Chapter 6 The Data Link layer

6.1 introduction, services

6.2 error detection, correction

6.3 multiple access protocols

6.4 LANs
- addressing, ARP
- Ethernet
- layer-2 switches
- VLANS

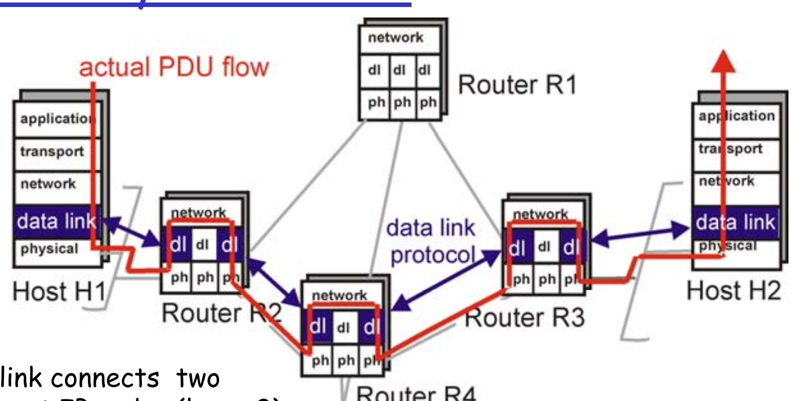
6.5 link virtualization: MPLS

6.6 data center networks

6.7 a day in the life of a web request

(play animation in .ppt slide on your own)

Link Layer: context

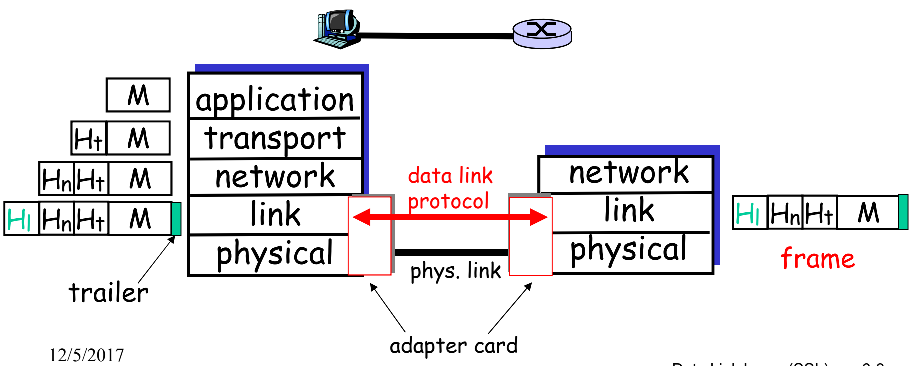


- A link connects two adjacent IP nodes (layer 3) along a path
  - An Ethernet switch (layer 2) is considered to be part of a link

- IP datagram transferred by different link protocols over different links which may provide different services

Link Layer: context

- unit of data: *frame*, which encapsulates an IP datagram
- IP expects no service guarantee from links
- Link can be
  - wire
  - wireless
  - LAN (layer 2)
  - WAN (virtual link)



Link Layer Services

- Framing
  - Encapsulate datagram with header and trailer
- Error Detection
  - errors caused by signal attenuation, noise.
  - receiver detects presence of errors
- Error Correction
  - receiver identifies *and corrects* bit error(s) without resorting to retransmission
- Link access
  - access protocol for shared channel access
  - *"MAC" addresses* used in frame headers to identify source, destination
    - different from IP addresses
    - **why both MAC and IP addresses?**

Link Layer Services (more)

- Half-duplex and full-duplex
  - with half duplex (shared channel), nodes at both ends of link can transmit, but *not* at same time
- Flow Control
  - pacing between sender and receiver(s)
- Reliable delivery between two physically connected devices
  - we learned how to do this already (chapter 3)
  - seldom used on low error-rate links (fiber, some twisted pair)
  - wireless links: high error rates

Q: why both link-level and end-end reliability?

Chapter 6 The Data Link layer

6.1 introduction, services

6.2 error detection, correction

6.3 multiple access protocols

6.4 LANs
- addressing, ARP
- Ethernet
- layer-2 switches
- VLANS

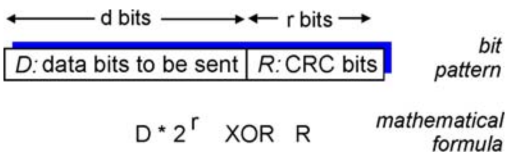
6.5 link virtualization: MPLS

6.6 data center networks

6.7 a day in the life of a web request
(play animation in .ppt slide on your own)

Cyclic Redundancy Check (CRC) - sender

- View data bits, D, as a binary number



- Choose r+1 bit pattern (generator), G
- **Goal**: choose r CRC bits, R, such that <D,R> is exactly divisible by G using modulo 2 arithmetic
- Modulo 2 arithmetic
  - there is no carry in addition, and no borrow in subtraction
  - addition and subtraction same as bitwise exclusive OR (XOR)

12/5/2017

Data Link Layer (SSL) 6-7

Cyclic Redundancy Check (CRC) - receiver

- **Bit string** <D,R> sent is exactly divisible by G
- Receiver knows G, performs division. If non-zero remainder, *error detected !*
- can detect all burst errors less than r+1 bits;
- longer burst errors are detectable with probability $1-(0.5)^r$

12/5/2017

Data Link Layer (SSL) 6-8

CRC Theory and Example

Want:

$(D*2^r)$ XOR R = nG

add R to both sides:

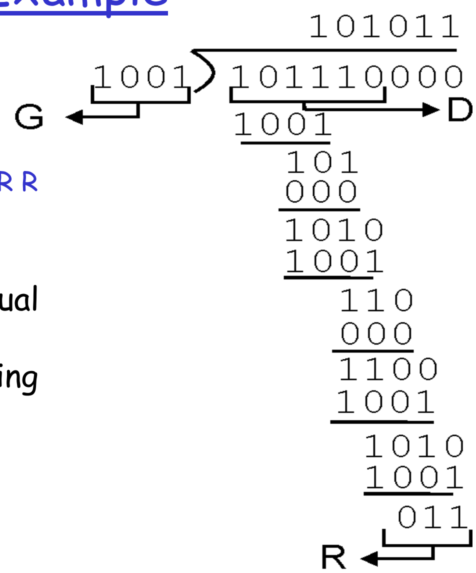
$D*2^r$ ~~XOR R XOR R~~ = (nG) XOR R

Equivalently:

the remainder from dividing $D*2^r$ by G is equal to R;

the desired CRC bit string is

$R = \text{remainder}[\frac{D*2^r}{G}]$



Chapter 6 The Data Link layer

6.1 introduction, services

6.2 error detection, correction

6.3 multiple access protocols

6.4 LANs

- addressing, ARP
- Ethernet
- layer-2 switches
- VLANS

6.5 link virtualization: MPLS

6.6 data center networks

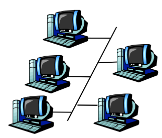
6.7 a day in the life of a web request

(play animation in .ppt slide on your own)

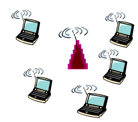
Links and Multiple Access Protocols

Two types of "links":

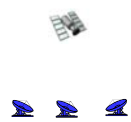
- point-to-point
  - fiber optic link
  - link between Ethernet switch and host
- broadcast (shared wire or medium)
  - old-fashioned Ethernet
  - shared coax cable in HFC (hybrid fiber cable), e.g., Spectrum
  - wireless (802.11 LAN and others), etc.



shared cable (e.g., old Ethernet)



shared RF (e.g., 802.11 WiFi)



shared RF (satellite)

humans at a party (shared air, acoustics)

Multiple Access protocols

single shared broadcast channel

- two or more simultaneous transmissions by nodes may interfere with each other
  - ***collision*** if a node receives two or more signals at the same time

- Need a protocol to determine when nodes can transmit
  - no out-of-band channel for coordination

MA Protocols: a taxonomy

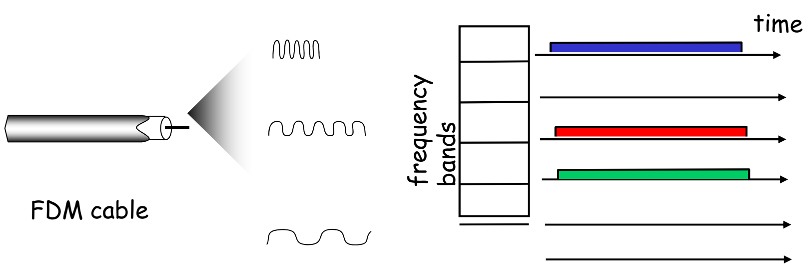
Three broad classes:

- Channel Partitioning (e.g., cell phones)
  - divide channel into smaller "pieces" (frequency bands, time slots, codes)
  - allocate a piece to each node for exclusive use
- Random Access (e.g., early Ethernet, 802.11 wifi)
  - shared channel , collisions allowed
  - "recover" from collisions
  - does not provide QoS
- "Taking turns" (e.g., token-ring LAN, FDDI)
  - nodes take turns
  - a node with more to send can take a longer turn

Channel Partitioning protocols

FDMA: frequency division ***multiple access****

- each station assigned a fixed frequency band (note: MIMO antenna can use multiple frequencies)
- unused transmission time in frequency bands go idle

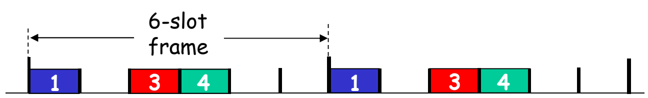


* *multiple transmitters*

Channel Partitioning protocols

TDMA: time division ***multiple access****

- each station gets fixed length slot (length = pkt trans time) in each frame
  - requires time synchronization
- unused slots go idle



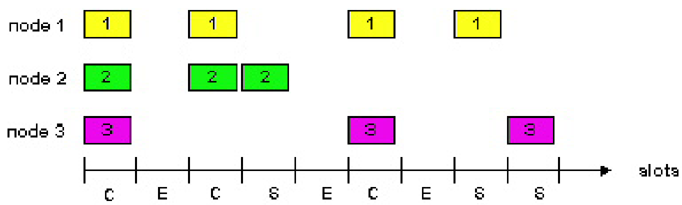
* *multiple transmitters*

Random Access Protocols

- When node has packet to send
  - transmit at ***full*** channel data rate
  - no *a priori* coordination among nodes
- two or more transmitting nodes ➜ "collision"
- random access MA protocol specifies:
  - how to detect collision
  - how to recover from collision (e.g., via delayed retransmissions)
- examples (chronological):
  - ALOHA
  - slotted ALOHA
  - CSMA, CSMA/CD, CSMA/CA

Slotted Aloha

- time is divided into equal size slots (pkt trans. times)
  - requires time synchronization
- node with new arriving pkt: transmit at beginning of next slot
- if collision: retransmit pkt in a future slot with probability p (or one of K slots at random), until successful.



Success (S), Collision (C), Empty (E) slots

12/5/2017

Slotted Aloha efficiency

Long-term fraction of time slots that are successful?

Suppose N nodes have packets to send

- each transmits in slot with probability *p*
- prob. successful transmission *S* is

by a particular node: $S = p\,(1-p)^{(N-1)}$

by any of N nodes:

S = Prob [one of N nodes transmits]

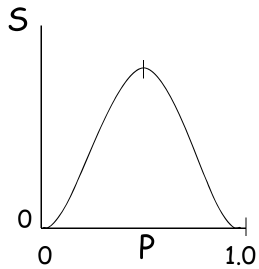
$= N\,p\,(1-p)^{(N-1)}$

… choosing optimum p, let N -> infinity

$= 1/e = .37$ as N -> infinity

Channel occupied by useful transmissions < 37% of time

12/5/2017



$$\frac{\partial S}{\partial P} = \frac{\partial}{\partial P}[NP(1-P)^{N-1}]$$

$$= NP\frac{\partial}{\partial P}(1-P)^{N-1} + (1-P)^{N-1}N$$

$$= -NP(N-1)(1-P)^{N-2} + N(1-P)^{N-1}$$

$$= N(1-P)^{N-2}\{-P(N-1)+1-P\}$$

$$= N(1-P)^{N-2}\{-NP+P+1-P\}$$

$$\frac{\partial S}{\partial P} = 0 \text{ when } P = \frac{1}{N} \text{ to maximize } S$$

My terminology : "Probability Division Multiplex"
Division of probability does not have to be fair, i.e.,
$P_1+P_2+ \ldots +P_N = 1$ is condition for maximum

$$S_{max} = NP(1-P)^{N-1}\Big|_{P=\frac{1}{N}}$$

$$= N\left(\frac{1}{N}\right)\left(1-\frac{1}{N}\right)^{N-1}$$

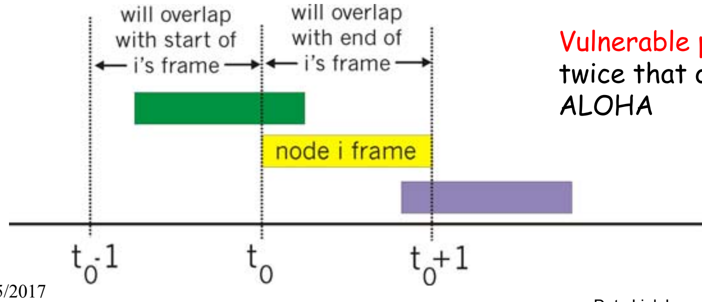
$$= \left(1-\frac{1}{N}\right)^{N-1} \xrightarrow{N\to\infty} e^{-1}$$

$$\frac{1}{e} \cong 0.368$$

which is maximum throughput (efficiency) of the slotted ALOHA protocol

Pure (unslotted) ALOHA

- unslotted Aloha: no time synchronization
- when frame arrives
  - send immediately (without waiting for beginning of slot)
- collision probability increases:
  - frame sent at t_0 can collide with another frame sent within $[t_0-1, t_0+1]$



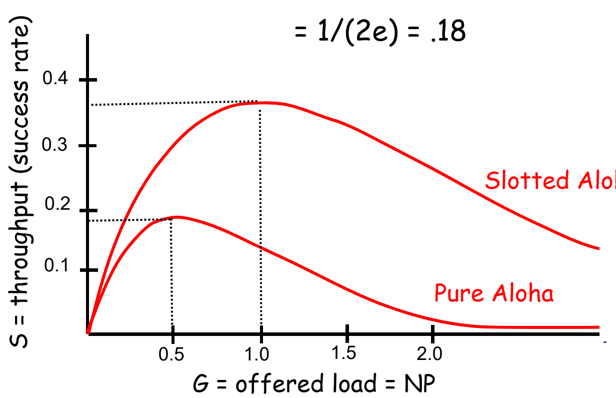
Vulnerable period is twice that of slotted ALOHA

Pure Aloha (cont.)

P(success by any of N nodes)

… choosing optimum P, let N -> infinity …

= 1/(2e) = .18



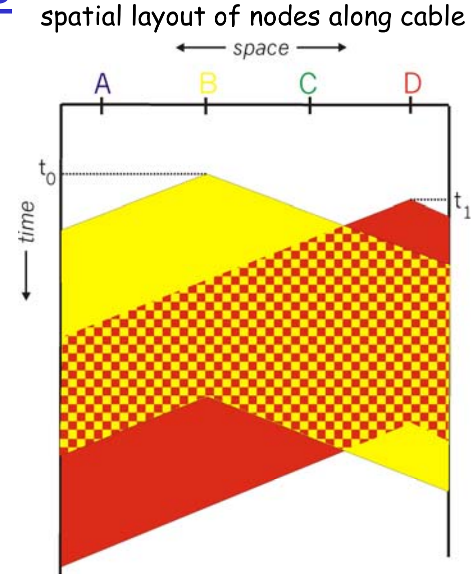
CSMA: Carrier Sense Multiple Access

CSMA: listen before transmit (for a channel with short propagation delay)

- If channel sensed idle: transmit entire packet
- If channel sensed busy, **defer** transmission;
  - *retry after some random interval*
- human analogy: don't interrupt when someone else is speaking

CSMA collisions

spatial layout of nodes along cable



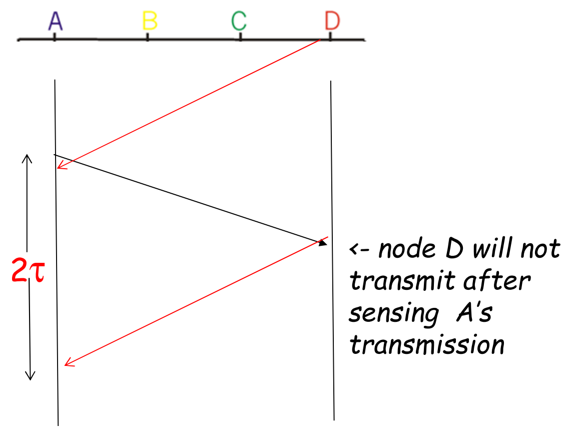
collisions *can* occur:
it takes time for two nodes to hear each other's transmission due to propagation delay

collision:
entire packet transmission time wasted

Vulnerable period of a transmission

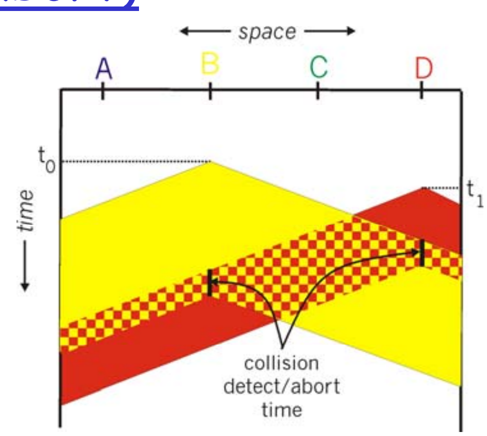
Let τ be the maximum one-way propagation delay between two nodes in a subnet

If sender A detects no collision after 2τ seconds, then it knows that its transmission will be successful



Vulnerable period is 2τ

CSMA/CD collision detection (& abort)



CSMA/CD

- carrier sensing, deferral as in CSMA
  - CD useful for channels where collisions are detectable within a *short* time
  - colliding transmissions aborted, reducing channel wastage
- collision detection is
  - *easy* in wired LANs: measure signal strength, compare transmitted and received signals
  - *difficult* in wireless LANs: received signal overwhelmed by local transmission signal
- high channel utilization possible by sending very long packets (relative to propagation delay)

CSMA/CD channel efficiency

Channel efficiency = t_{trans}/(contention period + t_{trans})

where t_{trans} is average transmission time of a frame

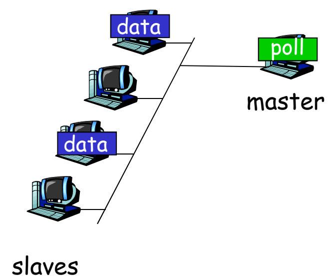
Let t_{prop} denote the maximum propagation delay between any two nodes. Then a good estimate of the average contention period is $2t_{prop}e$. (Why ?)

CSMA/CD channel efficiency = t_{trans} / ($2t_{prop}e$ + t_{trans})

"Taking Turns" MA protocols

Polling:

- master node "invites" slave nodes to transmit in turn
- concerns:
  - polling overhead
  - latency (for large N)
  - single point of failure (master)

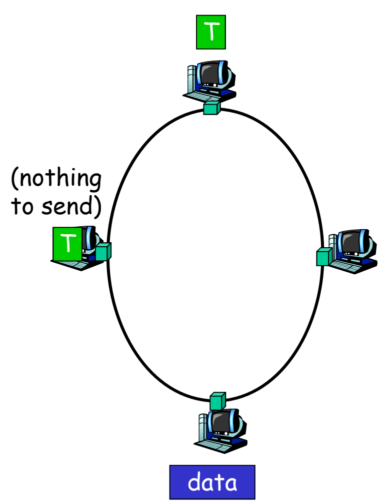


"Taking Turns" MA protocols

Token passing:

- control **token** (short msg) passed from one node to next sequentially.
- Data removed from ring by its sender

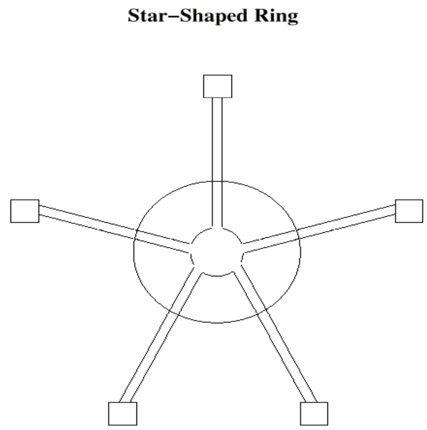
  => broadcast
- concerns:
- latency (for large N)
- single point of failure
  - ring interface is an active repeater
  - token loss



Solution: Star-shaped Ring Topology

Example:
Token ring (IEEE 802.5)

with wiring closet



Today's Ethernet uses a star topology

Generalization: A hierarchical ring (with multiple wiring centers to reduce cable length).

Chapter 6 The Data Link layer

6.1 introduction, services

6.2 error detection, correction

6.3 multiple access protocols

6.4 LANs
- addressing, ARP
- Ethernet
- layer-2 switches
- VLANS

6.5 link virtualization: MPLS

6.6 data center networks

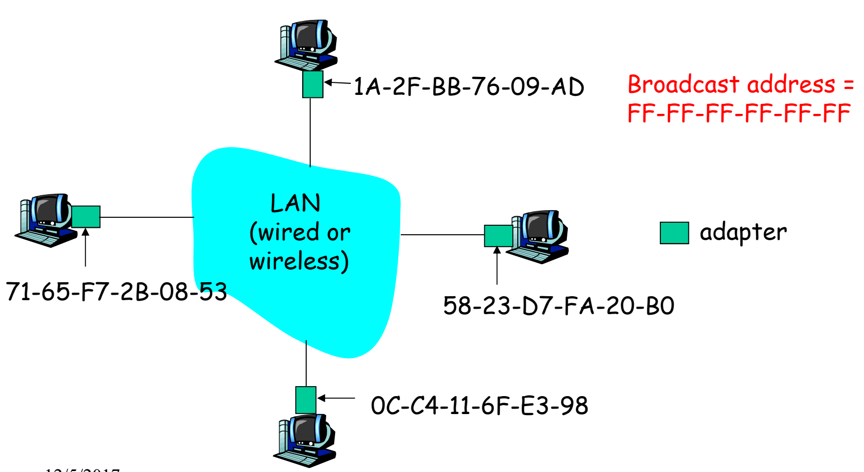
6.7 a day in the life of a web request
(play animation in .ppt slide on your own)

MAC and IP Addresses

- 32-bit IP address:
  - *network-layer* address
  - used to get datagram to destination IP subnet
- 48 bit MAC address (or LAN or *Ethernet* or link-layer address):
  - e.g.: 1A-2F-BB-76-09-AD (hexadecimal notation)
  - burned in NIC ROM (sometimes software settable)
  - used to *get frame from one interface to another interface in same subnet*
- MAC address necessary?

MAC Addresses

Each adapter on LAN has unique MAC address

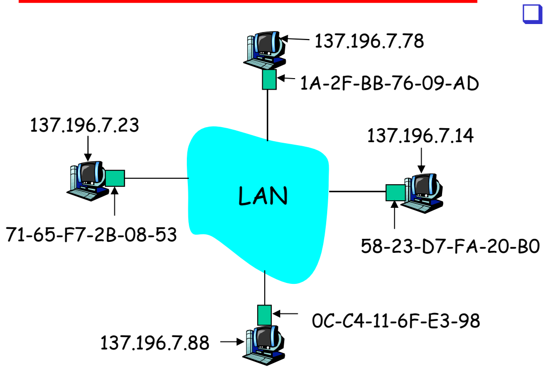


MAC Address vs. IP address

- MAC addresses are flat
  - MAC address allocation administered by IEEE
    - manufacturers buy blocks of MAC address space for a nominal fee
  - MAC addresses are portable
    - LAN card can be moved from one LAN to another, e.g., laptop
- IP's hierarchical address NOT portable
  - address depends on IP subnet to which node is attached
- analogy:
  - (a) MAC address: like Social Security Number
  - (b) IP address: like postal address

ARP: Address Resolution Protocol

Question: how to determine MAC address of interface B knowing B's IP address?



- Each IP node (host, router) on LAN has ARP table
- ARP table: 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)

ARP protocol: Same LAN

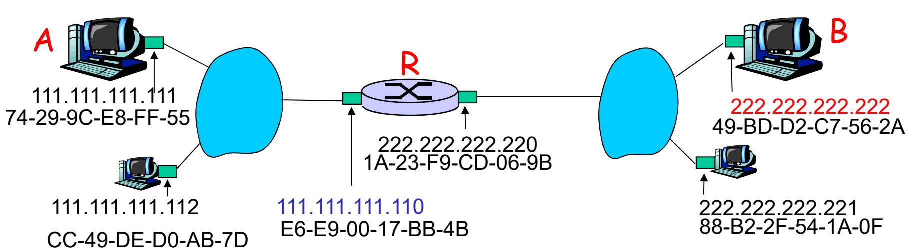
- A wants to send datagram to B, and B's MAC address *not* in A's ARP table.
- A broadcasts ARP query packet, containing B's IP address
  - Dest MAC address = FF-FF-FF-FF-FF-FF
  - all machines on LAN receive ARP query
- B receives ARP packet, replies to A with its (B's) MAC address
  - frame sent to A's MAC address (unicast)
- A caches IP-to-MAC address pair in its **ARP table**

  soft state
  - information that times out (goes away) unless refreshed
  - enhances performance but not necessary for correctness
- ARP enables "plug-and-play":
  - nodes create their ARP tables without any work by net administrator

Addressing: routing to another LAN

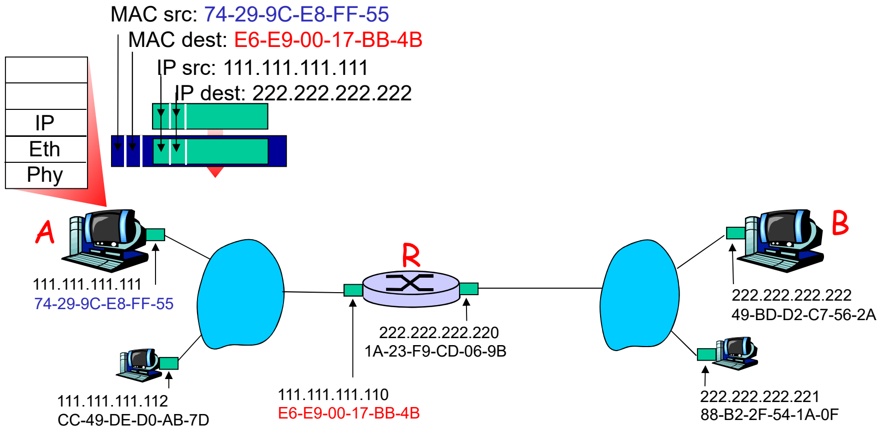
walkthrough: A sends datagram to B via R.

- focus on addressing - at both IP (datagram) and MAC layer (frame)
- A knows B's IP address
- A knows IP address of first-hop router, R
- A knows MAC address of first hop router's interface (how?)



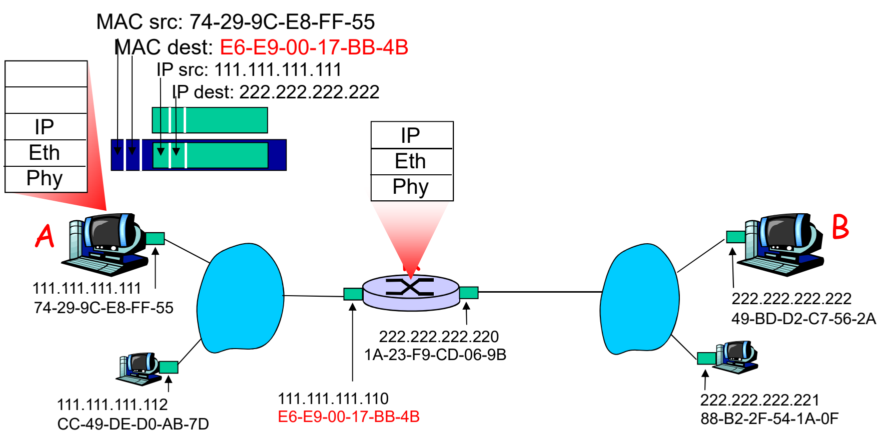
Addressing: routing to another LAN

- A creates IP datagram with IP source A, destination B
- A creates link-layer frame with R's MAC address as dest, frame contains A-to-B IP datagram



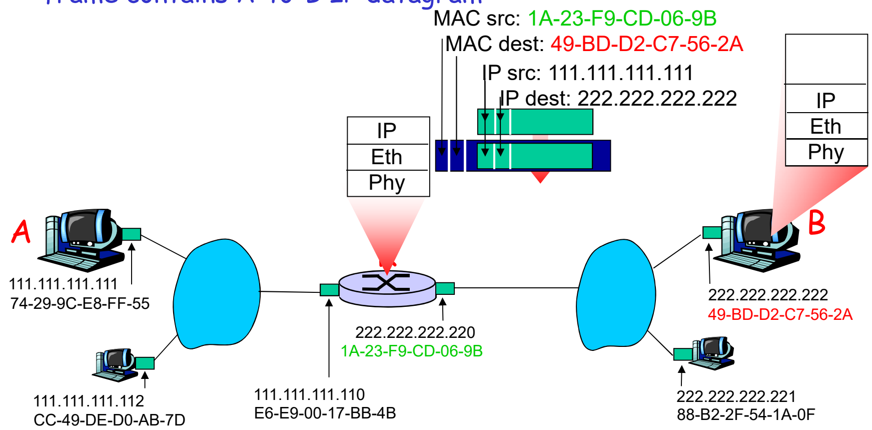
Addressing: routing to another LAN

- frame sent from A to R
- frame received at R, datagram passed up to IP



Addressing: routing to another LAN

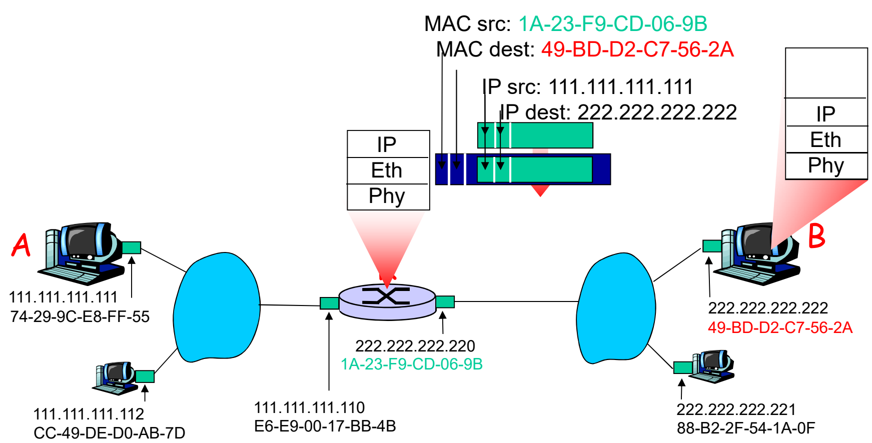
- R forwards datagram with IP source A, destination B
- R looks up B's MAC address
- R creates link-layer frame with B's MAC address as dest, frame contains A-to-B IP datagram



12/5/2017

Addressing: routing to another LAN

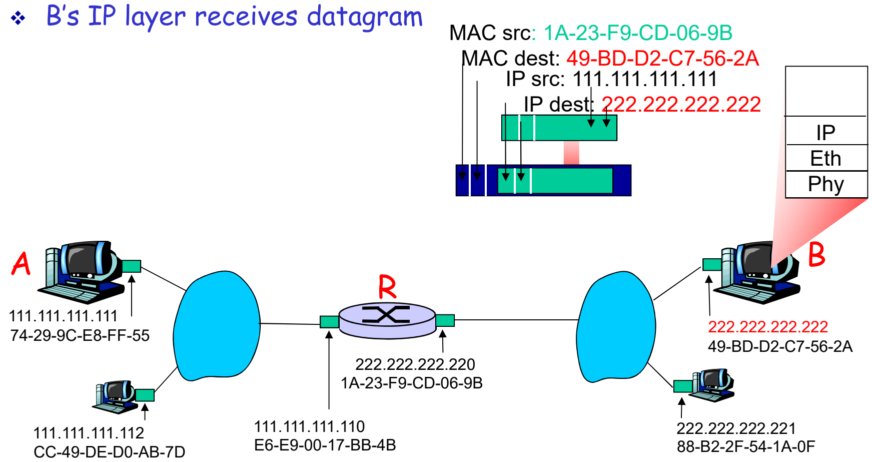
- R sends frame to B



12/5/2017

Addressing: routing to another LAN

- R sends frame to B
- B's IP layer receives datagram



Link layer, LANs

5.1 introduction, services

5.2 error detection, correction

5.3 multiple access protocols

5.4 LANs
- addressing, ARP
- Ethernet
- switches
- VLANS

5.5 link virtualization: MPLS

5.6 data center networks

5.7 a day in the life of a web request
(play animation in .ppt slides on your own)

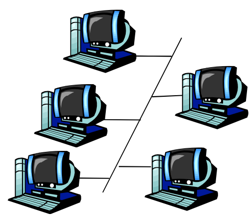
Ethernet

"dominant" wired LAN technology:

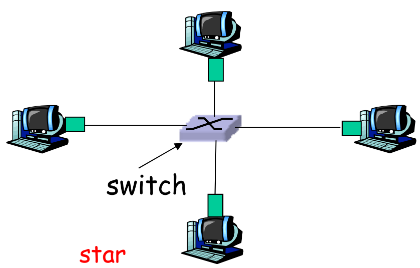
- cheap, $20 for NIC
- first widely used LAN technology
- simpler, cheaper than competitors
  - token-ring (16 Mbps), FDDI (100 Mbps), and ATM (155 Mbps)
- kept up with speed race: 10 Mbps – 10 Gbps

Star topology

- bus topology popular through mid 90s, and later star topology with hub at center
  - all nodes in same collision domain (their transmissions can collide with each other)
- today: star topology with active switch (layer 2) at center
  - no collision



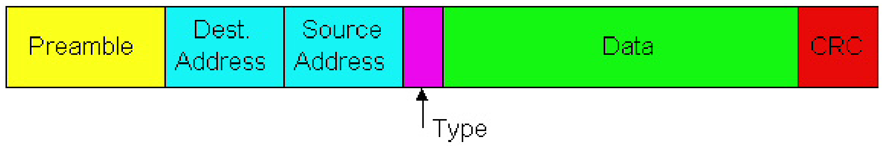
bus: coaxial cable



star

Ethernet Frame Structure

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

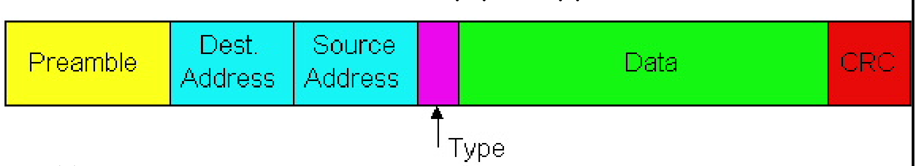


Preamble:

- 7 bytes with pattern 10101010 followed by one byte with pattern 10101011
- used to synchronize receiver, sender clocks
  - long preamble used due to "burst" nature of transmissions, unlike a synchronous point to point link

Ethernet Frame Structure (cont.)

- Addresses: 6 bytes
  - if adapter receives frame with matching destination address, or with broadcast address (eg ARP packet), it passes data in frame to network-layer protocol
  - else adapter discards frame
- Type: 2 bytes, indicates the higher layer protocol, ARP or IP (many others are supported such as Novell IPX and AppleTalk)
- CRC: 4 bytes, checked at receiver, if error is detected, the frame is simply dropped



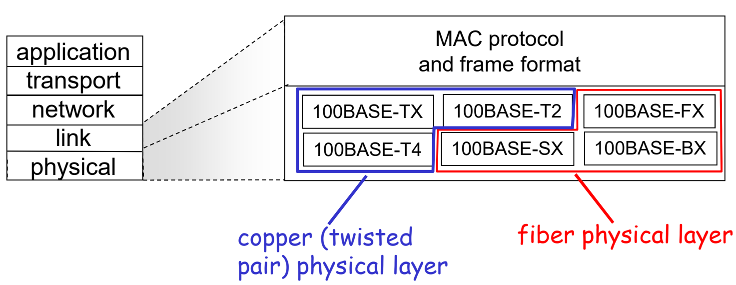
Unreliable, connectionless service

- Connectionless: No handshaking between sending and receiving adapters
- Unreliable: receiving adapter doesn't send acks or nacks to sending adapter
  - stream of datagrams passed to network layer can have gaps
  - gaps will be filled only if app is using TCP
- Ethernet's MAC protocol: CSMA/CD with binary backoff
  - Interval for random retransmission doubles after every additional collision

12/5/2017

802.3 Ethernet Standards: Link & Physical Layers

- *many* different Ethernet standards
  - different speeds: 2 Mbps, 10 Mbps, 100 Mbps, 1Gbps, 10Gbps
  - different physical layer media and technologies: coax cable, twisted pair, fiber
  - same frame format and MAC protocol



12/5/2017

Chapter 6 The Data Link layer

6.1 introduction, services

6.2 error detection, correction

6.3 multiple access protocols

6.4 LANs
- addressing, ARP
- Ethernet
- layer-2 switches
- VLANS

6.5 link virtualization: MPLS

6.6 data center networks

6.7 a day in the life of a web request

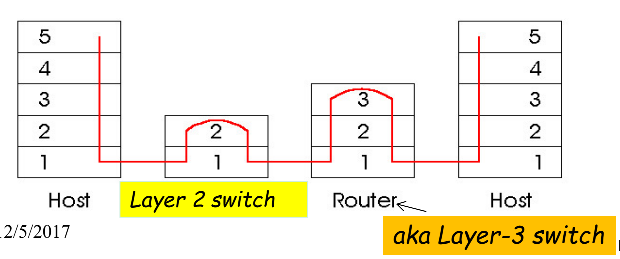
(play animation in .ppt slide on your own)

12/5/2017

Data Link Layer (SSL) 6-51

Layer-2 Switches vs. Routers

- both store-and-forward devices
  - routers: network layer devices examine network layer headers
  - layer-2 switches are link layer devices
- routers maintain forwarding tables, implement routing protocols
- layer-2 switches maintain switch tables, perform filtering and learning



12/5/2017

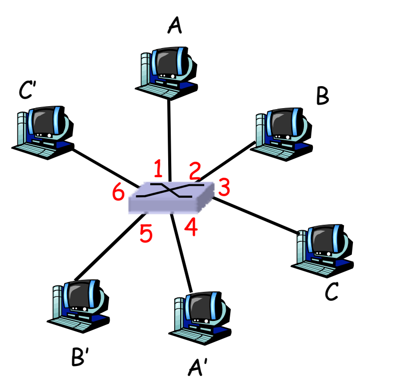
Data Link Layer (SSL) 6-52

Switch (layer 2)

- Link layer device
  - stores and forwards Ethernet frames
  - examines frame header and may selectively forward frame to just one outgoing interface (*instead of broadcast*)
  - it still uses CSMA/CD (just in case an outgoing interface is connected to a hub)
- plug-and-play, self-learning
  - switches do not need to be configured
- transparent
  - hosts are unaware of presence of switches

Switch: allows *multiple* simultaneous transmissions

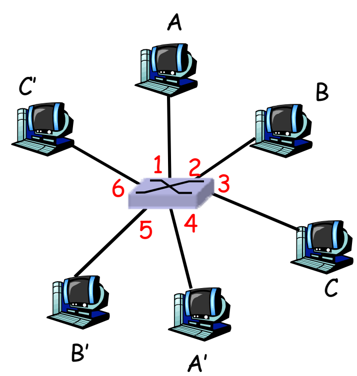
- hosts have dedicated, direct connection (*full duplex)* to switch
- a switch buffers packets
- *switching:* A-to-A' and B-to-B' simultaneously, without collisions
  - not possible with dumb hub



switch with six interfaces (1,2,3,4,5,6)

Switch Table

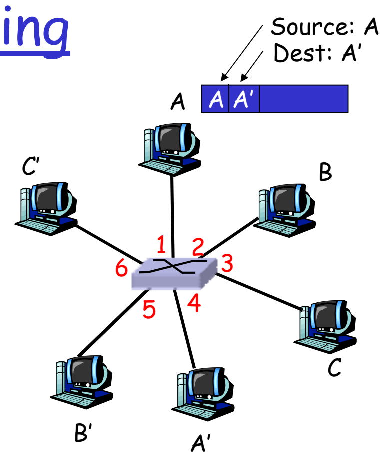
- Q: how does switch know that A' reachable via interface 4, B' reachable via interface 5?
- A: each switch has a switch table, each entry:
  - (MAC address of host, interface to reach host, time stamp)
- looks like a forwarding table for routing
- Q: how are entries created, maintained in switch table?
  - no routing protocol is used



switch with six interfaces (1,2,3,4,5,6)

Switch: self-learning

- switch *learns* which hosts can be reached through which interfaces
  - when frame received, switch "learns" location of sender (incoming LAN segment)
  - records sender/location pair in switch table



What is required to make this work for a network of switches?

MAC addr	interface	TTL
A	1	60

Switch table (initially empty, soft state)

Switch: frame filtering/forwarding

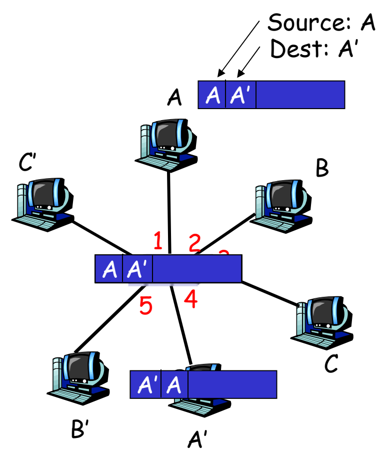
When frame received:

1. record interface associated with sending host
2. check switch table for *MAC destination address*
3. **if** entry in table found for *destination*
   then {
   if dest is on interface from which frame arrived
   then drop the frame
   else forward the frame on interface indicated
   }
   else flood

forward on all but the interface on which the frame arrived

Self-learning, forwarding: example

- destination A' unknown: *flood*
- destination A location known: *selective send*

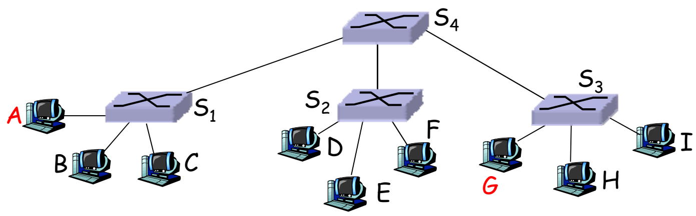


MAC addr	interface	TTL
A	1	60
A'	4	60

Switch table (initially empty)

Interconnecting layer-2 switches

- switches can be connected together

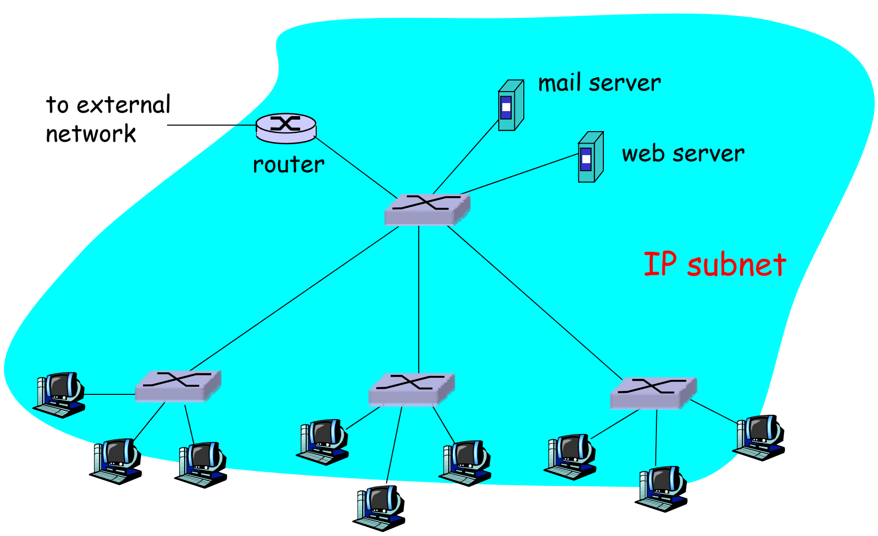


(note: some links are idled if physical topology has loops)

- Q: sending from A to G - how does S_1 know to forward frame destined to G via S_4 (and S_3) ?
- A: self learning (works exactly the same as in single-switch case)

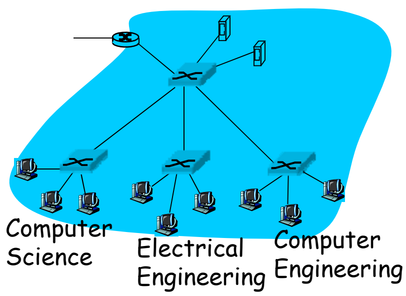
12/5/2017

Institutional network



12/5/2017

Scope of broadcast domain

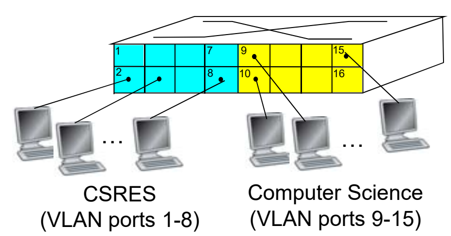


- a single *broadcast domain*
  - all layer-2 broadcast frames (ARP, DHCP, switch-table cache miss, etc.) cross entire LAN => security/privacy, efficiency issues
- multiple broadcast domains

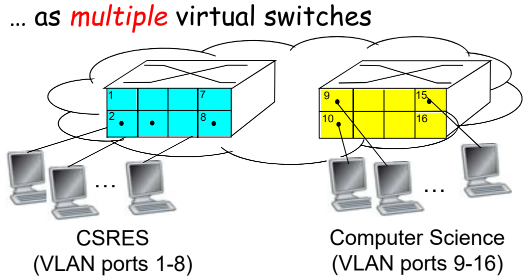
12/5/2017

Port-based VLANs

ports grouped by switch management software for a *single* physical switch to operate



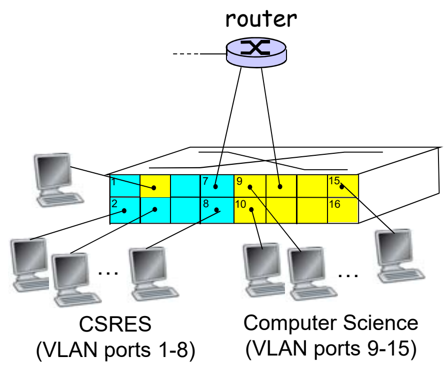
… as *multiple* virtual switches



12/5/2017

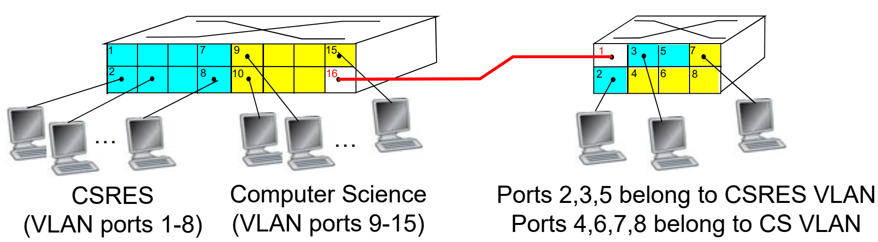
Port-based VLANs (cont.)

- *traffic isolation:* frames to/from ports of a VLAN can *only* reach its ports
  - can also define a **VLAN based on MAC addresses** of endpoints, rather than switch ports
- *dynamic membership:* ports can be dynamically assigned among VLANs
- *forwarding between VLANS:*



- done via a router (just as with separate switches)
  - in practice the router is built into the switch

VLANs spanning multiple switches



- *trunk ports:* carry frames between VLANs defined over multiple physical switches
  - frames forwarded within a VLAN between physical switches must carry VLAN ID info
  - *802.1q protocol* inserts/removes an *additional header field (4 byte VLAN tag)* for each frame forwarded between trunk ports

Chapter 6 The Data Link layer

Link Virtualization: A Network as a Link

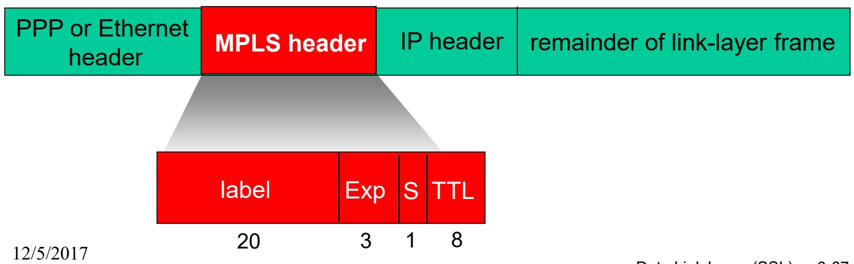
Virtual circuits provided by

- ATM, frame relay, which are packet-switching networks in their own right *(obsolete)*
  - with service models, addressing, routing different from Internet
- A subnet of MPLS capable routers

Each is viewed as a link connecting two IP nodes

Multiprotocol label switching (MPLS)

- *initial goal*: speed up IP forwarding by using fixed-length label (instead of variable-length IP prefix) to do forwarding
  - borrowed the idea from earlier Virtual Circuit approaches
  - MPLS routers insert (and remove) a MPLS header in between the link-layer and IP headers of a frame

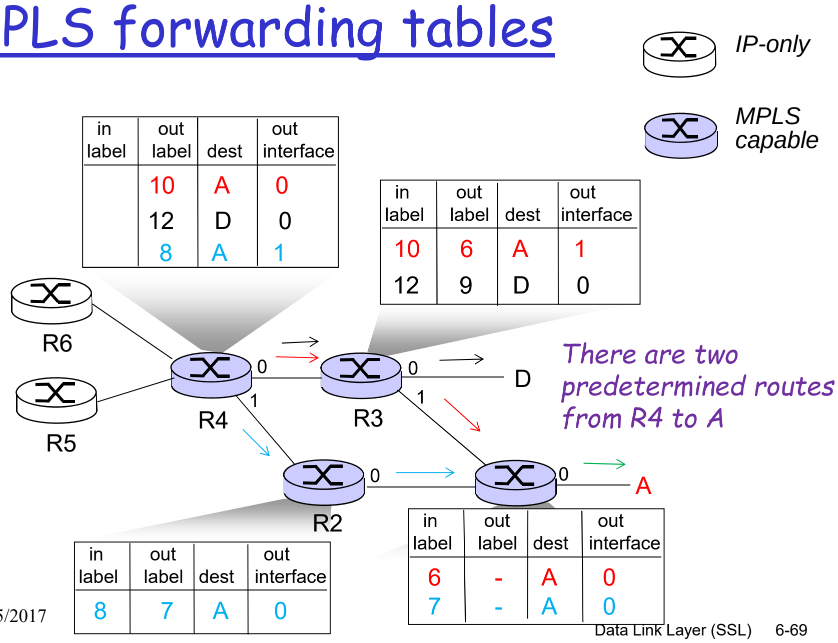


MPLS capable routers

- a.k.a. label-switched router
- forward packets to outgoing interface based only on label value (*does not inspect IP address*)
  - Much faster than longest prefix match
  - MPLS forwarding table distinct from IP forwarding tables
- ***flexibility***: MPLS forwarding decisions can *differ* from those of IP

Note: The router that serves as entrance to a MPLS tunnel filters packets - some packets do not enter tunnel and are forwarded using their IP destination addresses

MPLS forwarding tables



in label	out label	dest	out interface
	10	A	0
	12	D	0
	8	A	1

in label	out label	dest	out interface
10	6	A	1
12	9	D	0

in label	out label	dest	out interface
8	7	A	0

in label	out label	dest	out interface
6	-	A	0
7	-	A	0

MPLS applications

- Fast failure recovery - rerouting flows quickly to pre-computed backup paths (useful for VoIP)
- Traffic engineering – network operator can override IP routing and allocate traffic toward the same destination to multiple paths
- Resource provision for virtual links in private networks

Chapter 6 The Data Link layer

Data center networks

- 10's to 100's of thousands of hosts in close proximity supporting *cloud* applications
  - e-business (e.g. Amazon)
  - content-servers (e.g., YouTube, Akamai, Apple, Microsoft)
  - search engines, data mining (e.g., Google)
- challenges:
  - multiple applications, each serving massive number of clients
  - balancing load, avoiding *bottlenecks* in processing and networking

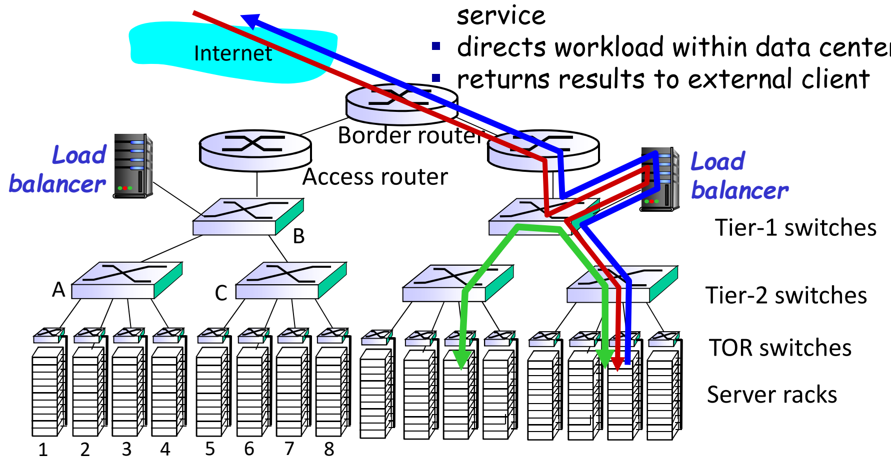
Inside a 40-ft Microsoft container, Chicago data center

Data center networks

Each LAN partitioned into smaller VLANs to localize ARP broadcast

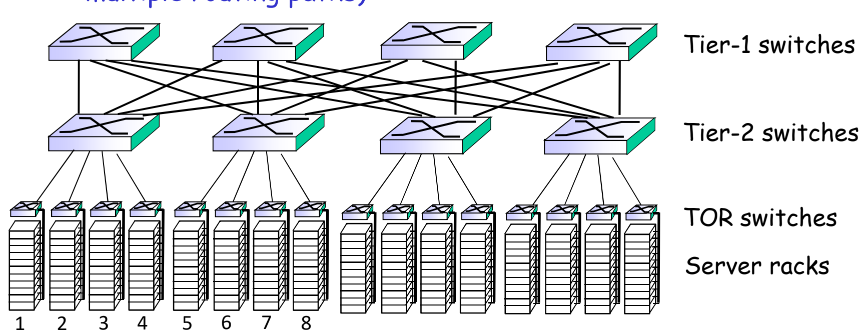
Load balancer:

- NAT functionality - hiding data center internals from outside
- receives external client requests for service
- directs workload within data center
- returns results to external client



Link layer below an access router

- Recent advances - rich interconnection among switches as well as duplication of switches:
  - increased reliability via redundancy
  - increased throughput between server racks (how to enable multiple routing paths)



focus of recent research: revisit routing for layer 2, congestion control, etc.

Chapter 6: Summary

- principles behind data link layer services:
  - error detection, correction
  - sharing a broadcast channel: multiple access
  - link layer addressing
- instantiation and implementation of various link layer technologies
  - Ethernet
  - switched LANS, VLANs
  - virtualized networks as a link layer: MPLS
  - data center networks
- synthesis: a day in the life of a web request
  (be sure to open Chapter6_A_Day_animation.ppt file in cs356/Slides folder on your own and see the animation)

The end