Network Layer:
Forwarding; Link Layer

Qiao Xiang, Congming Gao, **Qiang Su**

https://sngroup.org.cn/courses/cnns-xmuf25/index.shtml

12/2/2025

This deck of slides are heavily based on CPSC 433/533 at Yale University, by courtesy of Dr. Y. Richard Yang.

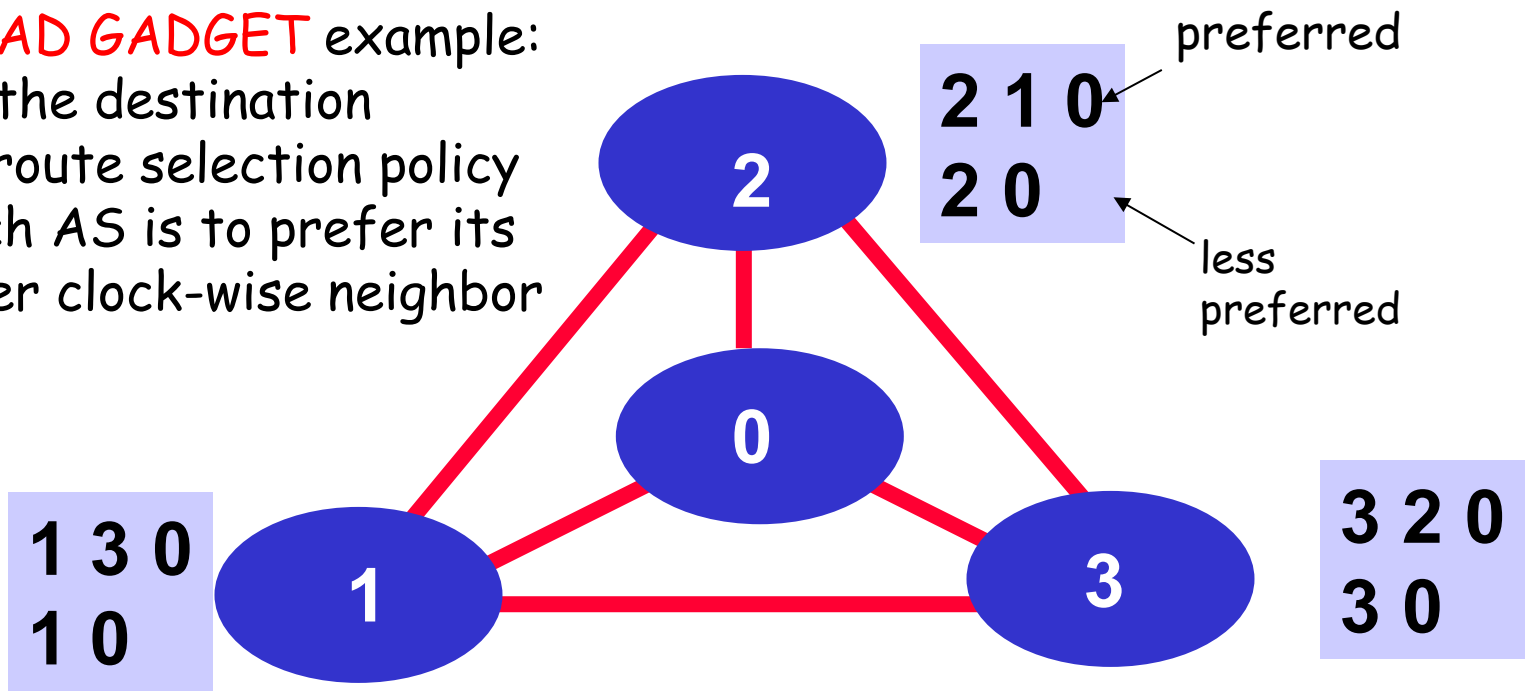
Admin

- Assignment 5: Due Dec 16, 2025

Recap: Policy Routing Stability

- A policy routing system can be considered as a system to aggregate local preferences, but aggregation may not be always successful.

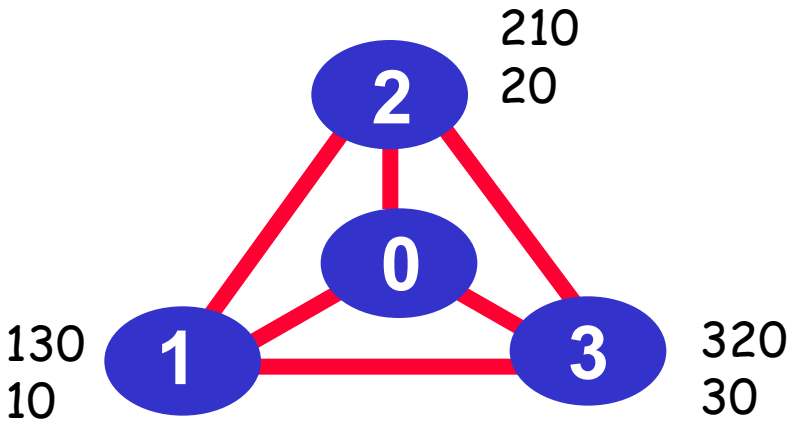
The BAD GADGET example:
- 0 is the destination
- the route selection policy of each AS is to prefer its counter clock-wise neighbor

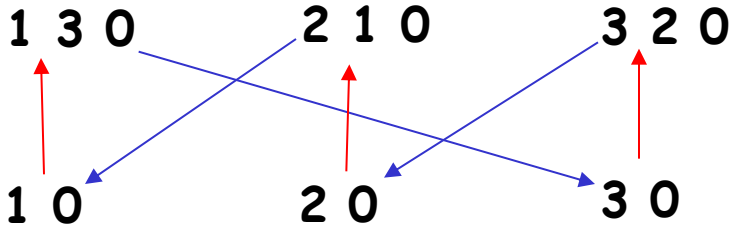


Policy (preferences) aggregation fails: routing instability !

Recap: Complete Dependency: P-Graph

- Complete dependency can be captured by a structure called P-graph
- Nodes in P-graph are feasible paths
- Edges represent priority (low to high)
  - A directed edge from path N_1P_1 to P_1
    - intuition: to let N_1 choose N_1P_1, P_1 must be chosen and exported to N_1
  - A directed edge from a lower ranked path to a higher ranked path
    - intuition: the higher ranked path should be considered first

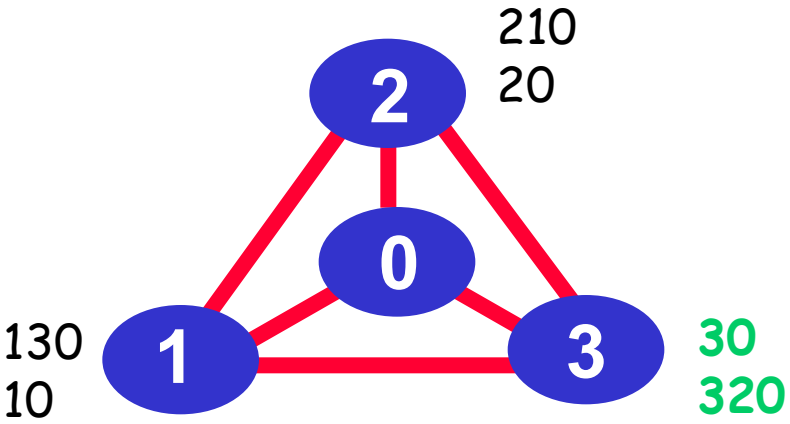


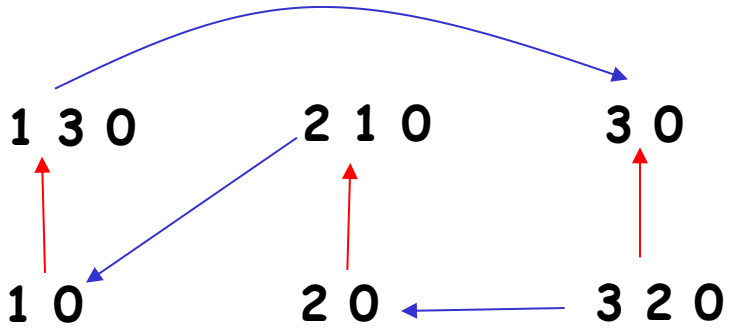


Any observation on the P-graph?

Recap: P-Graph and BGP Convergence

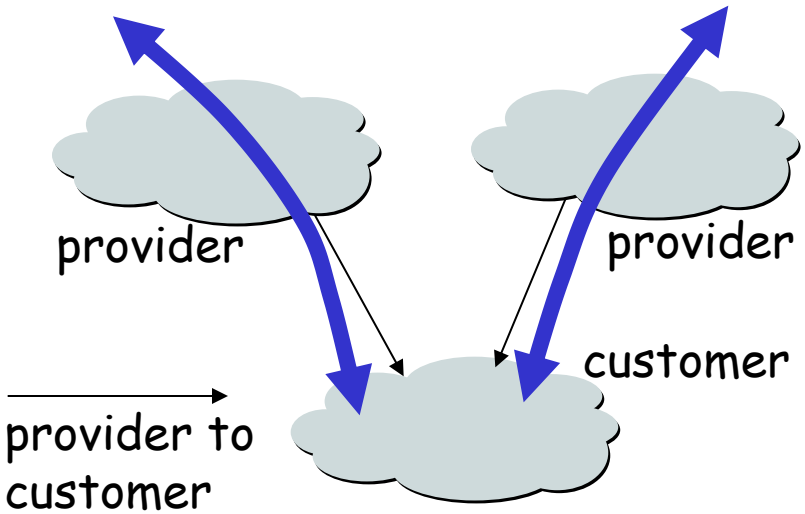
- If the P-graph of the networks has no loop, then policy routing converges.
  - intuition: choose the path node from the partial order graph with no out-going edge to non-fixed path nodes, fix the path node, eliminate all no longer feasible; continue
- Example: suppose we swap the order of 30 and 320



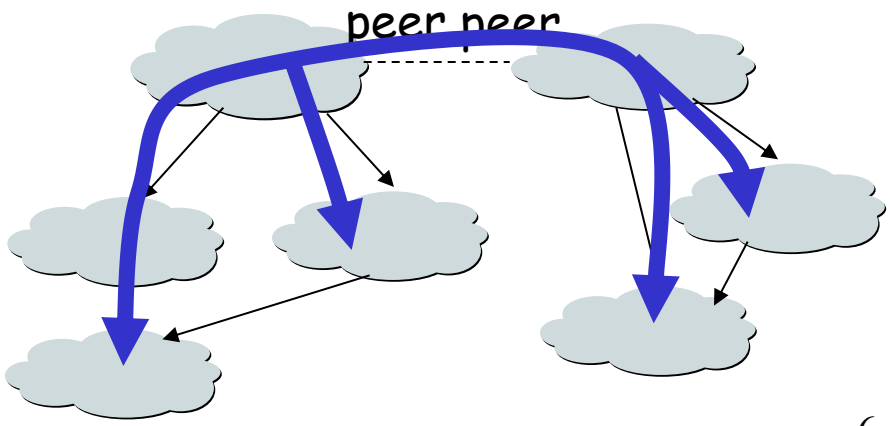


Recap: Internet Economy: Two Types of Business Relationship

- *Customer provider relationship*
  - a provider is an AS that connects the customer to the rest of the Internet
  - customer pays the provider for the transit service
  - e.g., XMU is a customer of CERNET and China Telecom

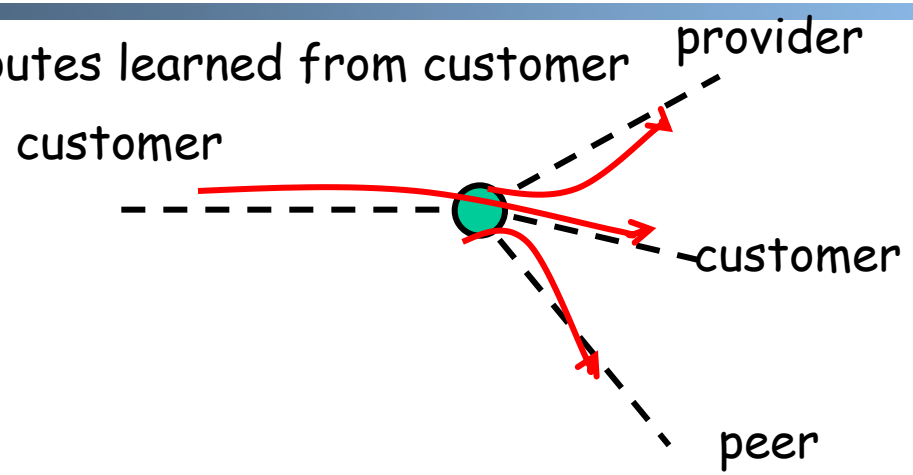


- *Peer-to-peer relationship*
  - mutually agree to exchange traffic between their respective **customers** only
  - there is no payment between peers



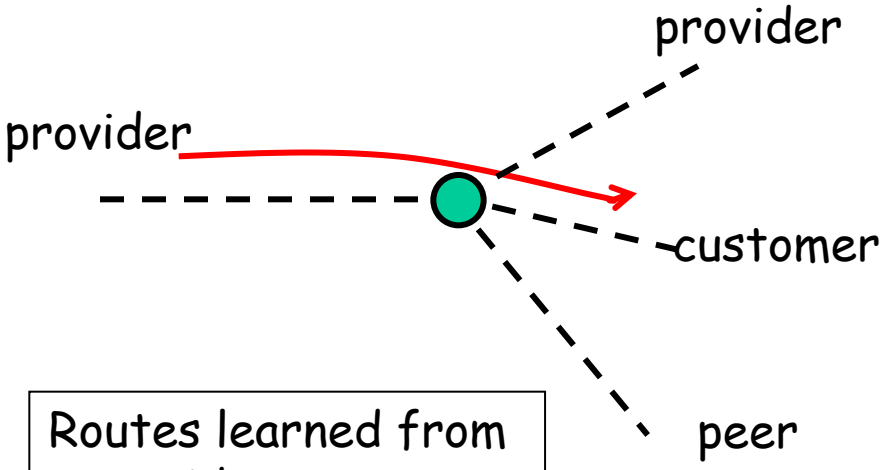
Recap: Export Policies and Economics

case 1: routes learned from customer



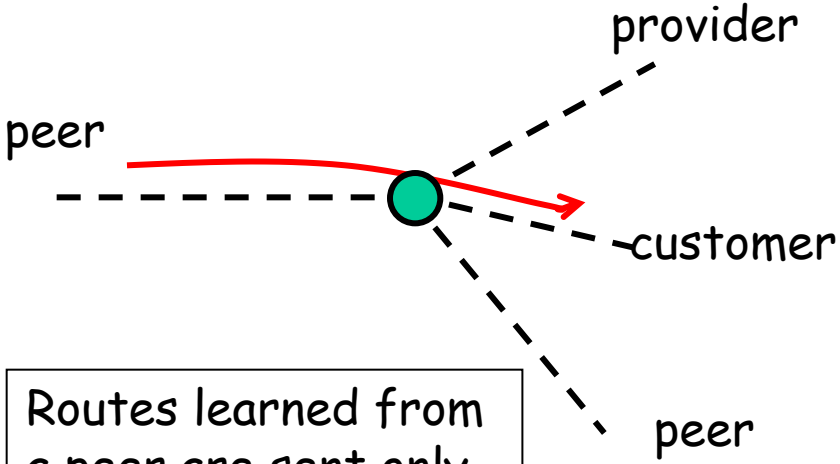
Routes learned from a customer are sent to all other neighbors

case 2: routes learned from provider



Routes learned from a provider are sent only to customers

case 3: routes learned from peer



Routes learned from a peer are sent only to customers

Outline

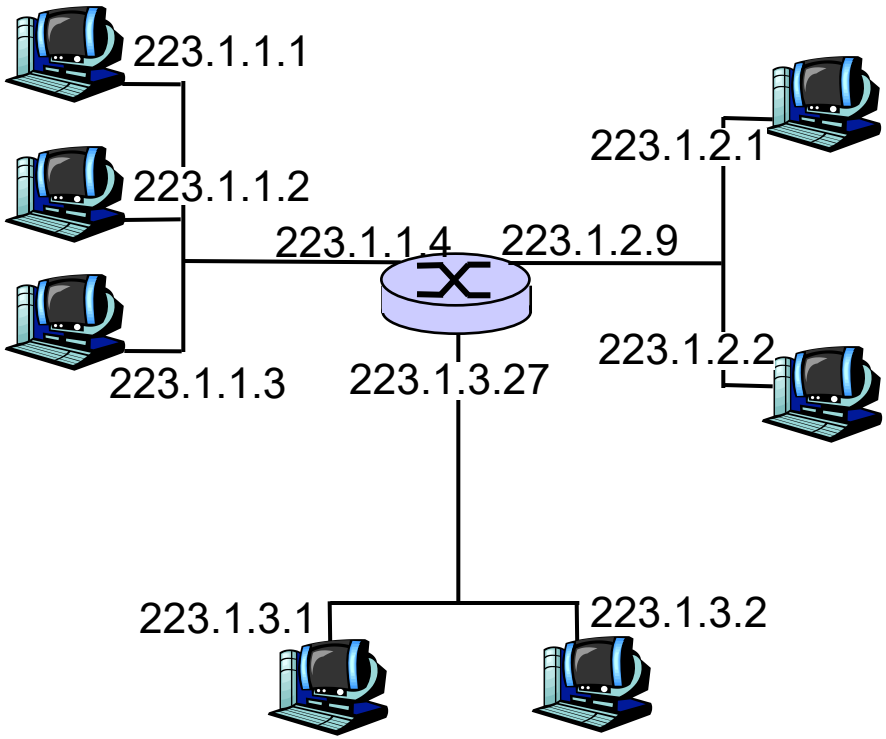
- Admin and recap
- Network control plane
  - Routing
    - Link weights assignment
    - Routing computation
      - Basic routing computation protocols
      - Global Internet routing
        - Basic architecture
        - BGP (Border Gateway Protocol): The de facto Inter-domain routing standard
          - Basic operations
          - BGP as a policy routing framework (control interdomain routes)
          - Policy/interdomain routing analysis
            - Global preference aggregation and Arrow's Theorem
            - Local preference aggregation
            - Economics and interdomain routing patterns
          - ***IP addresses for Interdomain routing***

IP Addressing Scheme: Requirements

- Uniqueness: We need an address to uniquely identify each destination
- Aggregability : Routing scalability needs flexibility in aggregation of destination addresses
  - we want to aggregate as a large set of destinations as possible in BGP announcements
- Current: the unit of routing in the Internet is a classless interdomain routing (CIDR) address

IP Address: Uniqueness

- IPv4 address: A 32-bit unique identifier for an *interface*
- *interface:*
  - routers typically have multiple interfaces
  - host may have multiple interfaces




```
%/sbin/ifconfig -a
```

223.1.3.2 = 11011111 00000001 00000011 00000010

223 1 3 2

e.g., /etc/sysconfig/network-scripts/ifcfg-enp0s25
%ifup

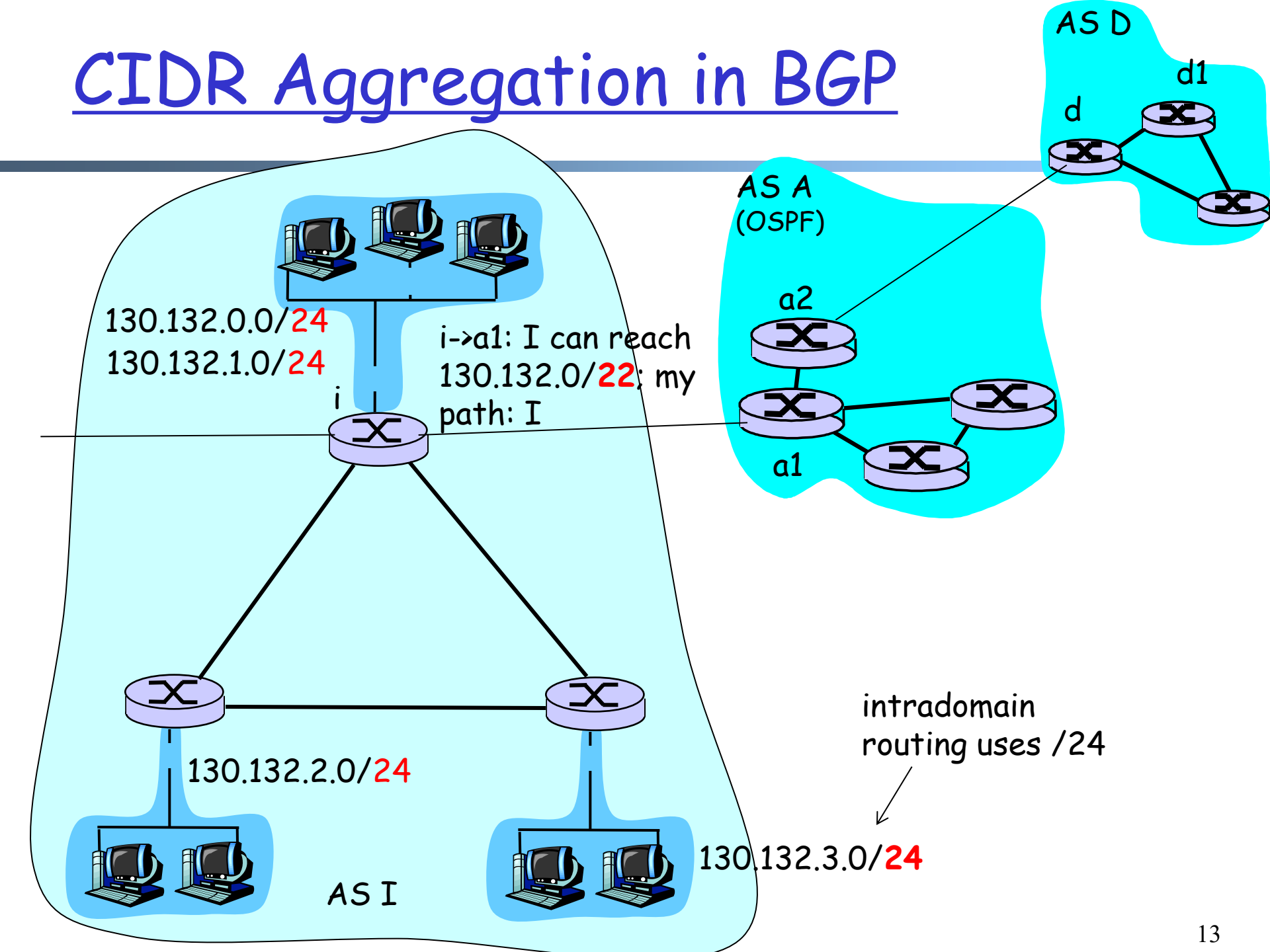
Classless InterDomain Routing (CIDR) Address: Aggregation

- A CIDR address partitions an IP address into two parts
  - A prefix representing the network portion, and the rest (host part)
  - address format: a.b.c.d/x, where x is # bits in network portion of address

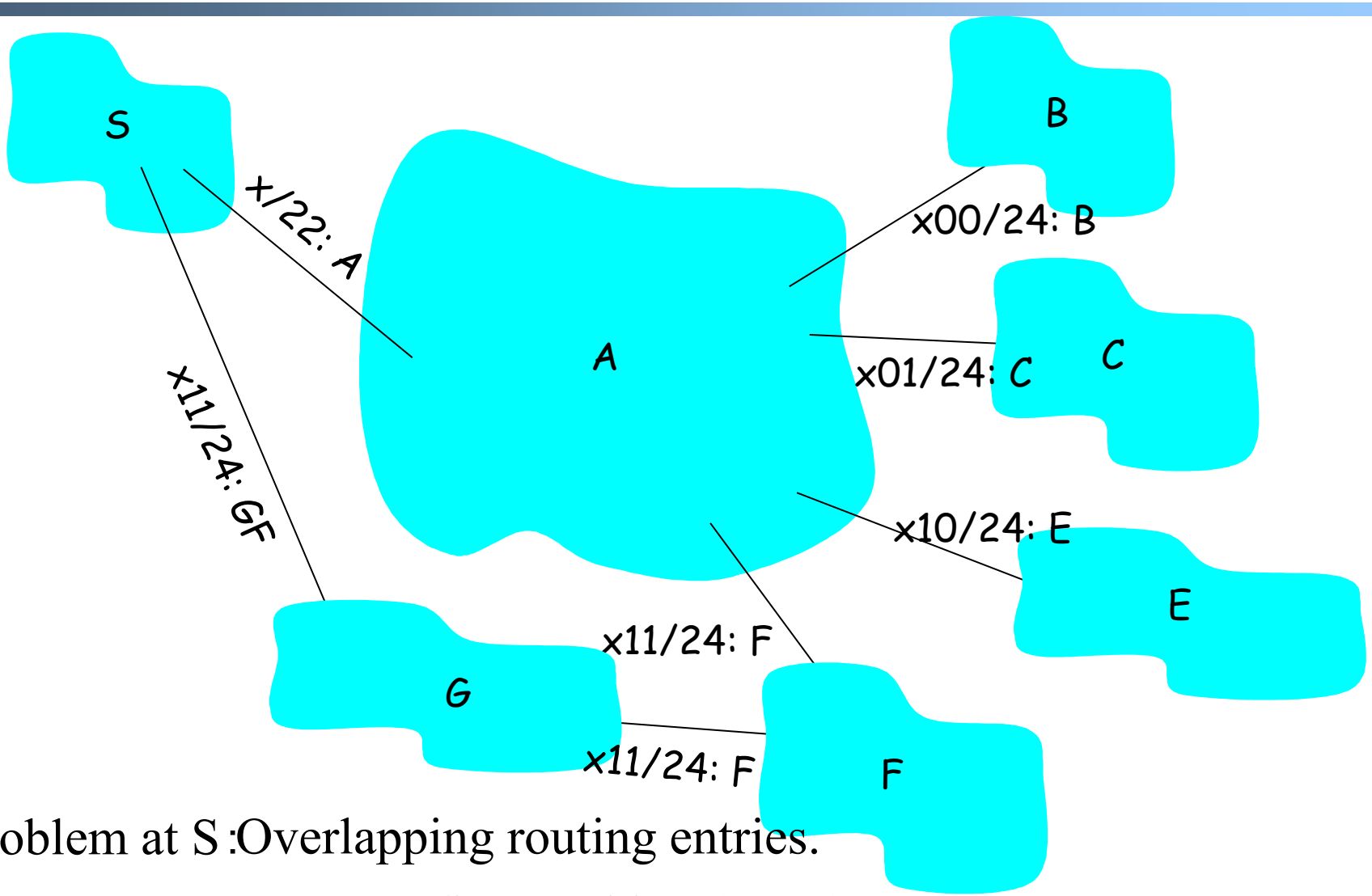
200.23.16.0/23

Some systems use mask (1's to indicate network bits), instead of the /x format

CIDR Aggregation in BGP



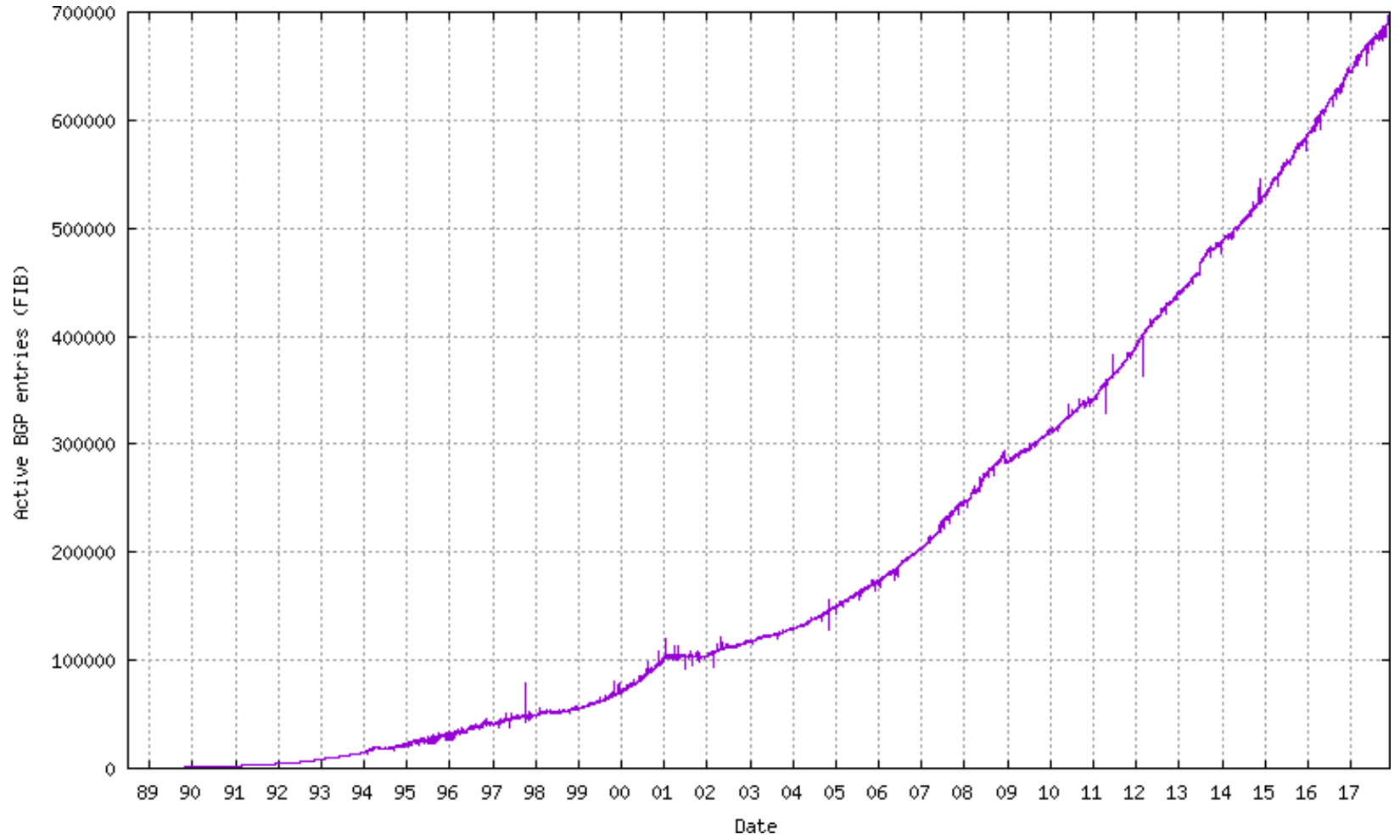
CIDR Aggregation in BGP



Problem at S :Overlapping routing entries.

Solution : Longest prefix matching (LPM)

Routing Table Size of BGP (number of globally advertised, aggregated entries)



Active BGP Entries (http://bgp.potaroo.net/as1221/bgp-active.html)
Internet Growth (http://www.caida.org/research/topology/as_core_network/historical.xml)

IP Addressing: How to Get One?

Q: How does an ISP get its block of addresses?

A: Local Internet Registry (LIR) or National Internet Registry (NIR)

https://www.iana.org/numbers

https://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xhtml

Use
%whois <IP address>
to check who is allocated the given address.

IP addresses: How to Get One?

Q: How does a *host* get an IP address?

A:

- Static configured
  - unix:
    %/sbin/ifconfig eth0 inet 192.168.0.10 netmask 255.255.255.0

- DHCP: Dynamic Host Configuration Protocol (RFC2131):
  dynamically get address from a DHCP server

DHCP Goal and History

- Goal: allow host to *dynamically* obtain its IP address from network server when it joins network
- History
  - 1984 Reverse ARP (RFC903): obtain IP address, but at link layer, and hence requires a server at each network link
  - 1985 Bootstrap Protocol (BOOTP; RFC951): introduces the concept of a relay agent to forward across networks
  - 1993 DHCP (RFC1531): based on BOOTP but can dynamically allocate and reclaim IP addresses in a pool, as well as delivery of other parameters
  - 1993 Errors in editorials led to immediate reissue as RFC1541
  - 1997 DHCP (RFC2131): add DHCPINFORM

DHCP: Dynamic Host Configuration Protocol

The often used DORA model (4 messages)

- host broadcasts "DHCP discover" msg
- DHCP server responds with "DHCP offer" msg
- host requests IP address: "DHCP request" msg
- DHCP server sends address: "DHCP ack" msg

Outline

- Admin and recap
- Network layer
  - Overview
  - Routing
  - Forwarding (put it together)

Network Forwarding: Putting it Together

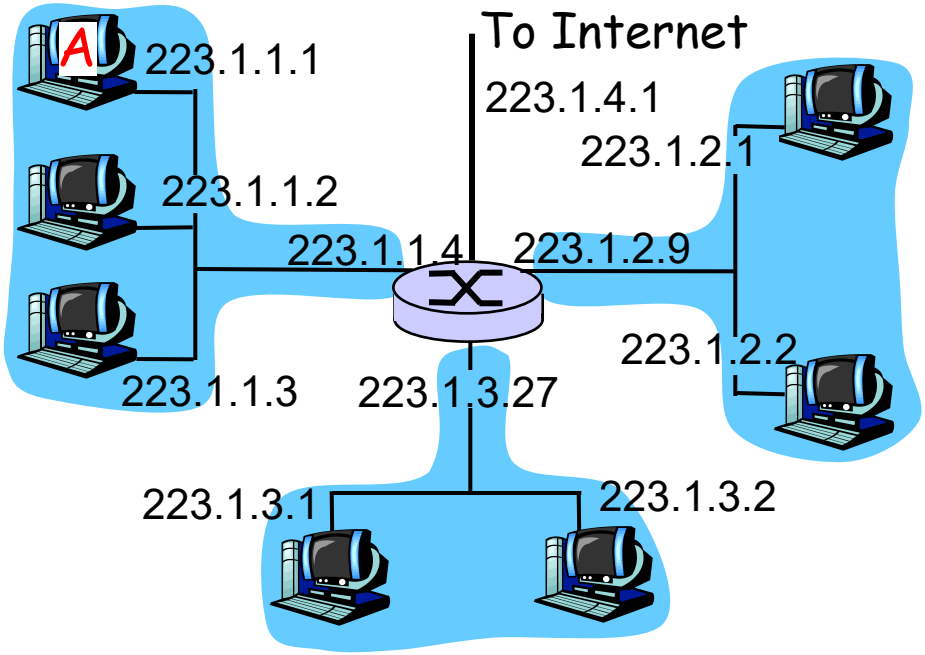
- Forwarding is also called the fast path (upon receiving each packet)

- Slow path: not per packet
  - Get IP address (DHCP, or static)
  - Setup/compute routing table

Forwarding: Example 1

	src	dst	
misc fields	223.1.1.1	223.1.1.3	data

- Setting: Host A network layer receives a packet above.
- Action:
  - Host A looks up destination in routing table
    - Exercise: Suppose A uses DHCP to obtain its address, how can A construct its routing table (routing information base, RIB)?



Host Routing Table Example: my Mac

- Mac
  - ifconfig –a
  - netstat –rn (man netstat to see description)

Routing tables

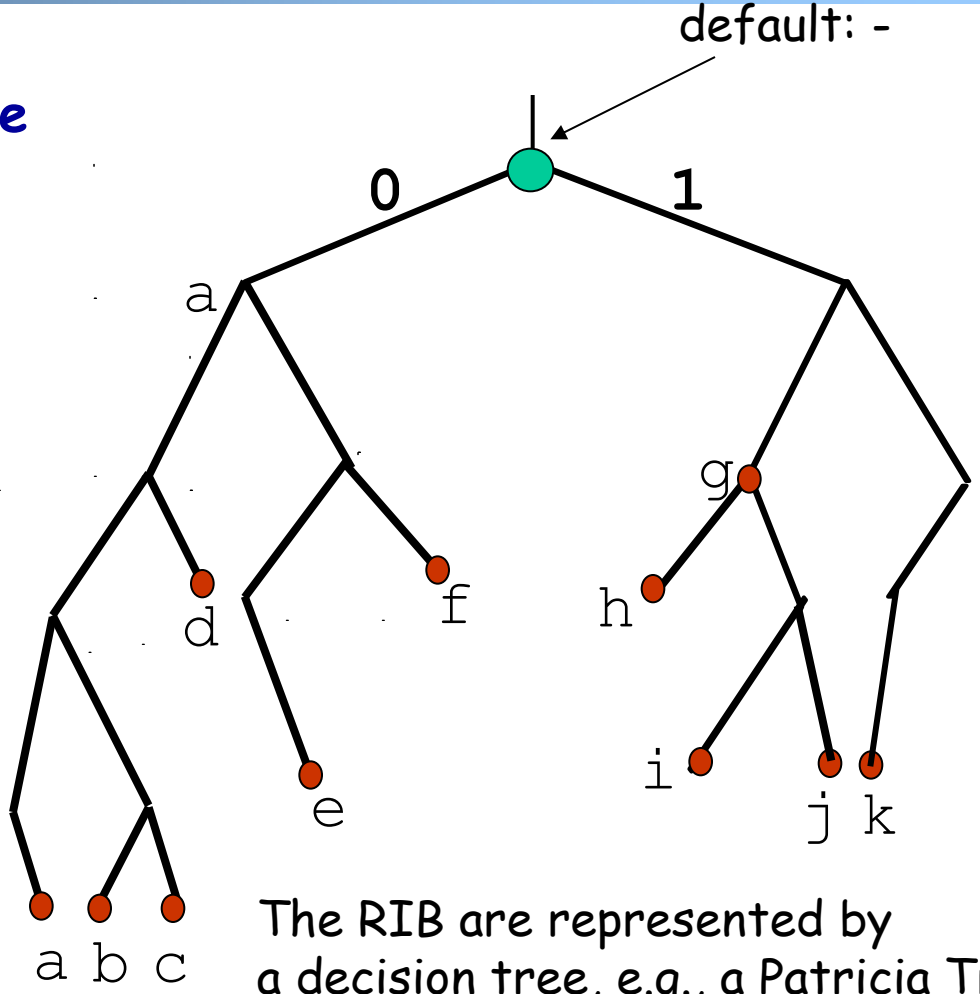
Internet:

Destination	Gateway	Flags	Refs	Use	Netif	Expire
default	172.27.16.1	UGSc	1470	0	en0	
127	127.0.0.1	UCS	1	0	lo0	
127.0.0.1	127.0.0.1	UH	4	3788	lo0	
169.254	link#4	UCS	106	0	en0	
169.254.1.229	link#4	UHLSW	1	0	en0	
169.254.5.209	f0:99:bf:1e:6f:de	UHLSW	1	0	en0	989
169.254.8.254	link#4	UHLSW	1	0	en0	
169.254.11.96	0:cd:fe:75:59:75	UHLSW	1	0	en0	1009
169.254.13.89	64:9a:be:af:34:53	UHLSW	1	0	en0	1145
169.254.16.49	link#4	UHLSW	1	0	en0	
169.254.19.58	link#4	UHLSW	1	0	en0	
169.254.19.82	link#4	UHLSW	1	0	en0	
169.254.21.198	link#4	UHLSW	1	0	en0	
169.254.22.67	0:23:12:12:bc:39	UHLSW	1	0	en0	31
169.254.23.4	link#4	UHLSW	1	0	en0	

…

CIDR Forwarding Look Up: Software

#	prefix	interface
a)	00001	
b)	00010	
c)	00011	
d)	001	
e)	0101	
f)	011	
g)	10	
h)	100	
i)	1010	
j)	1011	
k)	1100	



The RIB are represented by a decision tree, e.g., a Patricia Trie to look for the longest match of the destination address

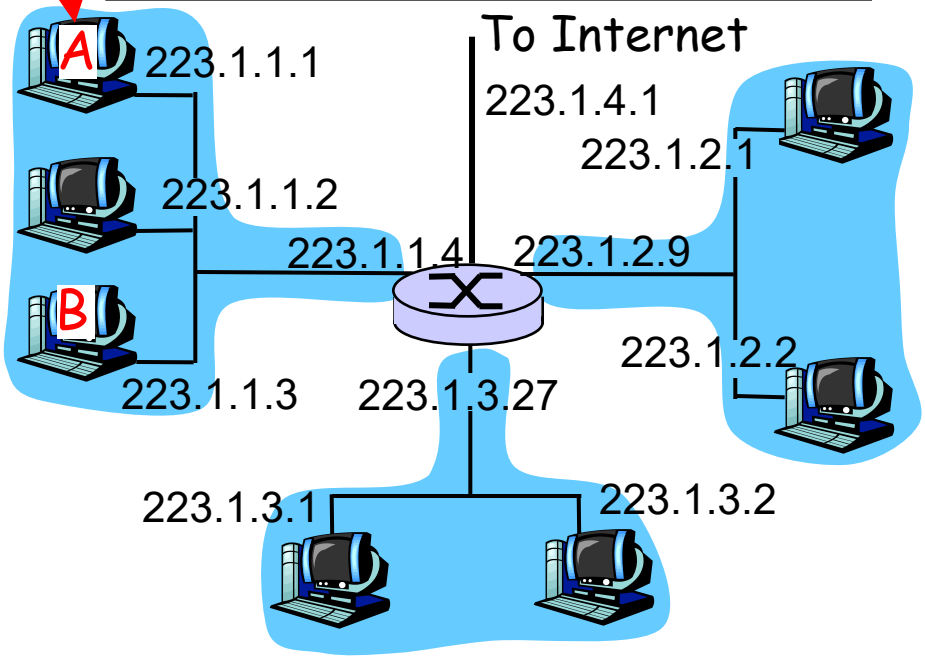
Putting it Together: Example 1: A->B

	src	dst	
misc fields	223.1.1.1	223.1.1.3	data

- Setting: Host A network layer receives a packet above.
- Action:
  - Host A looks up destination in routing table (on same subnet)
  - Hand datagram to link layer to send inside a link-layer frame
    - Key step: need to map B's IP address 223.1.1.3 to B's MAC address

forwarding table in A

Dest. Net.	next router	Nhops
223.1.1/24		1
223.1.2/24	223.1.1.4	2
223.1.3/24	223.1.1.4	2
0.0.0.0/0	223.1.1.4	-



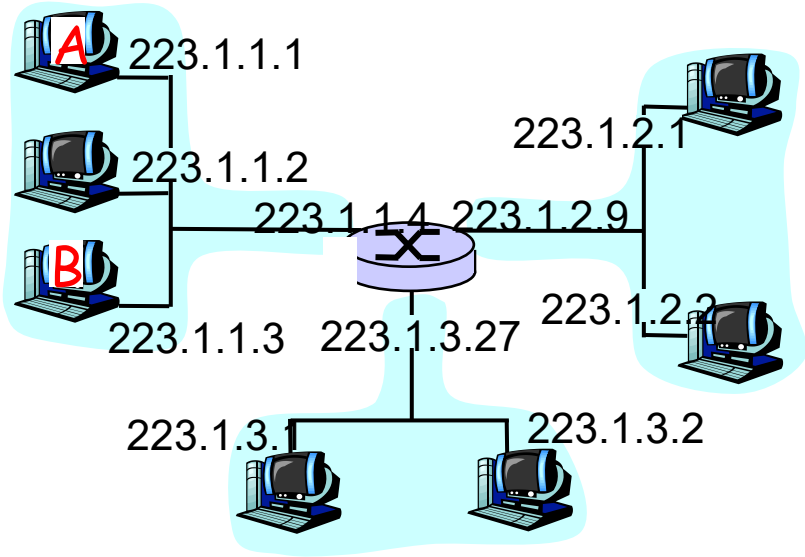
Comparison of IP address and MAC Address

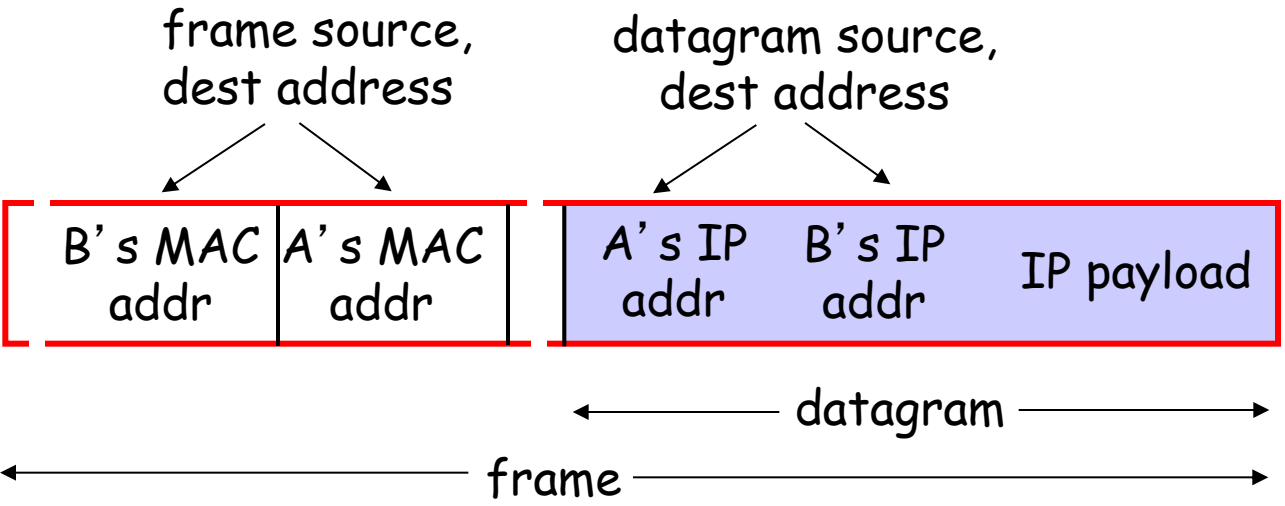
- IP address is locator
  - address depends on network to which an interface is attached
    - NOT portable
  - introduces features (e.g., CIDR) for routing scalability
- IP address needs to be globally unique (if no NAT)

- MAC address is an identifier
  - dedicated to a device
    - portable
  - flat
- MAC address does not need to be globally unique, but the current assignment ensures uniqueness

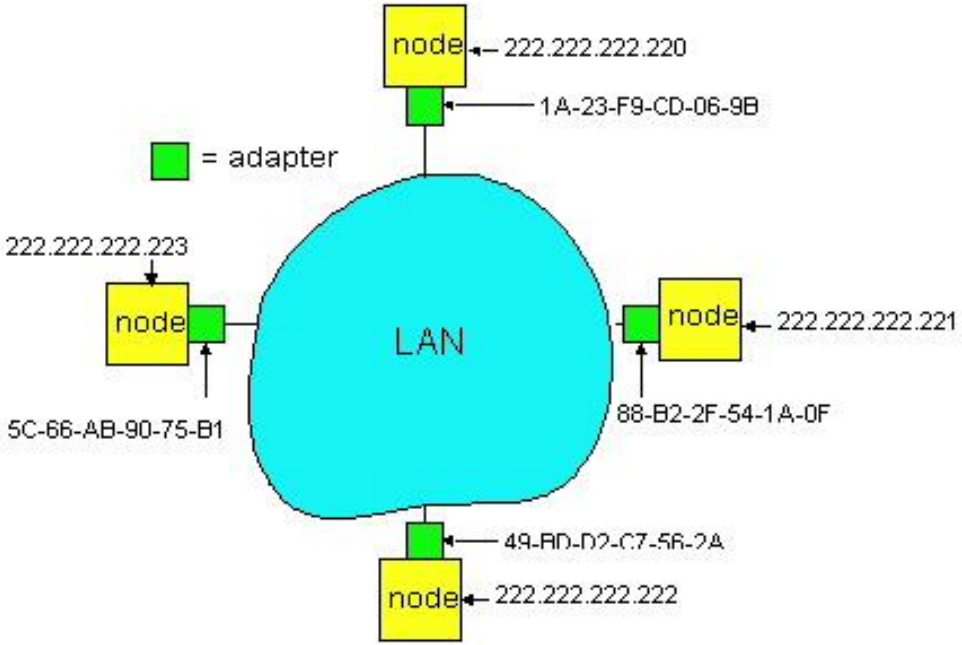
Issue

A finds the MAC address of B to construct





Recall: Address Resolution Table



- 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)

```
[yry3@cicada yry3]$ /sbin/arp
Address                  HWtype  HWaddress           Flags Mask            Iface
zoo-gatew.cs.yale.edu    ether   AA:00:04:00:20:D4   C                     eth0
artemis.zoo.cs.yale.edu  ether   00:06:5B:3F:6E:21   C                     eth0
lab.zoo.cs.yale.edu      ether   00:B0:D0:F3:C7:A5   C                     eth0
```

Recall: ARP Protocol

- ARP table by the ARP Protocol, which is a "plug-and-play" protocol
  - nodes create their ARP tables without intervention from net administrator

- A broadcast protocol:
  - source broadcasts query frame, containing queried IP address
    - all machines on LAN receive ARP query
  - destination D receives ARP frame, replies
    - frame sent to A's MAC address (unicast)

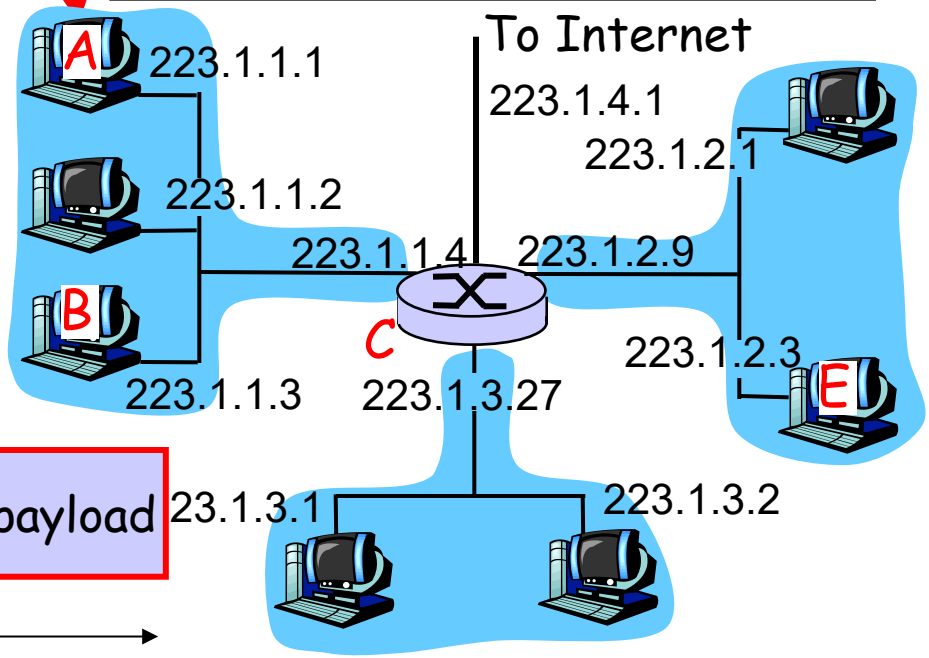
Putting it Together: Example 2 (Different Networks): A-> E

misc fields	223.1.1.1	223.1.2.3	data

- Setting: Host A network layer receives a packet above.
- Action:
  - Host A looks up destination in routing table
    - Find next hop should be 223.1.1.4
  - Hand datagram to link layer to send inside a link-layer frame

forwarding table in A

Dest. Net.	next router	Nhops
223.1.1/24		1
223.1.2/24	223.1.1.4	2
223.1.3/24	223.1.1.4	2
0.0.0.0/0	223.1.1.4	-



frame dst, src addr

datagram source, dest address

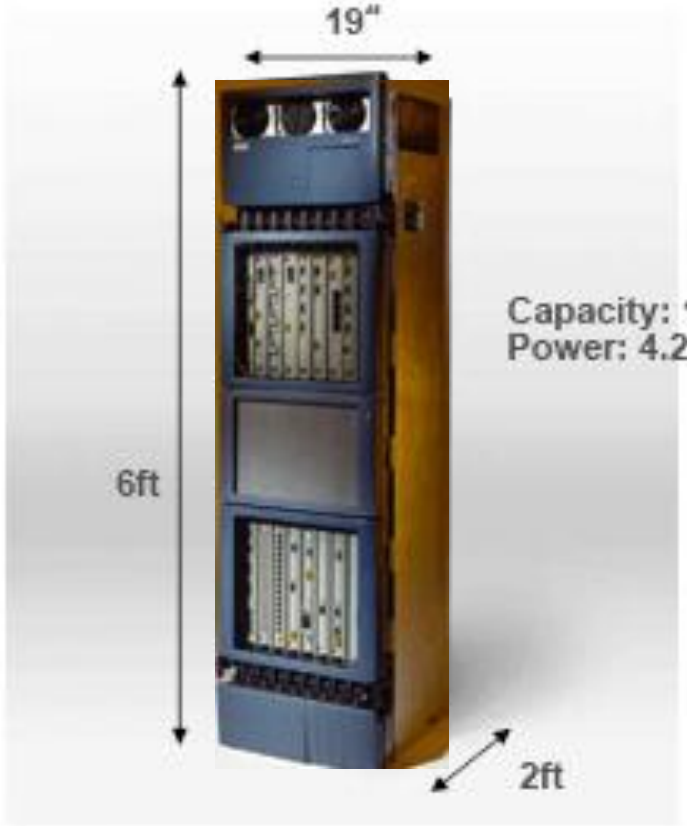
C's MAC addr	A's MAC addr	A's IP addr	E's IP addr	IP payload

datagram

frame

What A Router Looks Like: Outside

Cisco GSR 12416

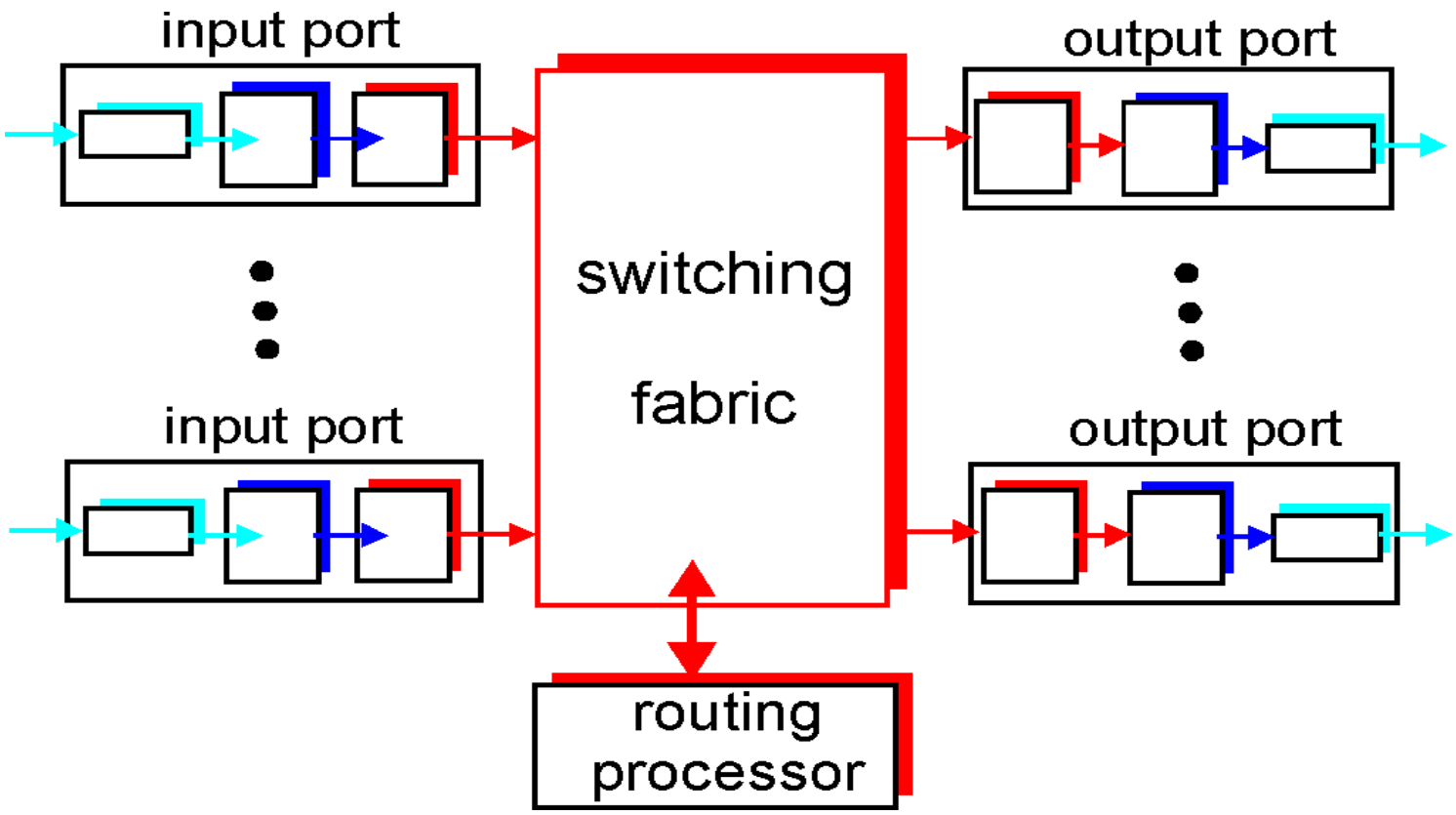


Capacity: 160Gb/s
Power: 4.2kW

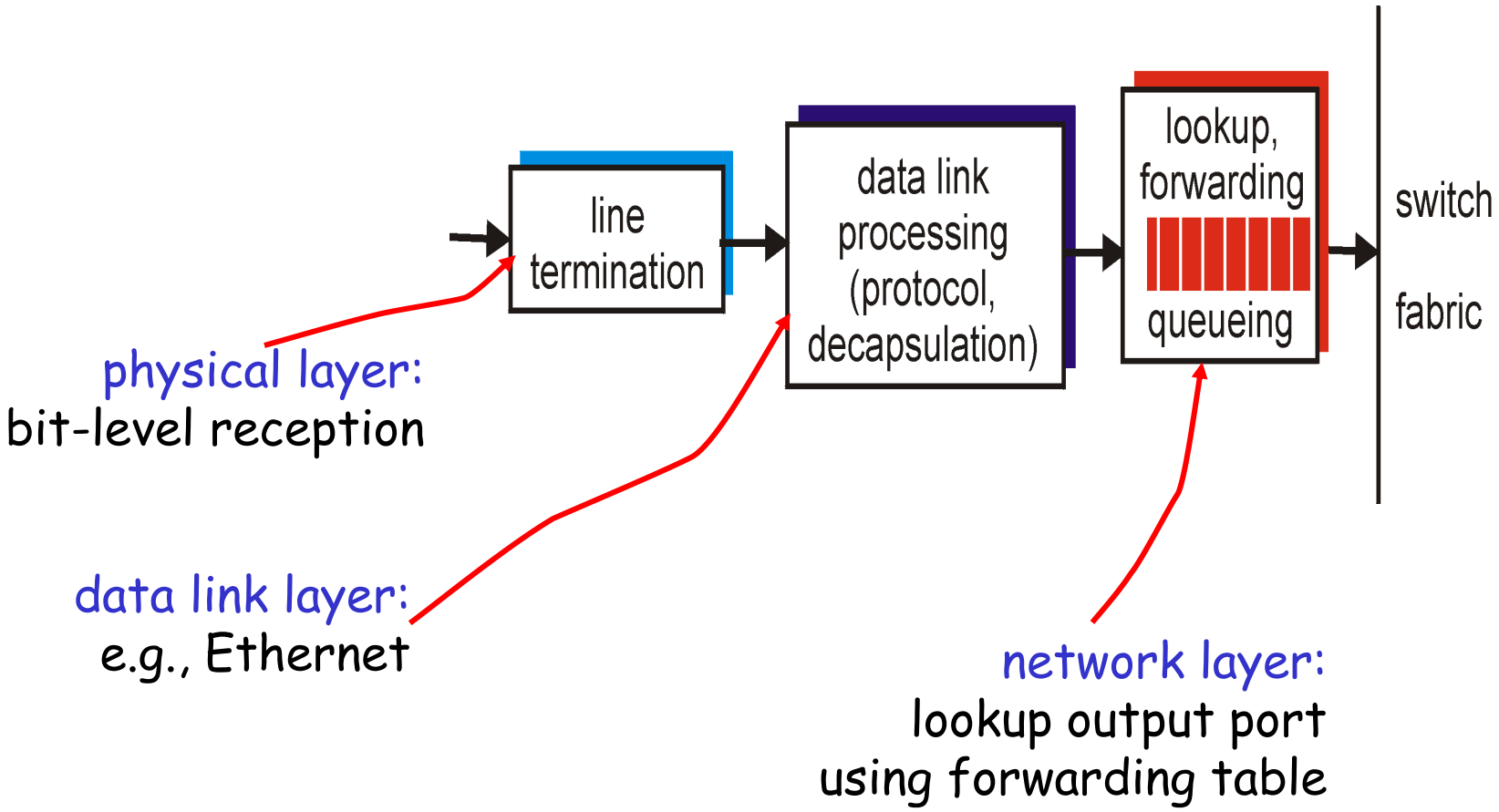
Juniper M160

Capacity: 80Gb/s
Power: 2.6kW

Look Inside a Router

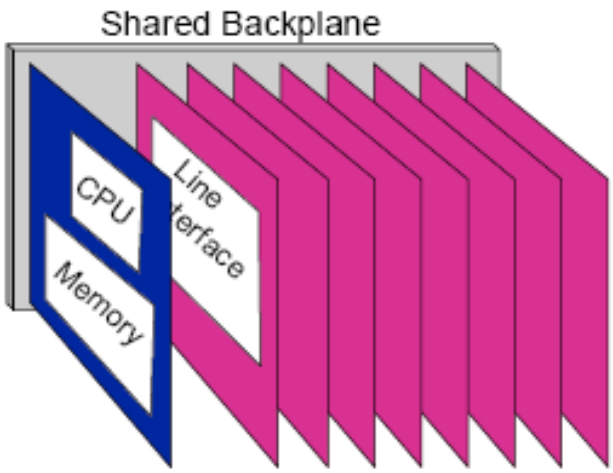


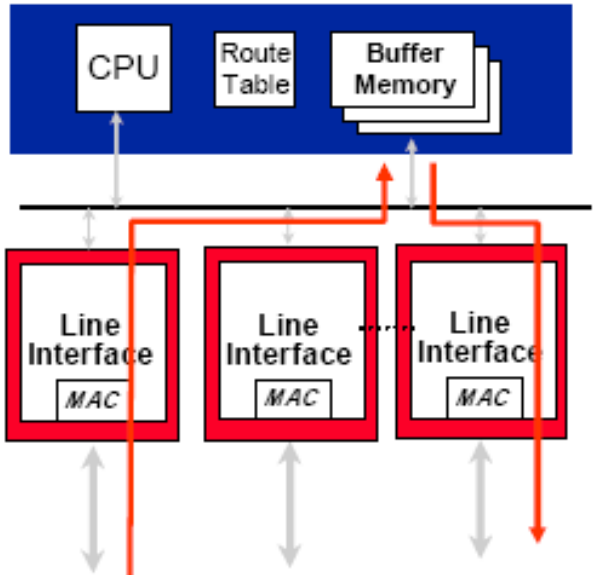
Look Inside a Router: Input Port



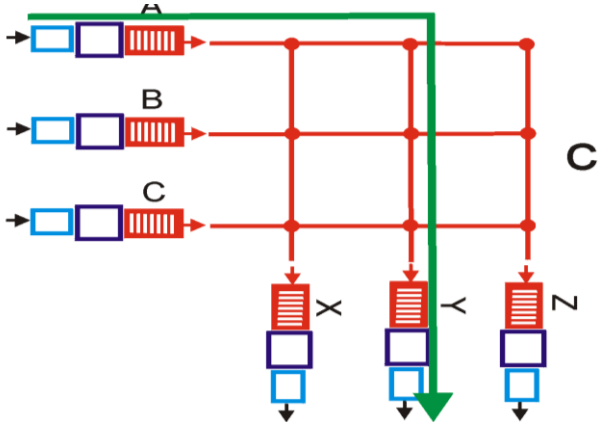
Look Inside a Router: Switching Fabric

Low End

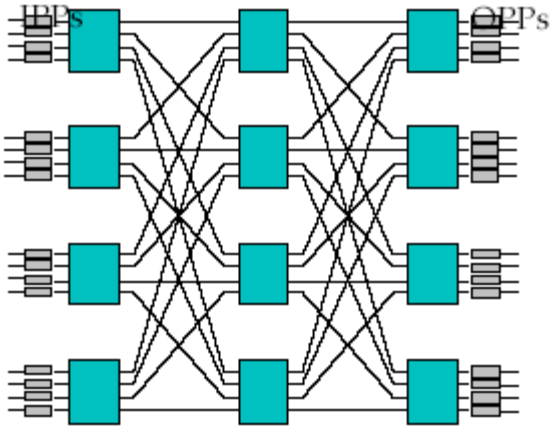


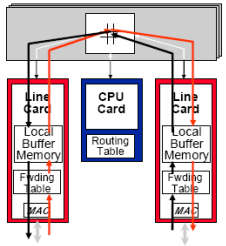


High End



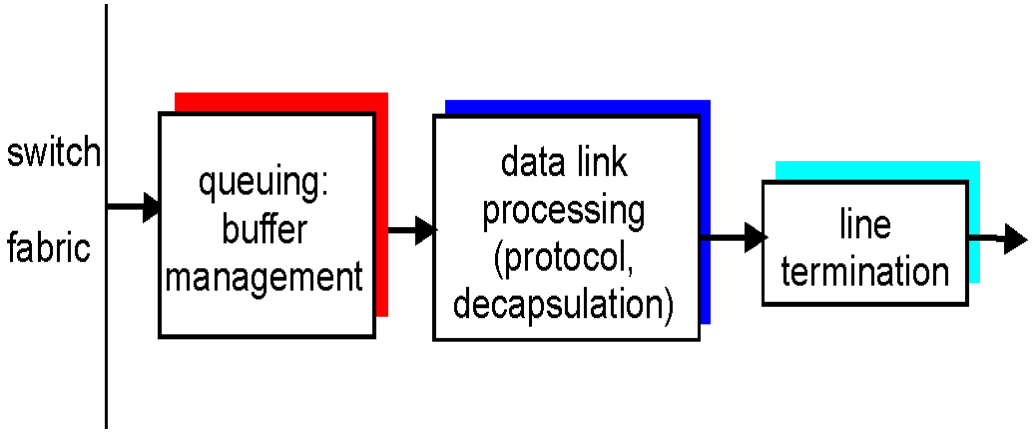
crossbar





Banyan

Look Inside a Router: Output Port



- *Buffering* required when datagrams arrive from fabric faster than the transmission rate
- *Queueing (delay) and loss due to output port buffer overflow !*
- *Scheduling and queue/buffer management* choose among queued datagrams for transmission

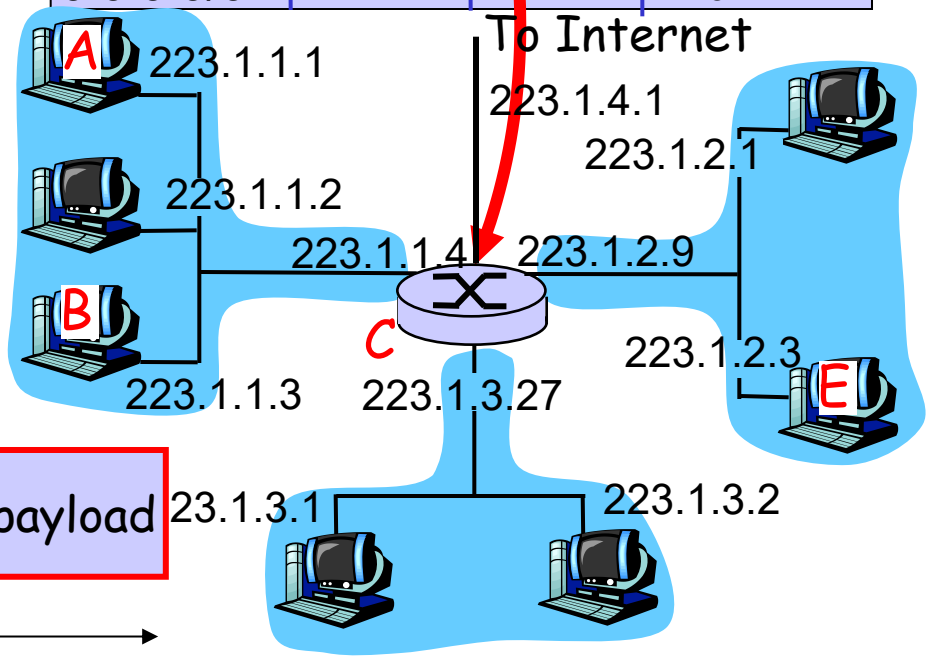
Putting it Together: Example 2 (Different Networks): A-> E

misc fields	223.1.1.1	223.1.2.3	data

- Setting: Packet above arrives at Router C's network layer.
- Action:
  - Router C conducts standard router actions
    - Assume packet correct, find next hop should be 223.1.2.9
  - Hand datagram to link layer to send inside a link-layer frame

forwarding table in router

Dest. Net	router	Nhops	interface
223.1.1/24	-	1	223.1.1.4
223.1.2/24	-	1	223.1.2.9
223.1.3/24	-	1	223.1.3.27
0.0.0.0/0	-	-	223.1.4.1



frame dst, src addr

datagram source, dest address

E's MAC addr	C's MAC addr	A's IP addr	E's IP addr	IP payload

datagram

frame

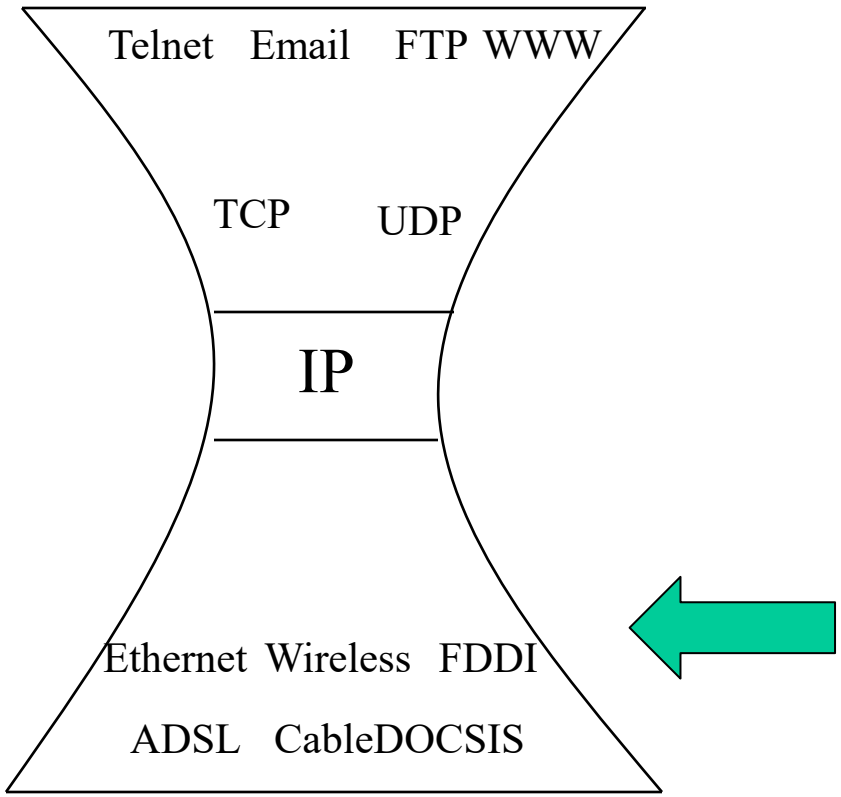
datagram arrives at 223.1.2.3!! (hooray!)

Summary of Network Layer

- We have covered the very basics of the network layer
  - routing and basic forwarding

- There are multiple other topics that we did not cover
  - Multicast/anycast
  - QoS
  - slides as backup
    just in case you need
    reading in the winter

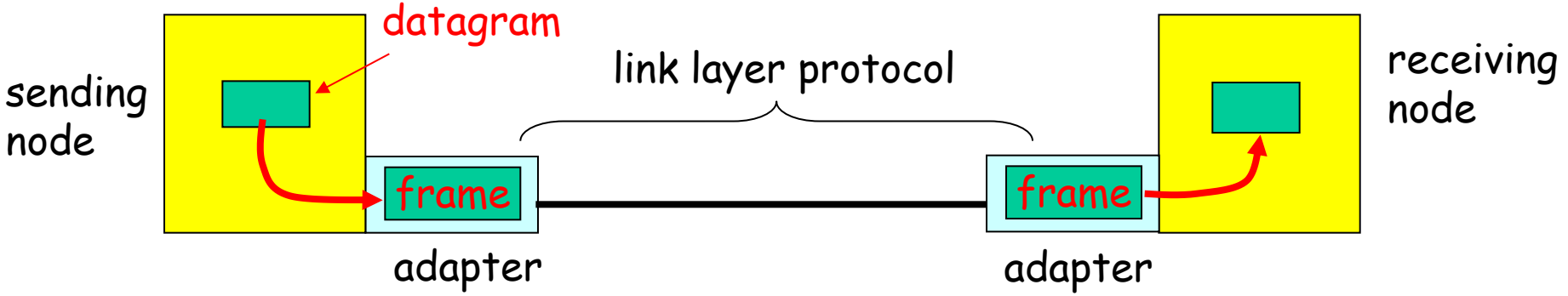
Roadmap: The Hourglass Architecture of the Internet



Link Layer Services

- Framing
  - encapsulate datagram into frame, adding header, trailer and error detection/correction
- Multiplexing/demultiplexing
  - frame headers to identify src, dest
- Reliable delivery between adjacent nodes
  - we learned how to do this already !
  - seldom used on low bit error link (fiber, some twisted pair)
  - common for wireless links: high error rates
- Media access control
- Forwarding/switching with a link-layer (Layer 2) domain

Adaptors Communicating



- link layer typically implemented in "adaptor" (aka NIC)
  - Ethernet card, modem, 802.11 card, cloud virtual switch
- adapter is semi-autonomous, implementing link & physical layers

- in most link-layer, each adapter has a unique link layer address (also called MAC address)