# E6998-02: Internet Routing

# Lecture 7
# Unix Forwarding and Routing

John Ioannidis

AT&T Labs – Research

`ji+ir@cs.columbia.edu`

# Announcements

Email: `<ji+ir@cs.columbia.edu>`

- Mail to anything else will not be answered.

Class web page: `http://www.cs.columbia.edu/~ji/F03/`

- Check frequently!
- Slides will be available there.
- As will additional reading material (papers, RFCs, source code, man pages, etc.).

Class BBoard: `coms6998-002-033@columbia.edu` (to post), or `https://www1.columbia.edu/sec/bboard/033/coms6998-002/`

Office hours: MW 15:00-16:00 in 464 CSC.

TA(s): Angelos Stavrou `<angel@cs.columbia.edu>`

TA office hours: TR 13:00-14:00 in the Mudd TA room.

# Summary of Lecture 3

- Address aggregation.
- Special addresses.
- Neighbor discovery.
- Router discovery.
- Multihoming.

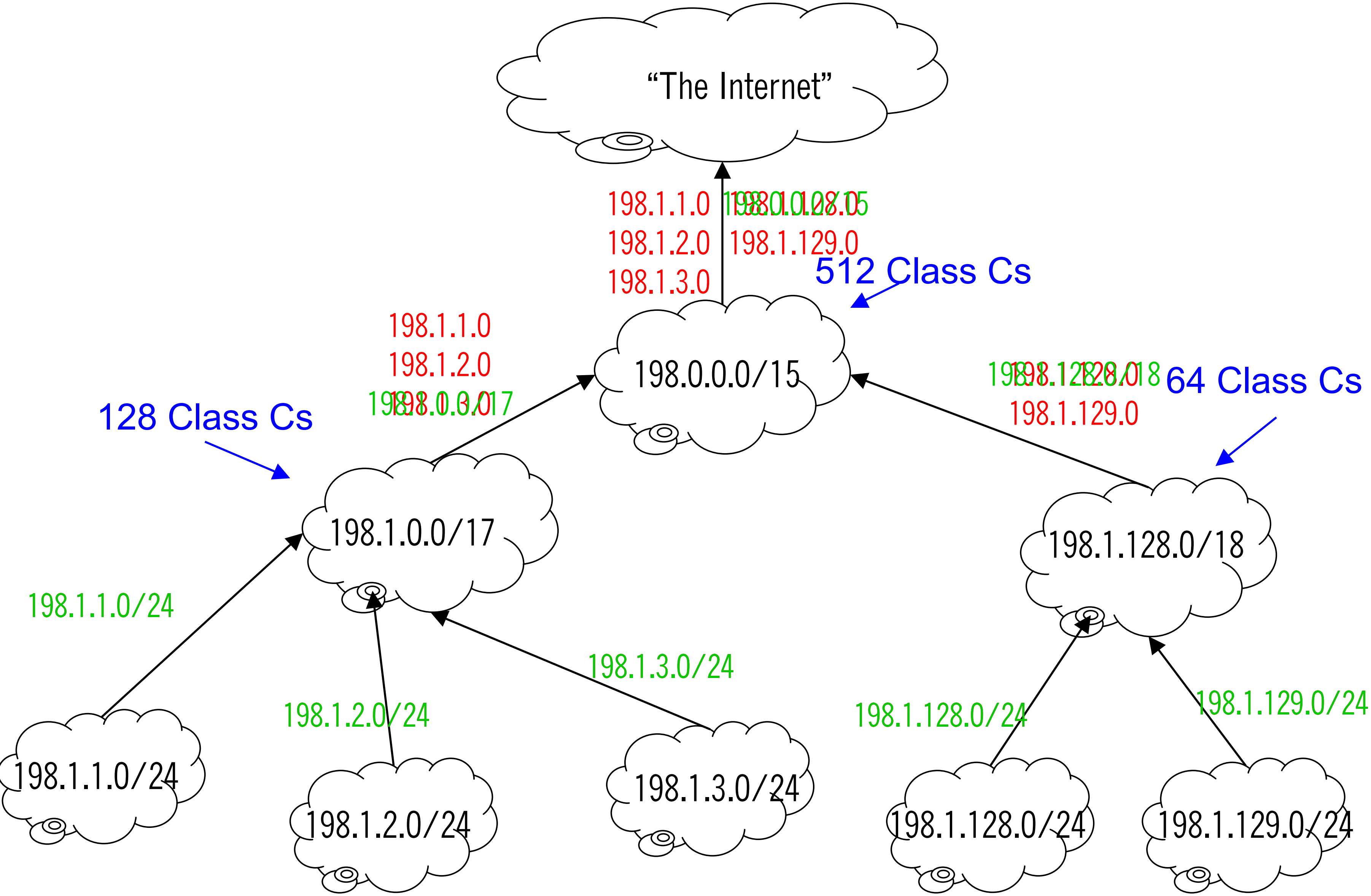
# Classless Interdomain Routing (CIDR)

- “Supernetting” (opposite of subnetting).
- Get rid of classes A/B/C.
- Give addresses in terms of prefixes.
- Netmask MUST have contiguous 1s, then contiguous 0s.
- Allows sites to be allocated the proper size of network.
- Allows ISPs to aggregate addresses of clients.
  - Reducing routing table size.
- “CIDR block” or “CIDR prefix”.

# CIDR Address Allocation

- Pre-CIDR allocations still routed, of course.
- ARIN/RIPE/APNIC have large allocations (/8s) to hand out.
- ISPs get addresses in large blocks from the Registries.
- Allocate chunks of these blocks to customers.
  - "Non-portable" address space: change ISP, change addresses.
  - Aggregation of addresses within ISP.

# Classful vs. CIDR Announcements



# Special Addresses

- 127.0.0.0/8 (“loopback”). Usually 127.0.0.1 on hosts.
- net.0 is “any host this subnet” (form of “anycast”).
- net.-1 is “all hosts this subnet” (directed broadcast).
- 255.255.255.255 is “local broadcast”.
- Multicast (224.0.0.0/4). Class E (240.0.0.0/4) still reserved.
- RFC1918 addresses (“site local”, “private-use”).
  - 10.0.0.0/8, 172.16.0.0/12, 192.168.0.0/16.
  - MUST NOT be routed outside an organization.
  - Used for NAT.
- draft-ietf-zeroconf-ipv4-linklocal-09.txt (“link local”)
  - 169.254.0.0/16.
  - MUST NOT be forwarded by a router (OK to bridge).
  - Used by auto-configuration process.

# Unicast, Anycast, Multicast, Broadcast

- RFC 791 does not mention any of these terms.
- Broadcast & Multicast originally Ethernet (etc.) notions.
- (net,-1) addresses are IP directed broadcasts.
  - "All hosts this subnet".
  - Routed normally until last subnet.
  - Then sent to all-ones MAC address (no ARP involved).
  - "Outside" directed broadcasts ("splattergrams") usually filtered at last-hop router or just answered by it.
- (net,0) addresses are IP directed anycasts.
  - "Any host this subnet".
  - Routed normally until last subnet.
  - Usually answered by last-hop router.
  - Router may know who the responsible host(s) are.
  - Not much use when sent in same subnet.

# {Uni,Any,Multi,Broad}cast cont'd

- All-ones address (255.255.255.255, "limited broadcast".
  - Stays in subnet.
  - "All hosts this subnet".
- All-zeroes address (0.0.0.0, "unspecified", INADDR_ANY).
  - As source, replaced by outgoing interface address.
  - As destination, same as loopback.
- IP Multicast (224.0.0.0/4).
  - On "target" subnet turned into Ethernet multicast.
  - Can be routed (we'll talk about this later).
- IP Anycast (not (net,0)).
  - Any address can be deemed anycast.
  - Routers determine what is anycast.
  - Suggested for critical services use (e.g., root DNS servers).

# IPv6 Addresses

- 128 bits.
- Representation (RFC2373, RFC1924):
  - Eight groups of four hex digits separated by colons.
  - Leading zeros dropped.
  - One contiguous set of 0000s replaced with ::
  - fe80:0000:0000:0000:280:c8ff:feca:a27b is the same as fe80::280:c8ff:feca:a27b.
  - ::1 is “loopback”, :: is “unspecified”.
  - Also, ::ffff:192.20.13.4
- 2000::/3 (addresses starting with the bits 001) are aggregatable addresses.
- Read RFC2373!

# Forwarding

- How to send an IP packet to a host on the same subnet?
  - "Same subnet" means equal subnet prefix (and different host part).
  - if ((src & netmask) == (dst & netmask)) { …
  - Find MAC address of destination (if not on p2p link).
  - Send packet.
- How to send an IP packet to a host on different subnet?
  - … } else { …
  - Find MAC address of appropriate router.
    - Have to know who the router is.
    - Entry in forwarding table.
  - Send packet.
  - Eventually a router attached to the dst subnet will get the packet.

# ARP

- Local (same subnet) forwarding.
- Address Resolution Protocol, RFC826
- Maps IPv4 addresses to MAC addresses.
- Ethertype 0x0806.
- Man pages: arp(4), arp(8)

```
03:10:59.738069 0:1:2:72:bd:3e ff:ff:ff:ff:ff:ff 0806 42: arp who-has 135.207.25.192 tell 135.207.25.36
                         0001 0800 0604 0001 0001 0272 bd3e 87cf
                         1924 0000 0000 0000 87cf 19c0
03:10:59.738190 0:e0:81:10:4b:64 0:1:2:72:bd:3e 0806 60: arp reply 135.207.25.192 is-at 0:e0:81:10:4b:64
                         0001 0800 0604 0002 00e0 8110 4b64 87cf
                         19c0 0001 0272 bd3e 87cf 1924 0000 0000
                         0000 0000 0000 0000 0000 0000 0000
```

# Gratuitous ARP, Proxy-ARP, RARP,

- When an interface comes up, it sends a “gratuitous ARP”.
  - Other stations update their ARP cache.
  - Can detect duplicate IP addresses.
- Proxy-ARP: poor man’s subnetting/routing.
  - Used to “subnet” on non-bit boundaries.
- RARP (Reverse ARP, ethertype 0x8035).
  - Used by booting station to find its IP address from its MAC address.
  - Needs a server.
  - How to get a station to report its IP address given its MAC address?

# NDP

- Neighbor Discovery Protocol.
- IPv6 ARP-equivalent.
- Uses UDP Multicast.
  - (ARP predates Multicast).
- RFC2461.

# Router Discovery

- (For hosts).
- Configured with a command:
  - `route add 135.207.4.0/24 135.207.25.36`
  - `route add default 135.207.31.1`
  - Default is the same as 0/0.
- Configured with DHCP/BOOTP at boot time.
- Simple routing protocol (e.g., RIP) used to announce routes.
- There is an ICMP message for router discovery (not used).

- IPv6: Router solicitation, also multicast based.

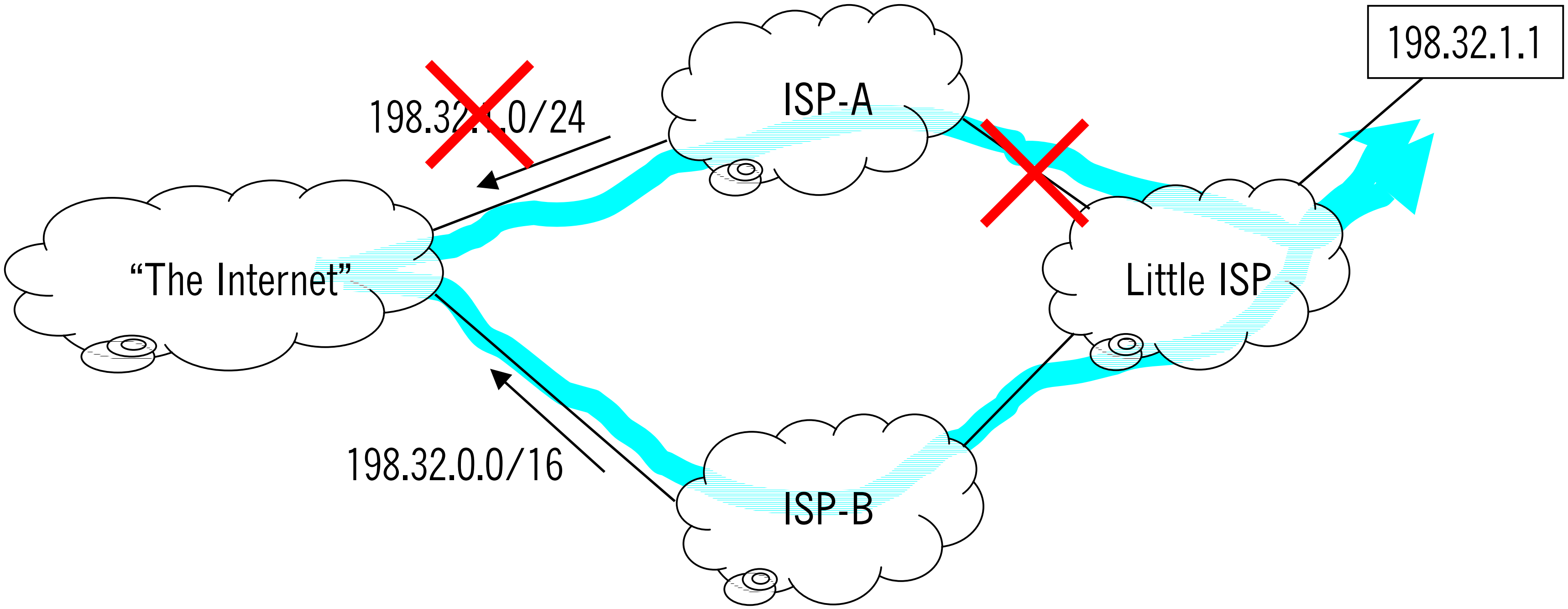
# Forwarding for Routers

- ***Forwarding*** vs. ***Routing***
  - Forwarding is selecting the next-hop machine for each outgoing packet.
    - Forwarding table, FIB.
  - Routing is the process of deciding the path from a source to a destination.
    - Routing table, RIB.
- Select the next-hop router.
  - Find the outgoing interface.
  - Find the MAC address of the next-hop router.
  - In Unix, you specify the IP address of the next-hop router.
- Longest-prefix first.
- Default routing (implied by longest-prefix rule: default has prefix of length 0).

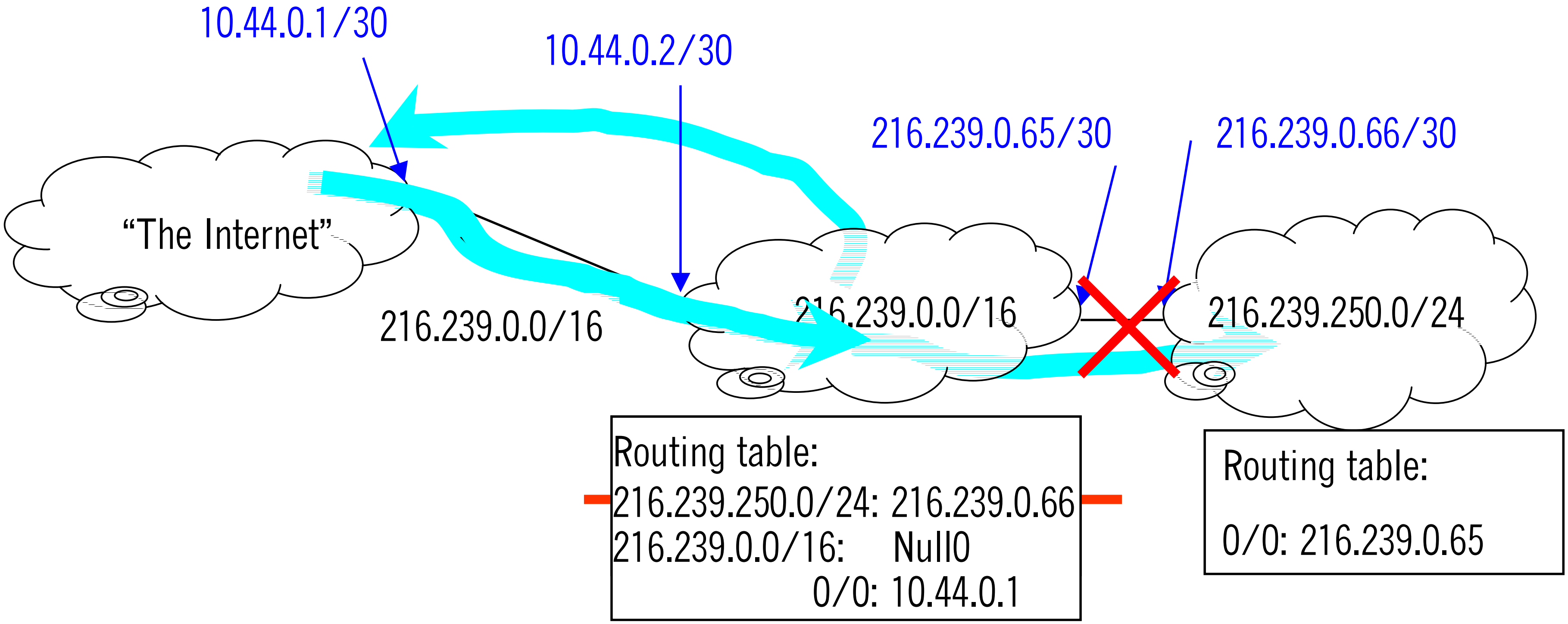
# Forwarding beyond the LAN

- IP routing is destination-address-based only.
- Routing protocol is used to derive the forwarding table.
  - Routers advertise prefixes that they know how to route to.
  - Lots of ways of doing this, hence lots of routing protocols.
- Routers forward to the next-hop router until destination is reached.
- Routers near the edges have "default" routes.
  - Also, static routes.
- Multiple forwarding entries may match an address.
  - Longest-prefix match wins.
- Default-free zone.

# Longest Path First



# Default Routing



10.44.0.1/30

10.44.0.2/30

216.239.0.65/30

216.239.0.66/30

“The Internet”

216.239.0.0/16

216.239.0.0/16

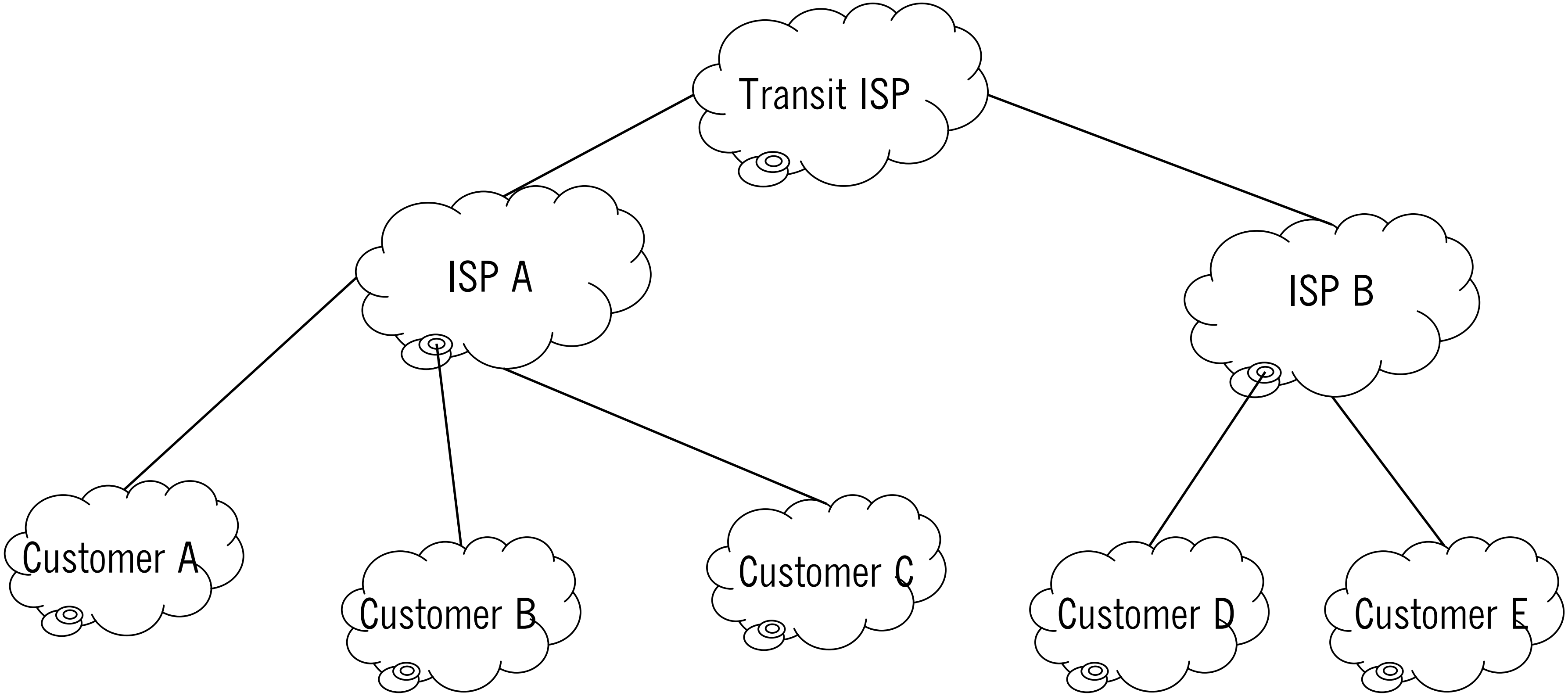
216.239.250.0/24

Routing table:
216.239.250.0/24: 216.239.0.66
216.239.0.0/16: Null0
0/0: 10.44.0.1

Routing table:
0/0: 216.239.0.65

- Add a bit bucket for own aggregate when doing default routing!

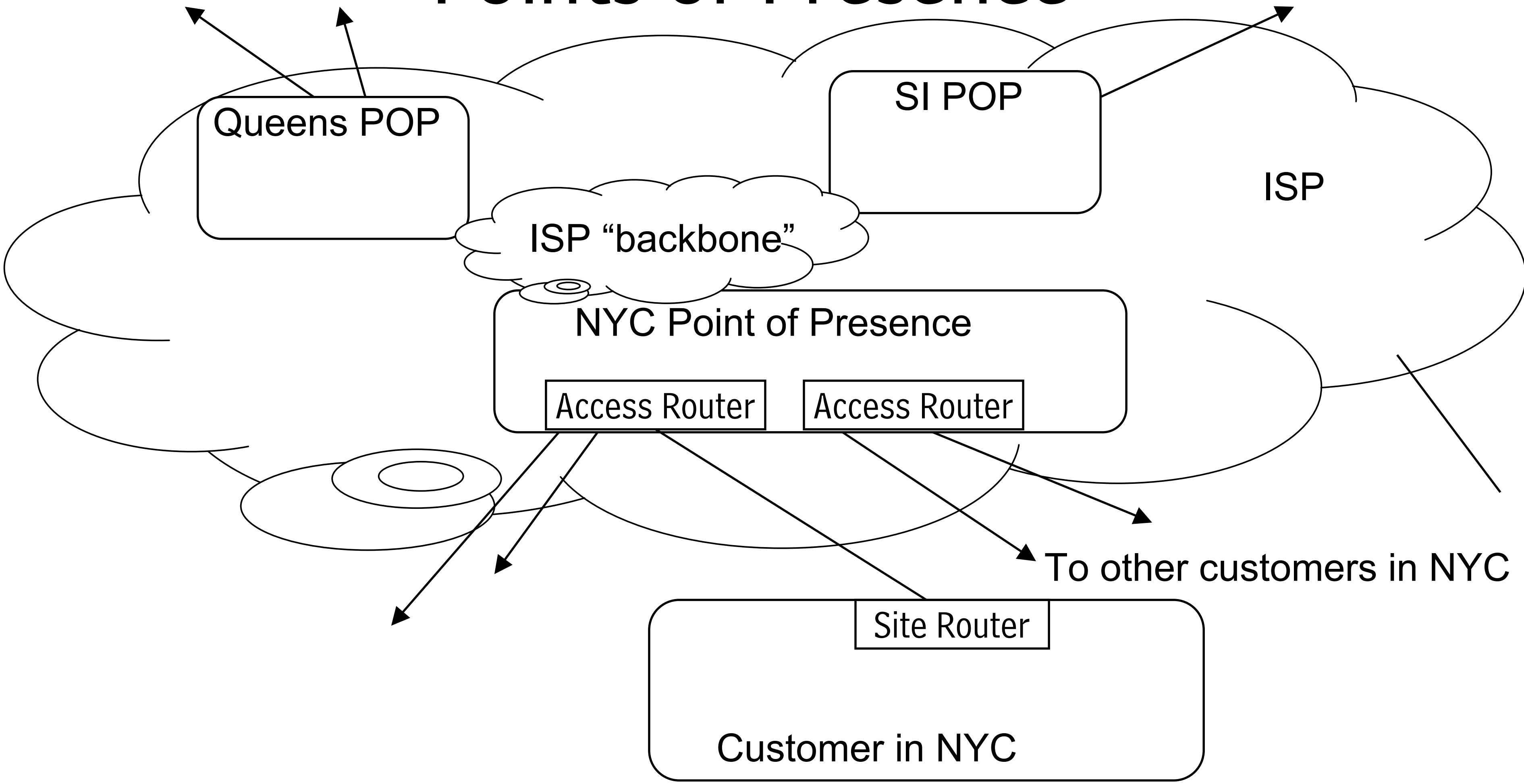
# Transit Networks



# Peering

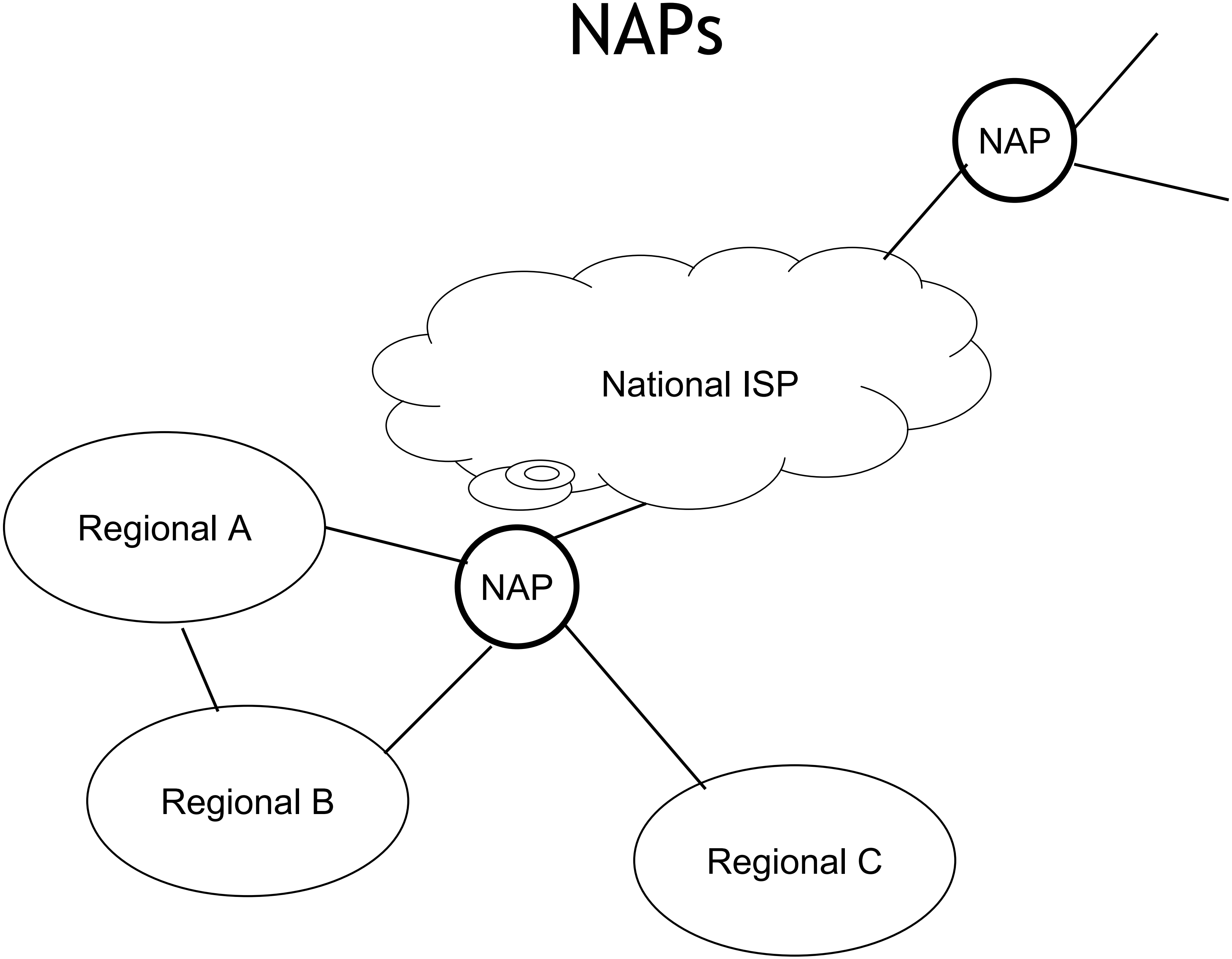
- The Internet is not a tree!
  - Unlike bridging, all links are active.
  - Routing determines which links are used for each pair of hosts.
- Network Providers exchange traffic at peering points.
- Regional Networks.
- Tier-1 networks.
- NAPs.
- Private peering.
  - Peering agreements.
  - Often very confidential.
- Route servers.
- Policy.
- We’ll keep coming back to this throughout the semester.

# Points of Presence



- Customers connect at POPs.
- POPs are connected by the ISP’s backbone.
- ISPs can be local, regional, national, global, etc.

# NAPs



# NAPs and Peering

- Small/Regional ISPs connect at NAPs.
- Large/National ISPs provide connectivity at NAPs.
- Mainly, they have private peering agreements.
- National ISPs provide both customer and transit traffic.

# Address Allocation

- Customers (sites, companies, organizations, universities, etc.) get a CIDR Block.
- Their provider is responsible for routing it.
  - Advertising the CIDR Block.
  - Getting packets to it.
- In the "before" time:
  - Customers got an allocation (class A/B/C) from the NIC, then the IANA.
  - Did not scale (a couple of people were doing the allocations).
  - Addresses were assigned without considerations for aggregation.

# Address Allocation, Cont’d

- Since CIDR.
  - Regional Registries (ARIN, RIPE, APNIC).
  - Registries get allocated /8s or shorter.
  - Registries allocate space to ISPs on a need-to-have basis.
  - ISPs allocate space to customers (who can also be smaller ISPs).
  - Most of the address space is non-portable (“belongs” to the ISP).
  - Much better for aggregation.
- Still a lot of old portable address space around.
- Customers who can justify portable space can still get it.
  - And guard it jealously.

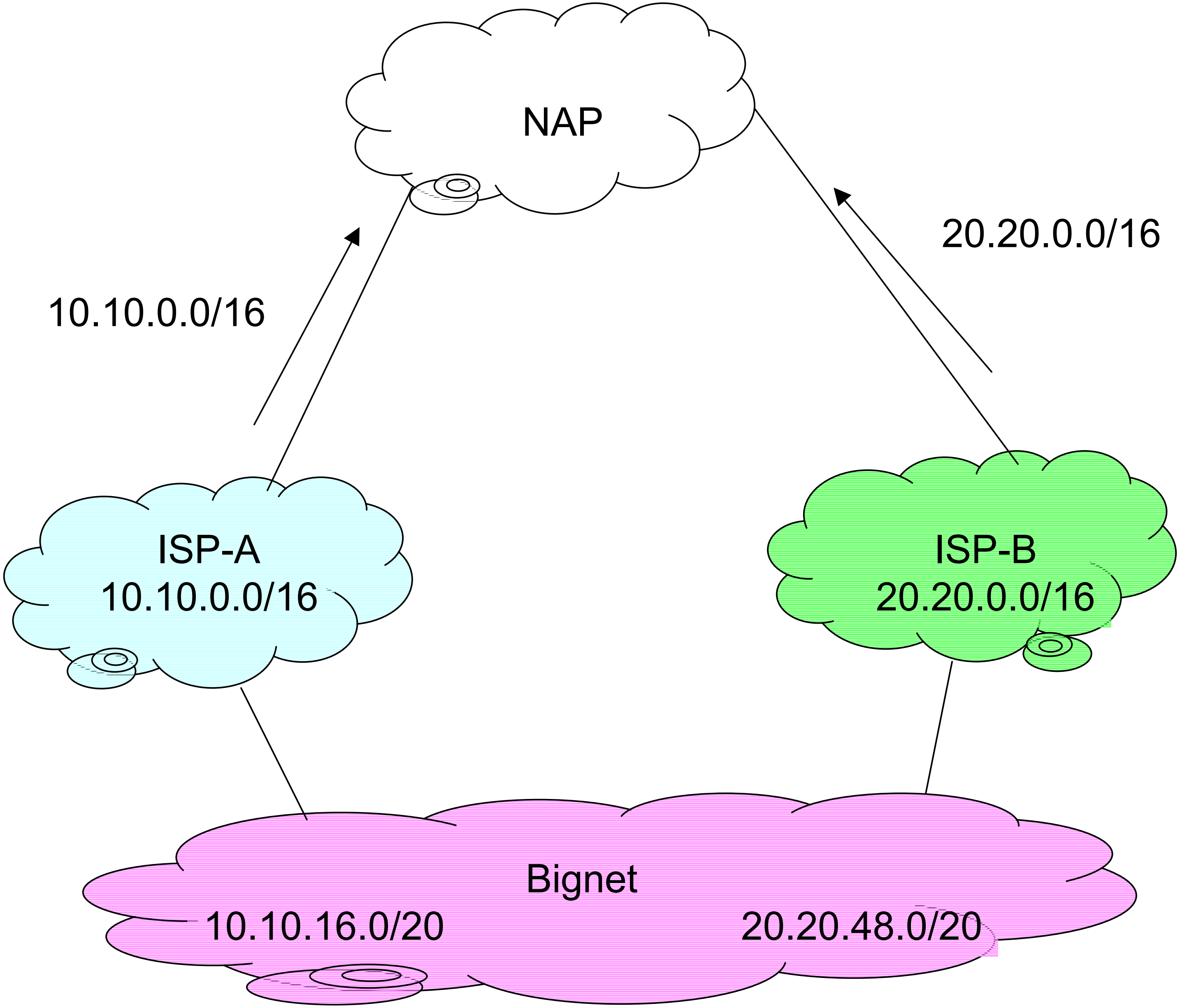
# Single-homed networks.

- Customers with only one provider are called “single-homed”.
- Their ISP is their default route.
  - No need to run routing protocols.
- Can have portable or non-portable address space.
- ISP advertises their address space.
- What happens when they change providers?
  - Portable space: no problem.
  - Non-portable space:
    - Renumber (big pain).
    - Steal the previous providers address space.
      - It happens all the time.

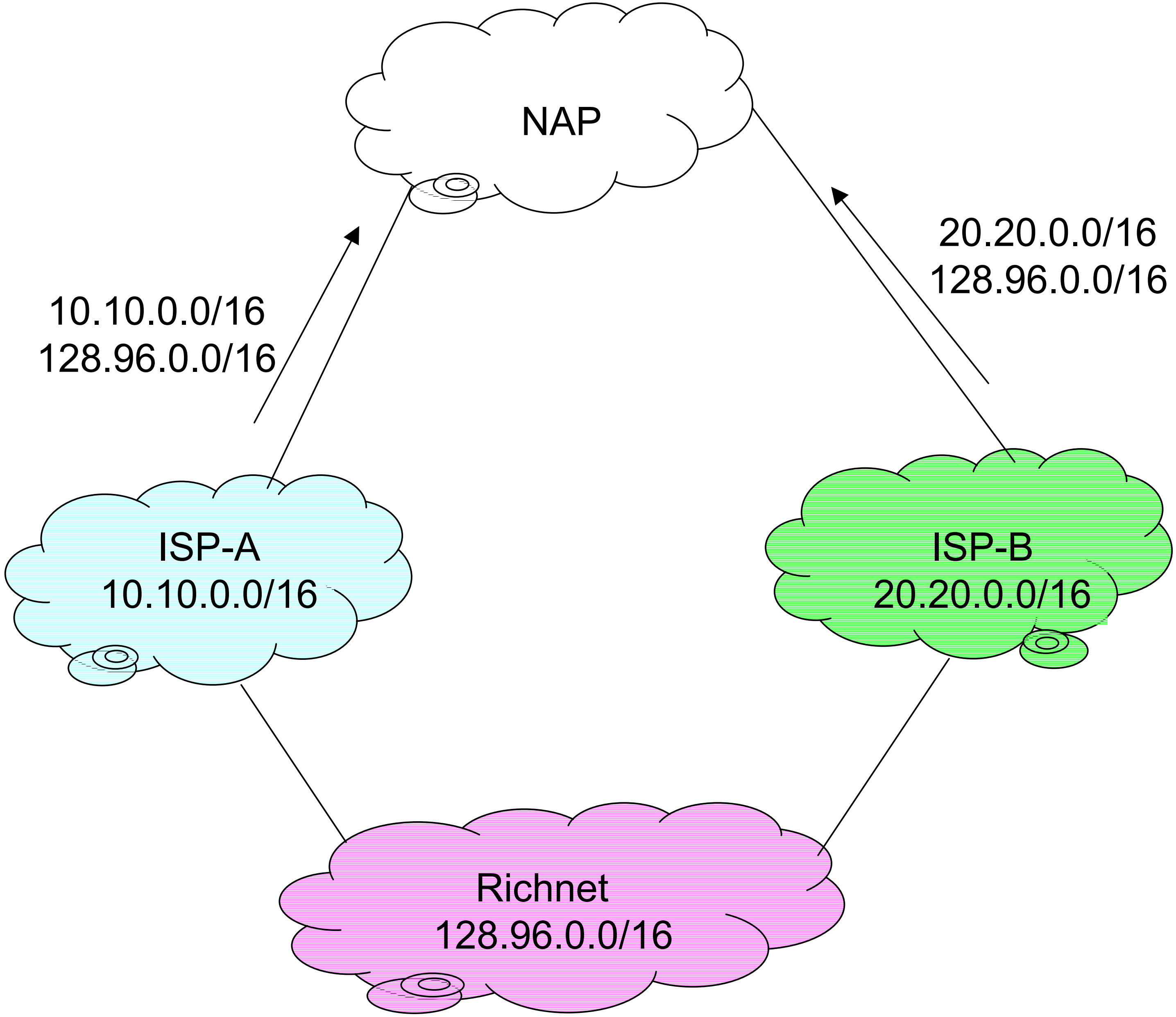
# Multihoming

- A node can have interfaces connected to multiple networks.
  - If it forwards between interfaces, it’s called a router.
  - If it does not, it’s called a multi-homed host.
- A network can also be multihomed.
  - Have service from more than one ISP.
- Multihomed networks create interesting routing problems.
  - Address space usage.
  - Issues with aggregation.
  - Traffic engineering.
  - Policy.
  - Reachability.

# Multihoming II



# Multihoming III


NAP
10.10.0.0/16
128.96.0.0/16
20.20.0.0/16
128.96.0.0/16
ISP-A
10.10.0.0/16
ISP-B
20.20.0.0/16
Richnet
128.96.0.0/16