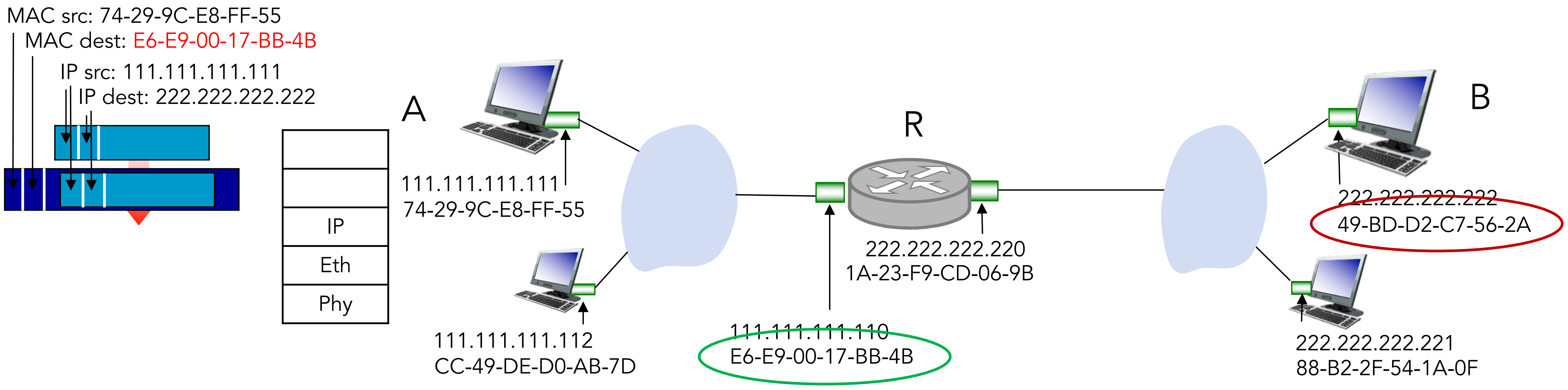
- *So, what if it doesn't have an entry for a given destination you need?*

ARP protocol (in the same LAN)

- A wants to send datagram to B
  - B's MAC address not in A's ARP table
  - A broadcasts ARP query packet, containing B's IP address
    - Destination MAC address = FF-FF-FF-FF-FF-FF
    - All nodes on LAN receive ARP query
  - B receives ARP packet, replies to A with its (B's) MAC address
    - Response frame sent to A's MAC address (unicast)
  - A caches IP-to-MAC address pair in its ARP table until info is too old
- ARP is "plug-and-play" – nodes create their ARP tables without intervention from net administrator
- ARP logically straddles between the link and network layers
  - Encapsulated within a frame, with link-layer addresses in it …

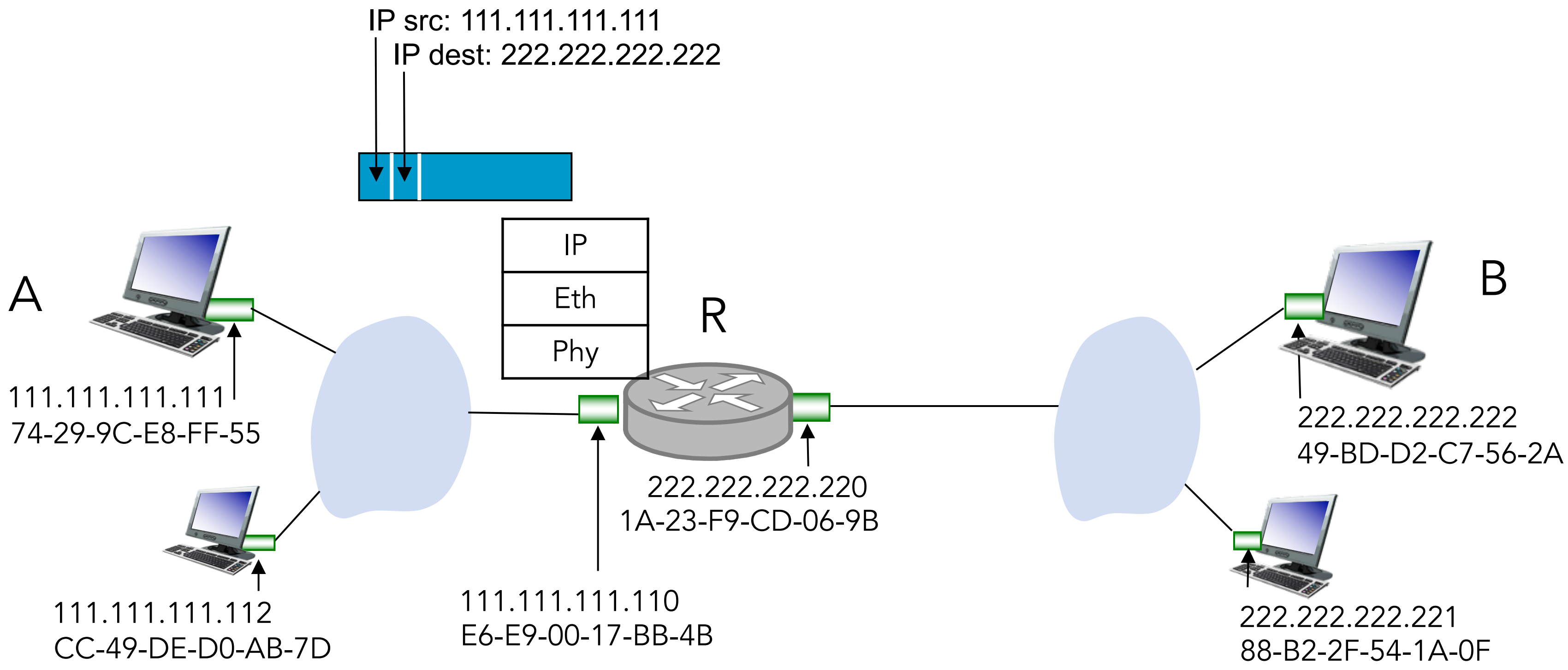
Addressing: routing to another LAN

- A creates IP datagram with IP source A, destination B
  - Needs a MAC but can't be that of B or the local devices won't pick it up and pass it to their network layer
- To get to B, datagram must get to router R
  - A creates link-layer frame with R's MAC address as destination address, frame contains A-to-B IP datagram



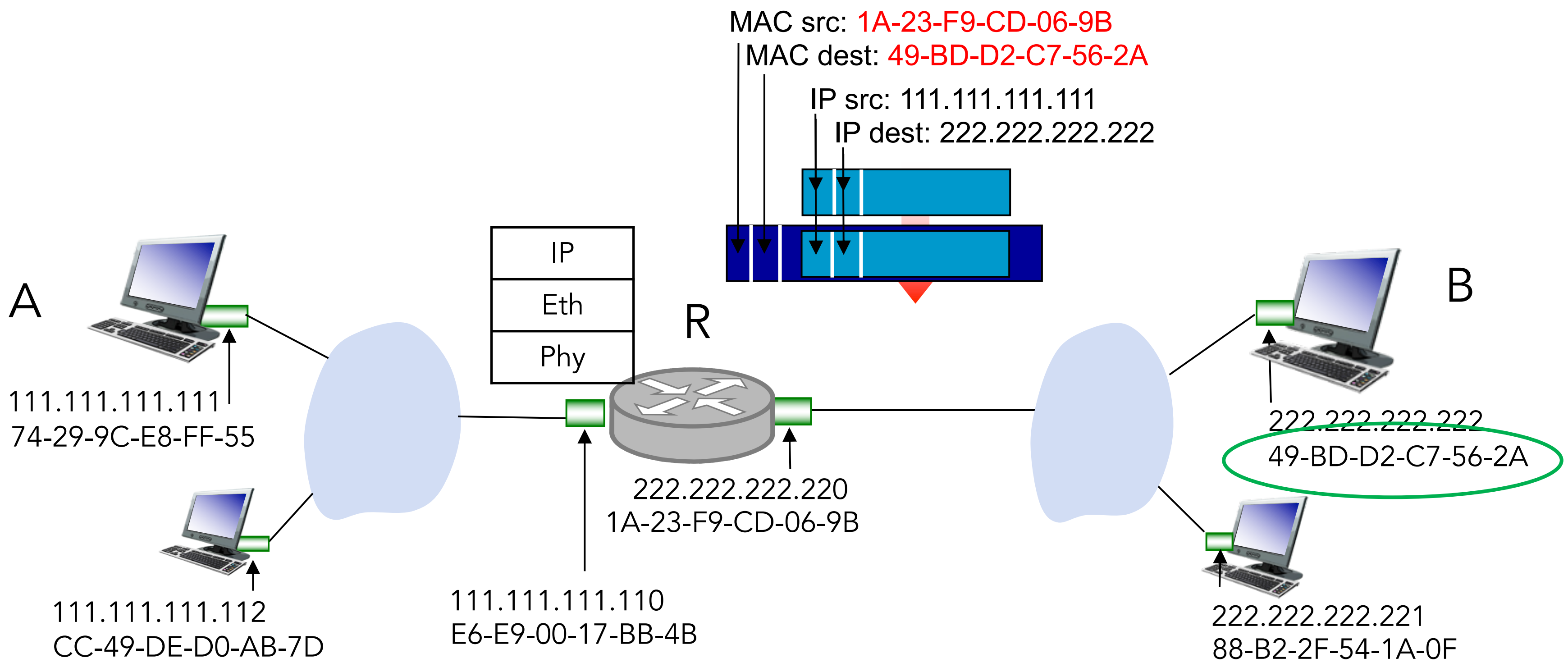
Addressing: routing to another LAN

- Frame sent from A to R, R gets it, and passes datagram up to IP



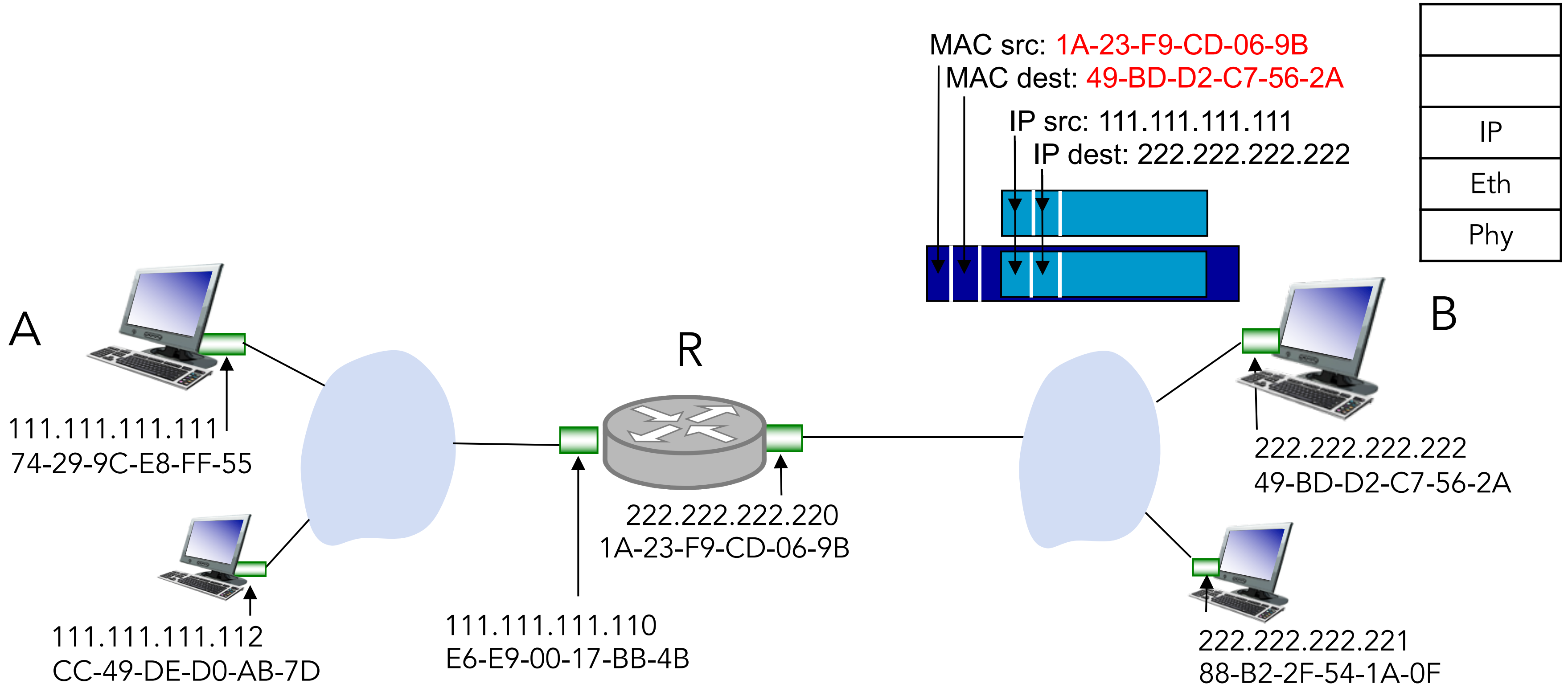
Addressing: routing to another LAN

- To determine the right interface to use, IP forwarding table
- Pass it to the interface, creating a link layer frame with B's MAC address as destination; the frame has A-to-B IP datagram



Addressing: routing to another LAN

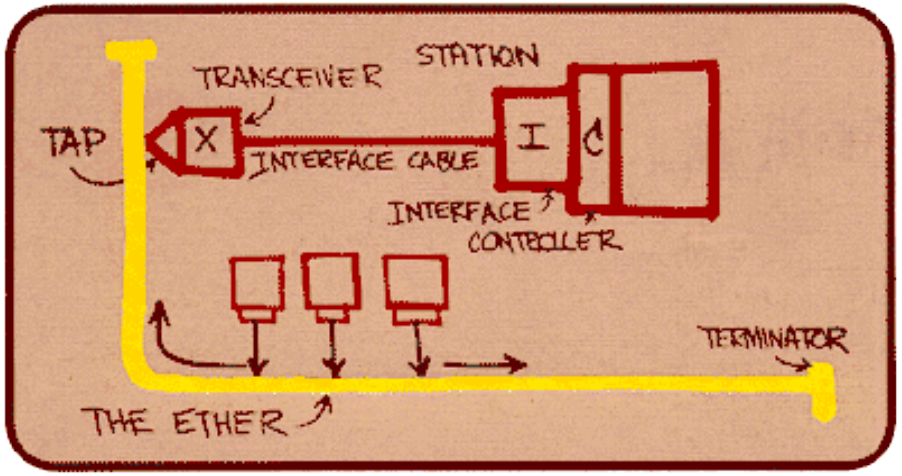
- When the frame arrives, the destination passes the datagram up



Ethernet

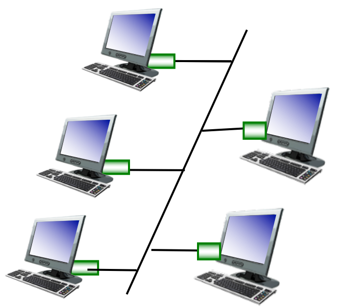
- The dominant wired LAN technology
  - Alternatives included token ring, FDDI and ATM
- Some reasons for domination
  - First widely used LAN technology
  - Simpler, cheap
  - Kept up with speed race (higher data rate was the most compelling reason to change to another LAN technology): 10 Mbps – 10 Gbps

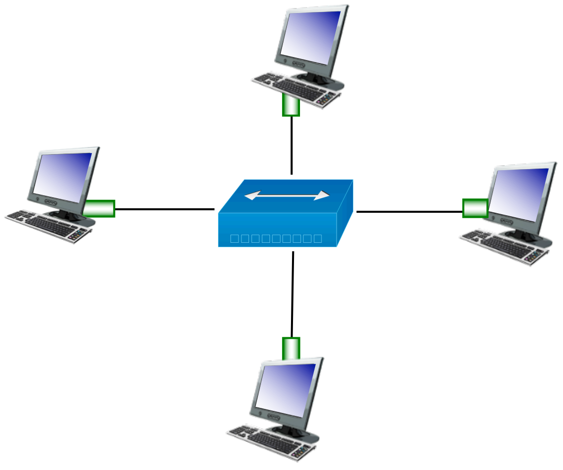
Metcalfe's Ethernet sketch
(1970s while at Xerox)



Ethernet: physical topology

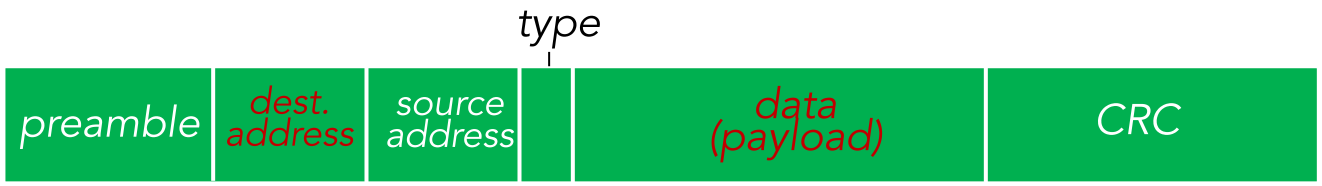
- Initially and through the mid 90s, a bus
  - All nodes in same collision domain (can collide with each other)
- Late 90s a move from bus to star with a hub in the center
  - Hub – physical-layer device, gets bit, recreates it, boots its energy and drops it on all interfaces
- *Both bus and hub-based start, broadcast*
- Early 2000s, a switch instead of hub
  - Active, store-and-forward switch in center
  - Each "spoke" runs a (separate) Ethernet protocol (nodes do not collide with each other)





Ethernet frame structure

- Sending adapter encapsulates IP datagram (or other network layer protocol packet) in Ethernet frame



- Data (46-1,500B) – Carries the datagram
  - If >1,500B – fragment, if <46B, has to be padded; all passed on to network layer which uses the length of IP datagram to remove stuffing
- Destination addresses (6B) – destination MAC address
  - If adapter receives frame with matching dst address, or with broadcast address FF-FF-FF-FF-FF-FF, it passes data in frame to network layer
  - Otherwise, adapter discards frame

Ethernet frame structure

- Source address (6B) – MAC address of the source
- CRC (4B) – Cyclic redundancy check at receiver (drop on error)
- Type (2B) – Indicates higher layer protocol for demultiplexing
  – Mostly IP but others possible, e.g., Novell IPX, AppleTalk
  – Also ARP has its own type (0806)
- Preamble (8B)
  – 7B ‘10101010’ followed by 1 byte ‘10101011’ (Start Frame Delimiter)
  – First 7B to wake up and synchronize receiver’s and sender clock rates, the SFD or its last 2 bits to announce important things are coming up

Ethernet: unreliable, connectionless

- Connectionless – No handshaking between sending and receiving NICs, just send what you have
- Unreliable – Receiving NIC doesn't send acks or nacks back if CRC check passes or fails
  - Data in dropped frames, gaps, are recovered only if initial sender uses higher layer rdt (e.g., TCP), otherwise dropped data lost
- Connectionless and unreliable → a simple and cheap Ethernet

- Not one but many different Ethernet standards
  - Common MAC protocol and frame format
  - Different speeds: 2, 10 and 100 Mbps, 1, 10 and 40 Gbps
  - Different physical layer media: fiber, cable, twisted-pair copper

10GBASE-T
10BASE-2
1000BASE-LX

Ethernet switches (late 1990s–today)

- Avoid broadcast and collision – relay msg to the correct port
- Switches are store-and-forward devices, like routers
  - More complex hardware – parallel packet processing and queueing
- If packet is addressed to an unknown address, broadcast to all
- Switch allows a subnet to grow very large, as packet flows are isolated from each other and can happen in parallel
- No special configuration is required on nodes or in switch
  - It's entirely "plug and play"
  - Switch automatically learns MACs of recent senders on each port
- If a port is connected to another switch, then it will relay traffic for many MAC addresses

Switches avoid collision entirely

- Switch ports have output queues (like a router), so switch will wait to send a packet until the port is free
  - Worst that can happen is that a packet is dropped due to a full queue
- Early Ethernet on bus or hub-based start topologies
  - A broadcast channel needing CSMA/CD to deal with collisions
- Today's Ethernet uses switched-base start topology
  - With a store-and-forward switch
  - Modern switches are full-duplex so a node and a switch can send frames to each other without interference
  - No collisions – no need for a MAC protocol
  - Switch isolates links from each other, so easy to mix legacy and new equipment

Switches and routers

Network admins sometime have a choice, e.g. connecting departments within a university

Ethernet switch

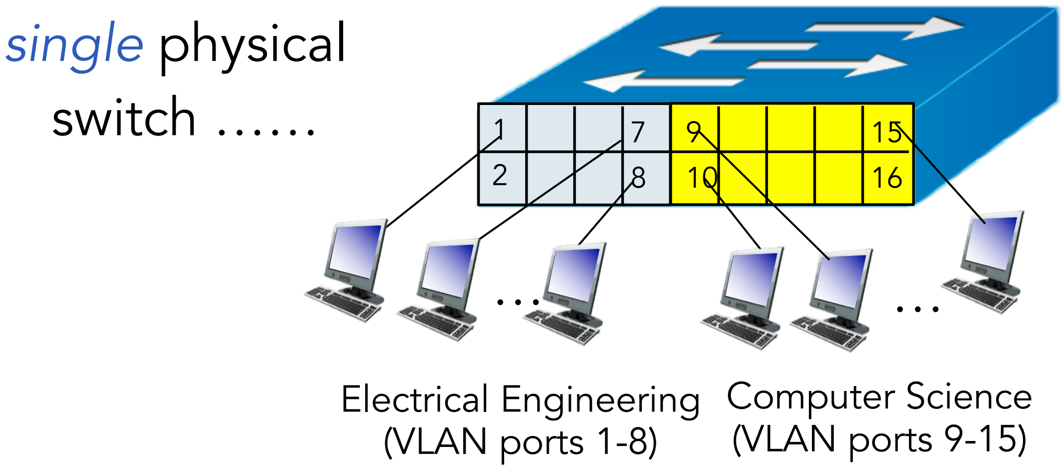
- Chooses an outbound link using packet's dst MAC address
- Forwarding rules are learned by inspecting traffic
- Redundant links are not allowed
- No configuration required, just "plug and play"
- ARP and DCHP (broadcast) traffic must be sent to all switch ports (on all connected switches)

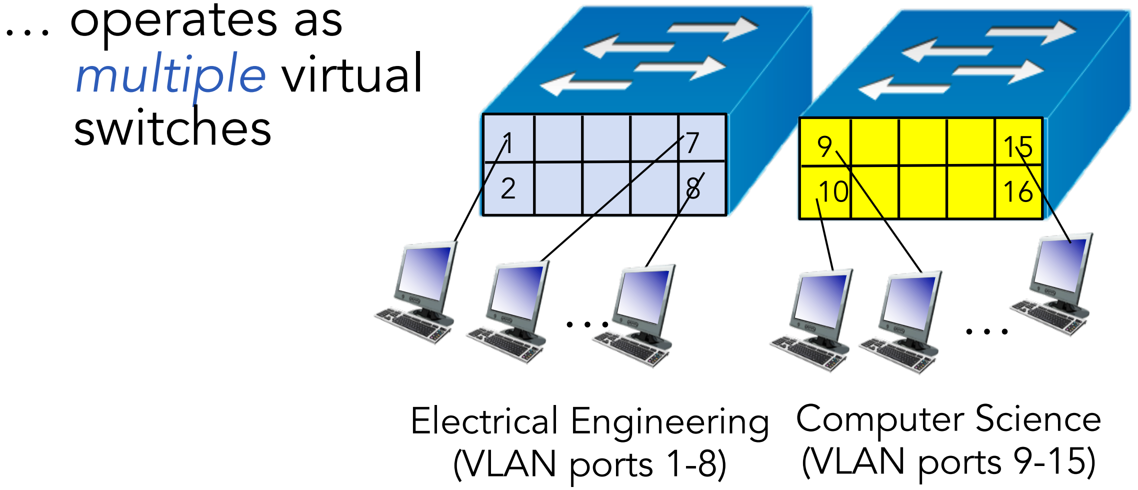
Router

- Chooses an outbound link using packet's dst IP address
- Forwarding rules are decided by IGP and BGP
- Routing algorithm chooses shortest among multiple paths
- Router must be configured to assign its IP addresses, IGP, etc
- Gives adms greater control over where traffic is sent (traffic eng)
- Isolates Ethernet broadcasts

VLANs (Virtual LANs)

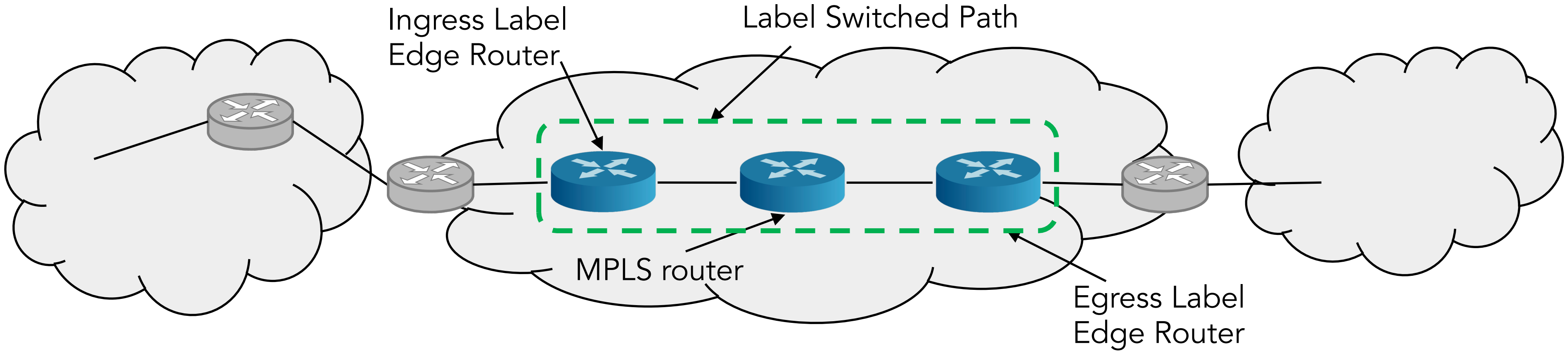
- A switch that supports VLANs can be divide it into multiple virtual switches
  - Let large switches to be flexibly configured
  - Often configured by port, can also assign MAC addresses to VLANs
- Typically used to isolate private subnets for security
- Traffic from one VLAN cannot flow into another VLAN





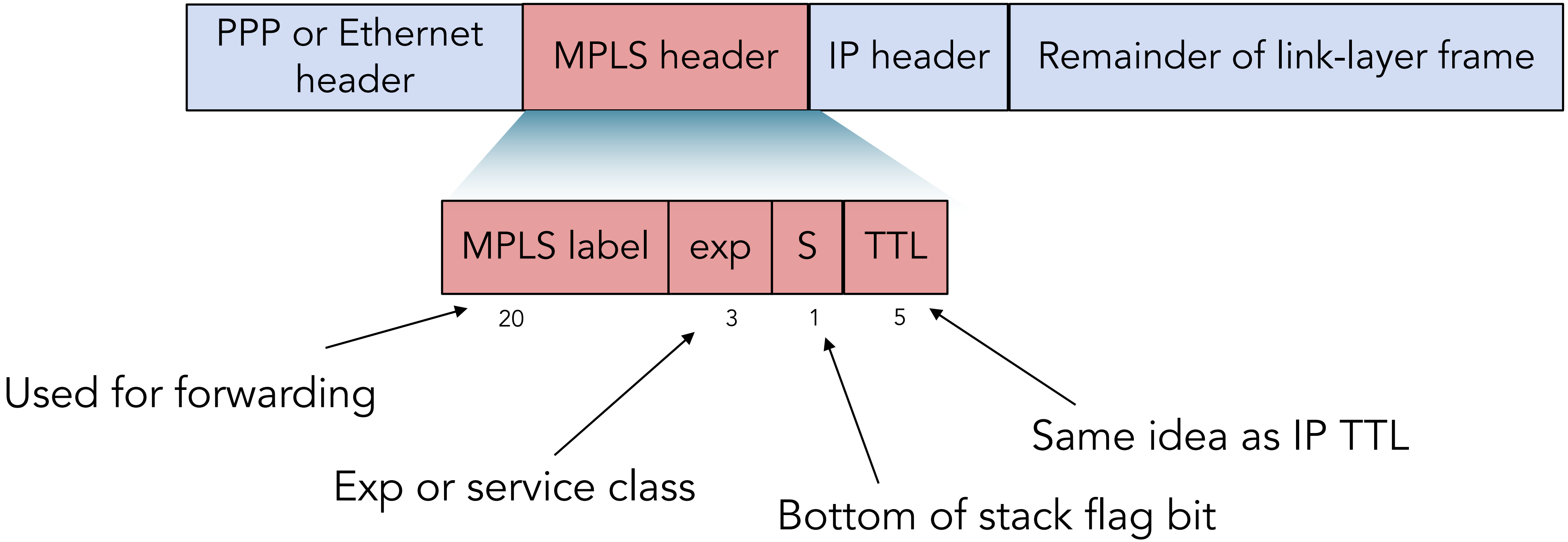
Multiprotocol label switching (MPLS)

- Initially proposed in mid 90s to improve IP router forwarding
  - Idea, use a fixed-length label rather than shortest prefix matching
- MPLS routers or label-switched routers
  - Exchanged labeled packets over Label Switched Paths (LSPs)
  - MPLS forwarding table distinct from IP forwarding tables
  - Ingress LER is a tunnel entry points, adds the label stack
  - Egress LER removes the label stack and does the classic IP lookup



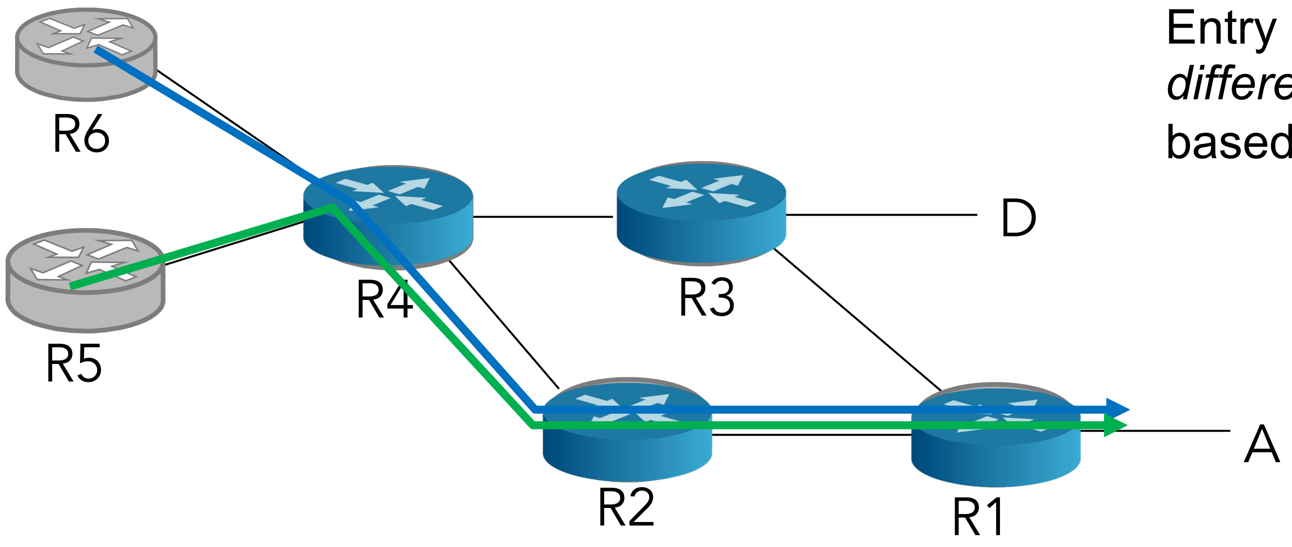
MPLS headers

- Labeled packets are tagged with one or more Label Stack Entries (LSEs), MPLS header, inserted between the frame header (link layer) and the IP packet (network)



MPLS versus IP paths

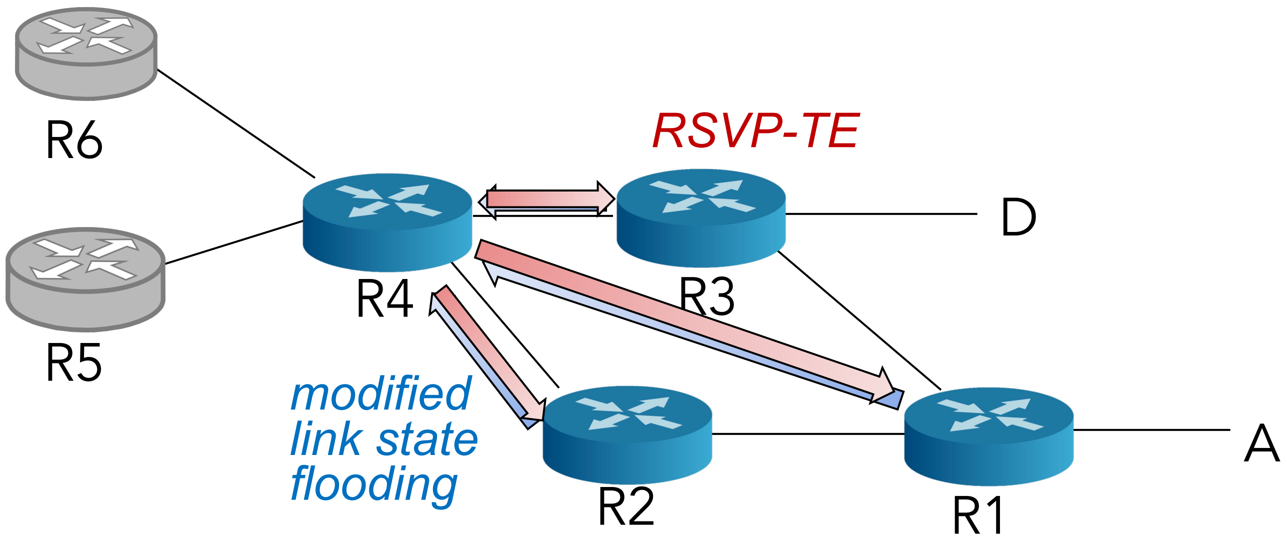
- IP routing: path to destination determined by dst address alone
- MPLS forwarding decisions can differ from those of IP
  - Use dst and src addresses to route flows to same dst differently (traffic engineering)
  - Re-route flows quickly if link fails: pre-computed backup paths



Entry router (R4) can use *different* MPLS routes to A based, e.g., on source address

MPLS signaling

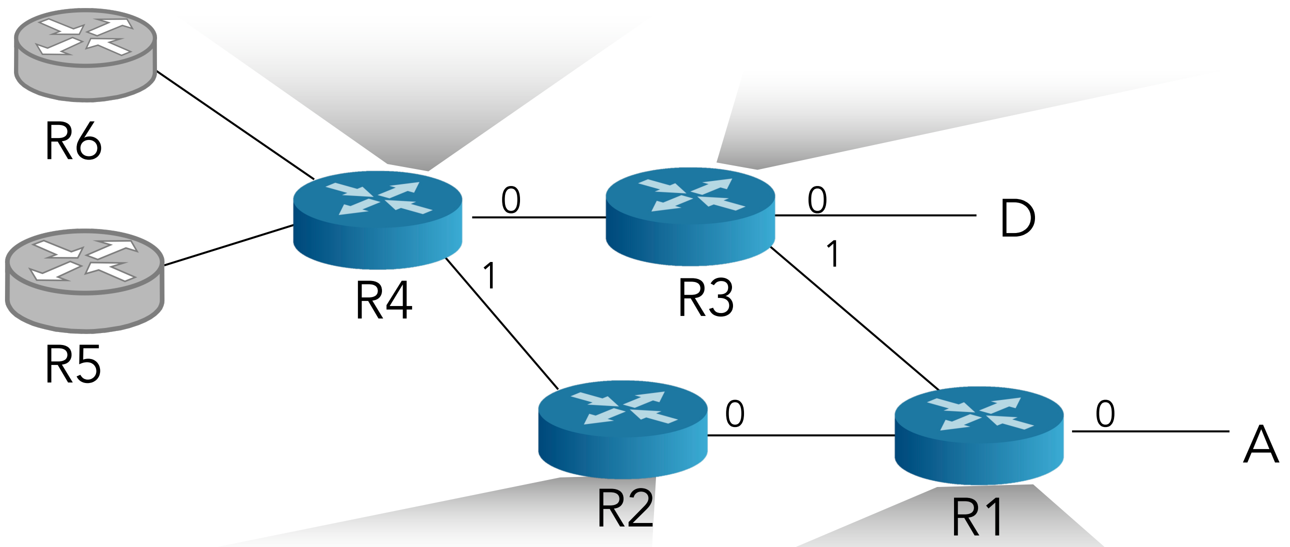
- Modify OSPF, IS-IS link-state flooding protocols to carry info used by MPLS routing,
  - e.g., link bandwidth, amount of "reserved" link bandwidth
  - Entry MPLS router uses RSVP-TE signaling protocol to set up MPLS forwarding at downstream routers



MPLS forwarding tables

In label	Out label	Dest	Out i/f
	10	A	0
	12	D	0
	8	A	1

In label	Out label	Dest	Out i/f
10	6	A	1
12	9	D	0

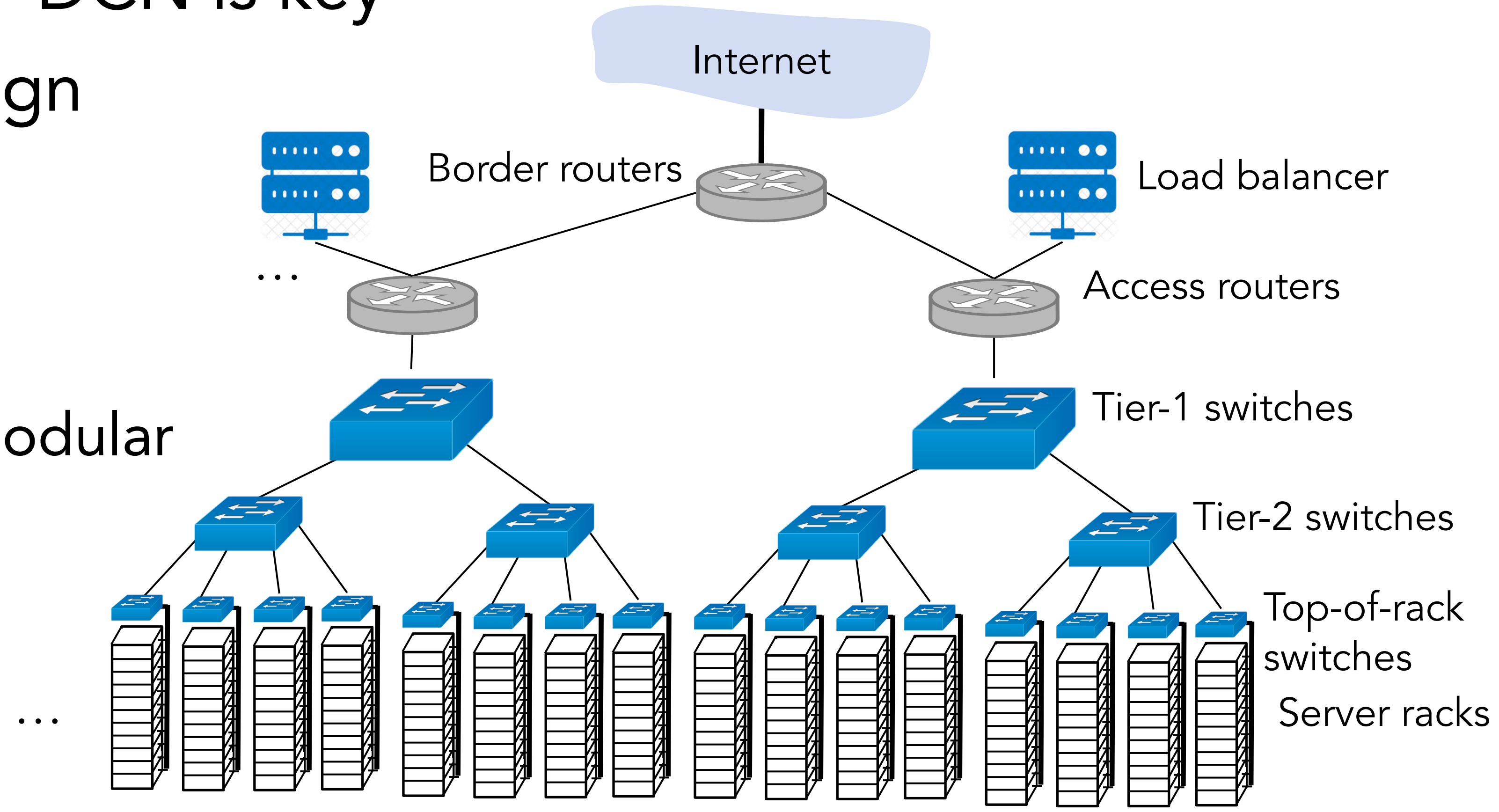


In label	Out label	Dest	Out i/f
8	6	A	0

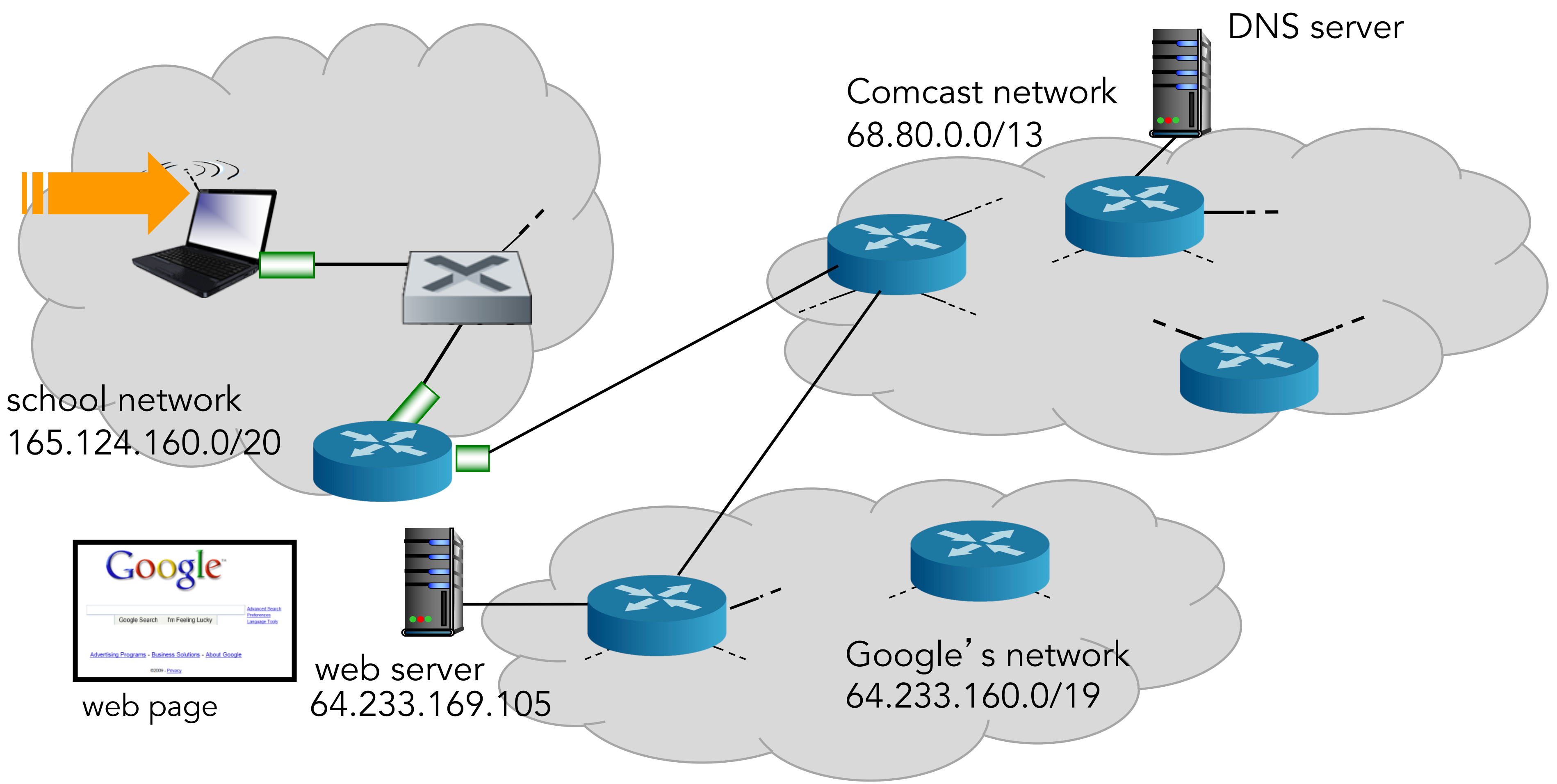
In label	Out label	Dest	Out i/f
6	-	A	0

Data center networks

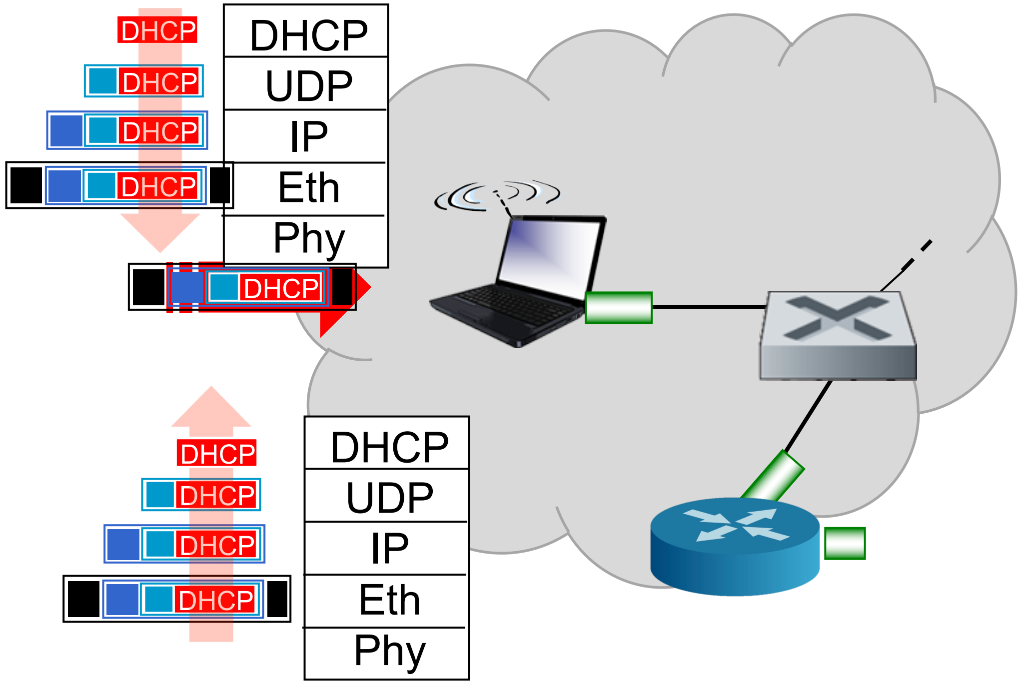
- Google, Microsoft, Amazon, … building big data centers
  – Housing 10s or 100s of thousands of machines
- A fast and efficient DCN is key
- Much work on design of DCN
  – Topologies
  – Routing
  – Container-based modular data centers
  – Reliability



A recap of sorts … visiting www.google.com?

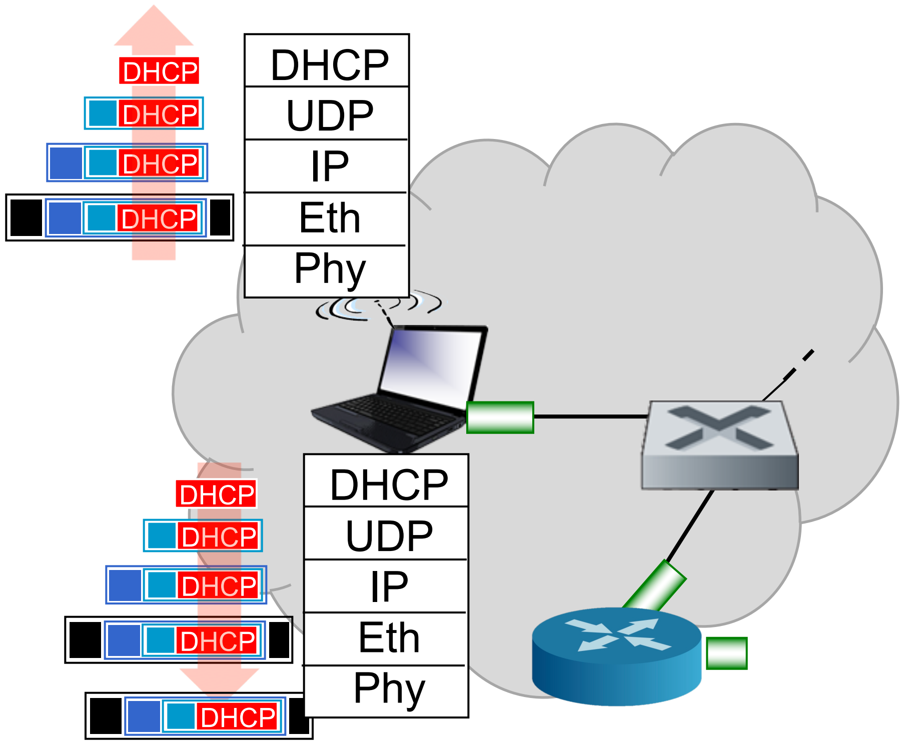


What happens when visiting www.google.com?



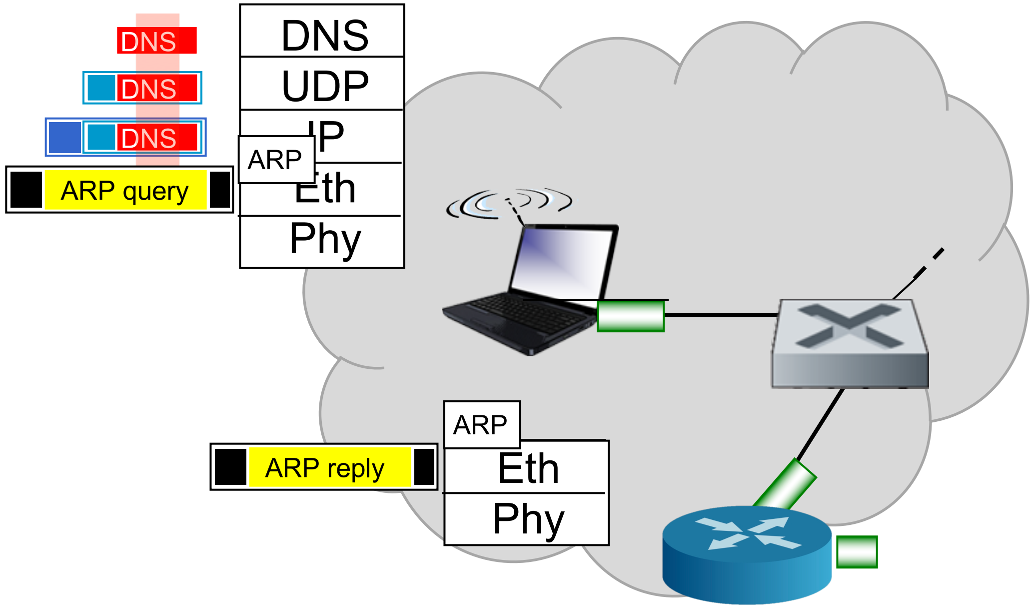
- Connecting laptop needs to get its own IP address, address of first-hop router, address of DNS server: use DHCP
- OS creates a DHCP request encapsulated in UDP, encapsulated in IP, encapsulated in 802.3 Ethernet
- Ethernet frame broadcast (dest: FF:FF:FF:FF:FF:FF) on LAN, received by router running DHCP server
- Ethernet unpacked to get IP, unpacked to get UDP, unpacked to get DHCP

Connecting to network (continued)



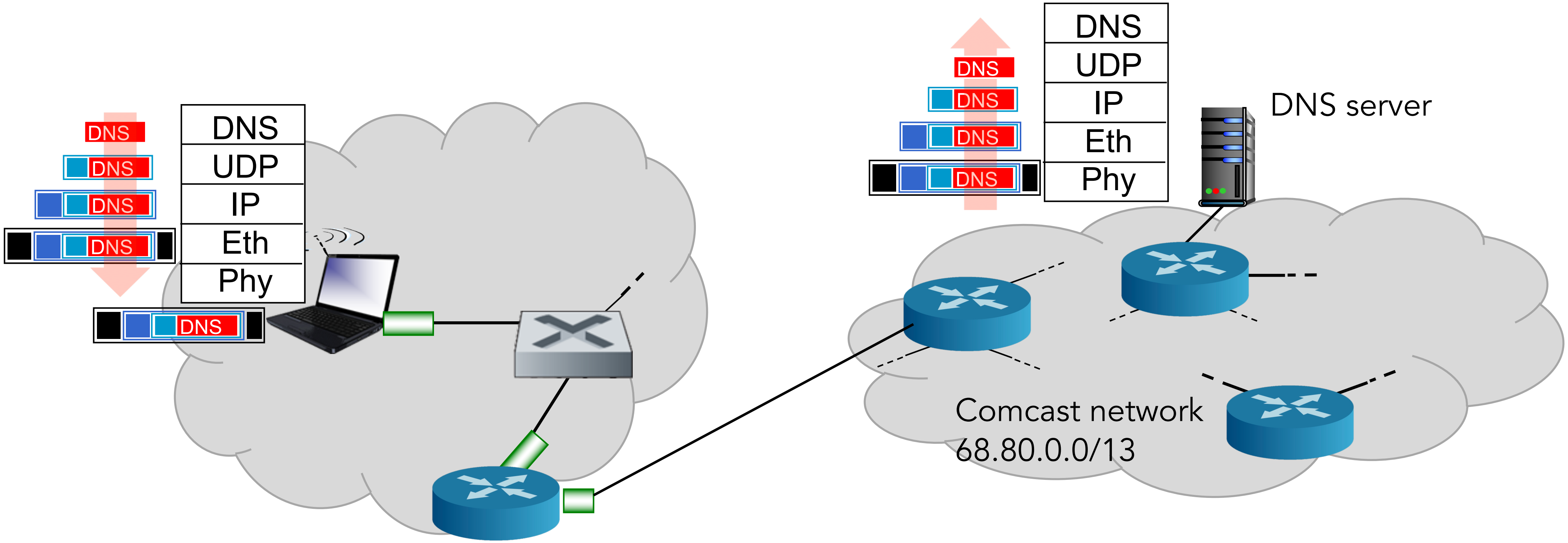
- DHCP server creates DHCP ACK containing client's IP address, IP address of first-hop router for client, subnet mask, & IP address of DNS server
- Encapsulation at DHCP server, frame forwarded (switch learning) through LAN, demultiplexing at client
- Client receives DHCP ACK

ARP (before DNS, before HTTP)



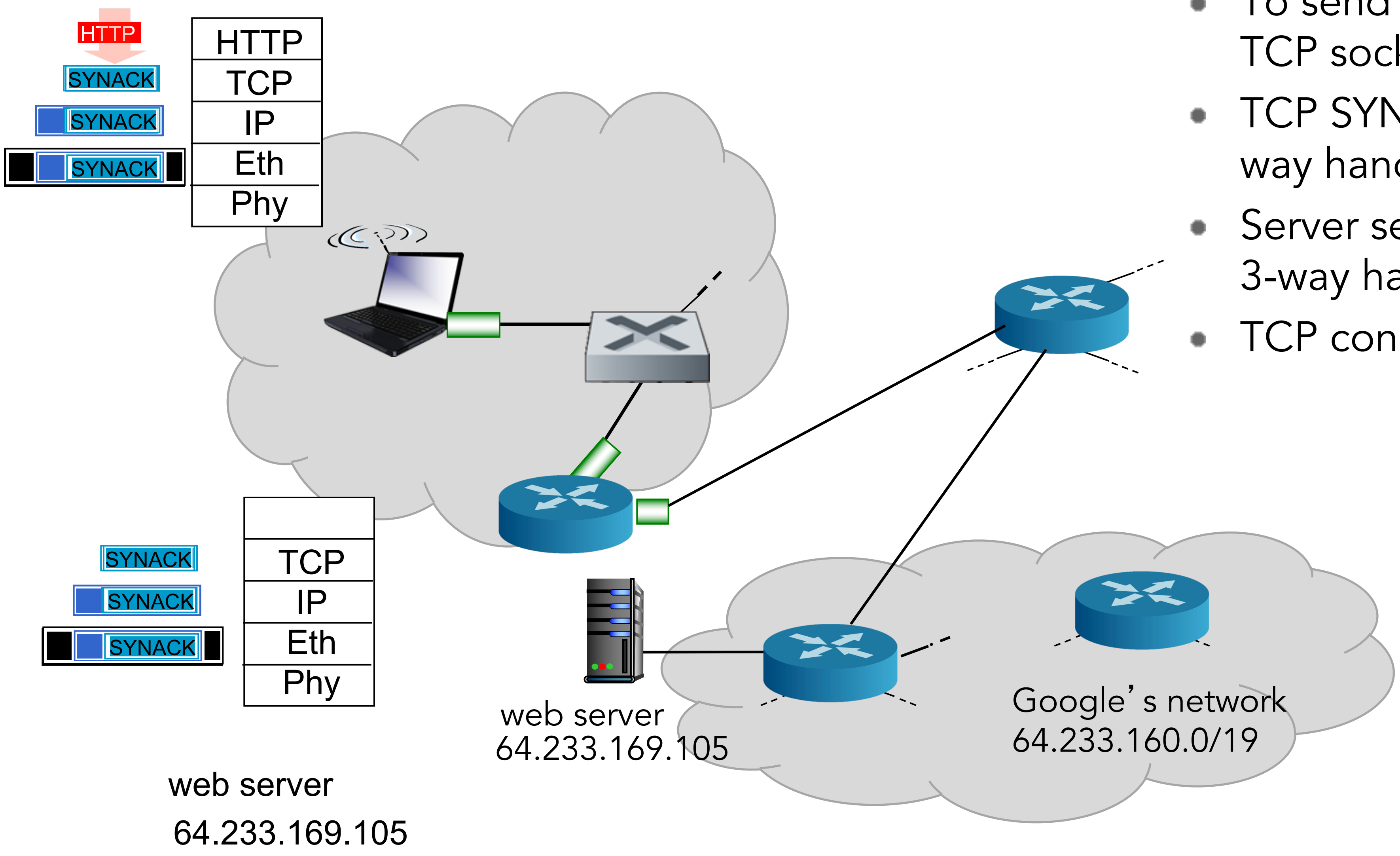
- Before sending HTTP request, we need to send a DNS request to get IP address of www.google.com
- Create DNS request, inside UDP segment, inside IP datagram, inside Ethernet frame, but we don't yet have the MAC address of the router to set as the first-hop Ethernet destination
- Client broadcasts ARP query listing the router's IP address. Router replies with its MAC address (on that subnet)

DNS



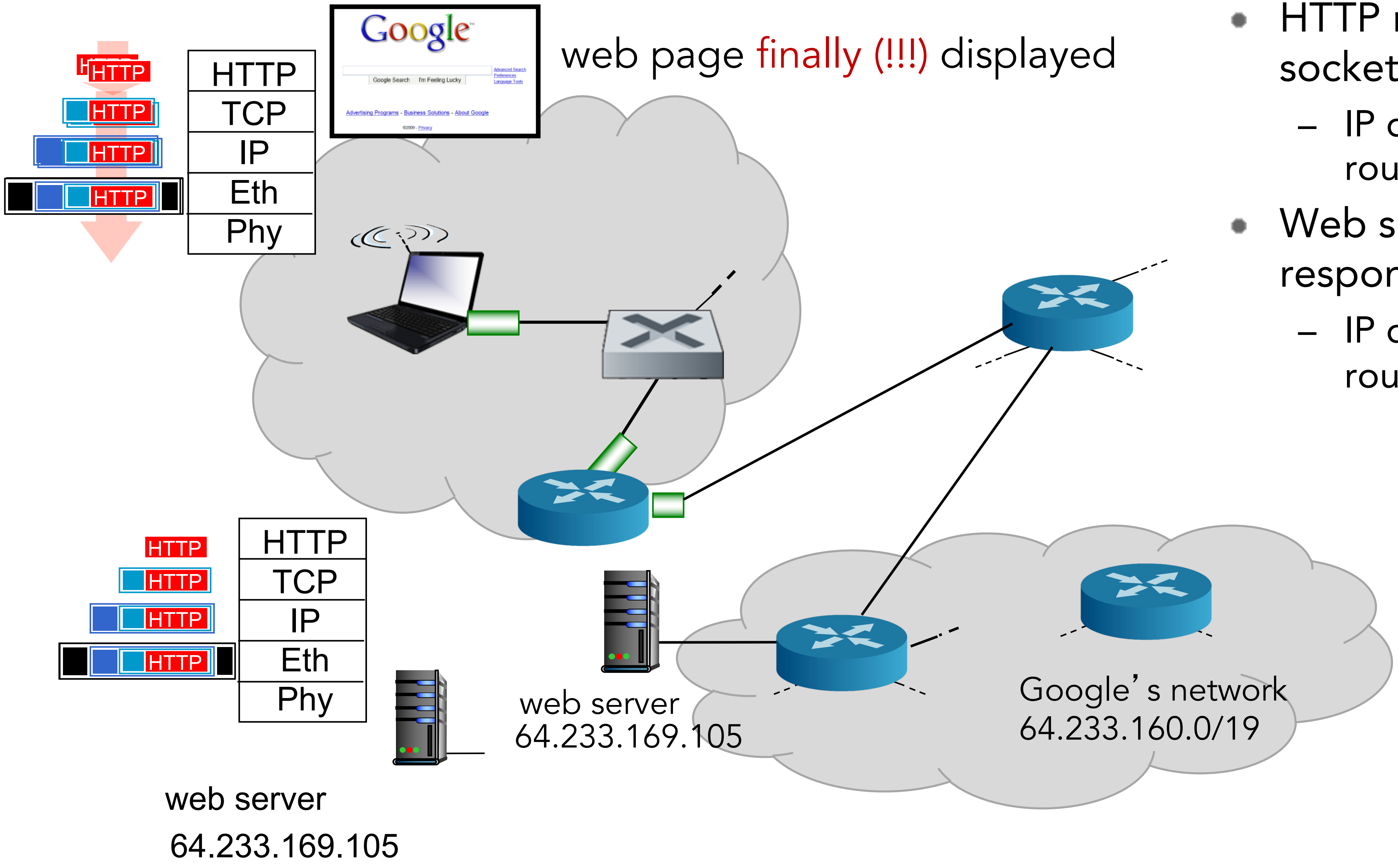
- IP datagram containing DNS query forwarded via switch from client to 1st hop router
- … from campus network into Comcast network, routed (tables already created by RIP, OSPF, IS-IS and/or BGP routing protocols) to DNS server
- DNS server replies with IP address of www.google.com

TCP



- To send HTTP request, client opens TCP socket to server on port 80
- TCP SYN packet sent -- step 1 of 3-way handshake
- Server sends SYN-ACK – step 2 of 3-way handshake
- TCP connection established!

HTTP request/reply



- HTTP request sent into TCP socket
  - IP datagram containing request routed to google.com
- Web server replies with HTTP response (incl. web page)
  - IP datagram containing response routed back to client

The End