

E6998-02: Internet Routing

Lecture 3

Routing and Addresses

John Ioannidis

AT&T Labs – Research

`ji+ir@cs.columbia.edu`

Announcements

Lectures 1-4 are available.

Homework 1 due 9/17 at 3am.

Email your homework to `ji+hw1@cs.columbia.edu`.

Only plain-ASCII or PDF files will be accepted!

Class BBoard: at `coms6998-002-023@columbia.edu` and
<https://www1.columbia.edu/sec/bboard/023/coms6998-002/>

AIM discussion group.

Noel Chiappa guest lecturer on 9/19. BE HERE!

Everybody, please send email (`ji+ir@cs.columbia.edu`)
telling me if you're taking or auditing the class, and if
you're a CVN student.

Why Layer-3 Routing?

- Why can't we have the entire Internet work with learning bridges and ST?
- Heterogeneity in network technologies.
- Require bridges to know about all end stations!
 - Bad for scaling.
- Abstraction layers.
- Some limitations of bridging (loops are bad).
- Structure (nets of nets of nets...).
- Scaling needs hierarchy.
 - Only way we know how to do it!

Internet Address Format

- IP Addresses are 32 bits long.
 - “Dotted-decimal notation”: 0x803b1014 is 128.59.16.20
 - Shortcuts (rare): 127.0.0.1 same as 127.1
- The old days (1822 format): net.port.port.PSN
 - ARPANET: net=10, MILNET: net=26
 - COLUMBIA-20.ARPA was 10.0.0.89
 - ipswitch.bellcore.com was 10.7.0.89
- Network-host split.
 - “Classful” addressing.

Two-level Address Hierarchy

- Why split the address into a network part and a host part?
 - Hosts on the same LAN/organization share the same network part.
 - Forwarding decisions can thus be made by looking only at the network part.
 - Therefore, only the network part (or “network number”) needs to be exchanged between routers.
- Address aggregation.
 - Representing groups of addresses by only one (their common prefix).
- Hierarchy allows scaling!

Address Classes (old stuff)

- First few bits determine “address class” and net-host boundaries:
 - 0... (0.0.0.0-127.255.255.255): Class A, 8-bit nets.
 - 10... (128.0.0.0-191.255.255.255): Class B, 16-bit nets.
 - 110... (192.0.0.0-223.255.255.255): Class C, 24-bit nets.
 - 1110... (224.0.0.0-239.255.255.255): Class D, multicast.
 - 1111... (240.0.0.0-255.255.255.255): Class E, reserved.
- Classes A/B/C:
 - Wasteful allocation of address space (not enough Class A/Bs).
 - Routing table explosion (too many “Class Cs”).

Old Address Allocation Plan

- An organization gets a Class A, B, or C depending on its expected size.
 - ARPA=10, ATT=12, MIT=18, UCB=32, Stanford=36...
 - Columbia=128.59, Bellcore=128.96
 - GIP-Altaïr=192.27.52
- Turns out Class Cs were too small, Class B and A way too large.
- ~1992, no one was getting any more Class As, and we were afraid we were soon going to run out of Class Bs.
- People hated Class Cs because they had to get several of them, and that increased the size of the routing tables.

Subnetting

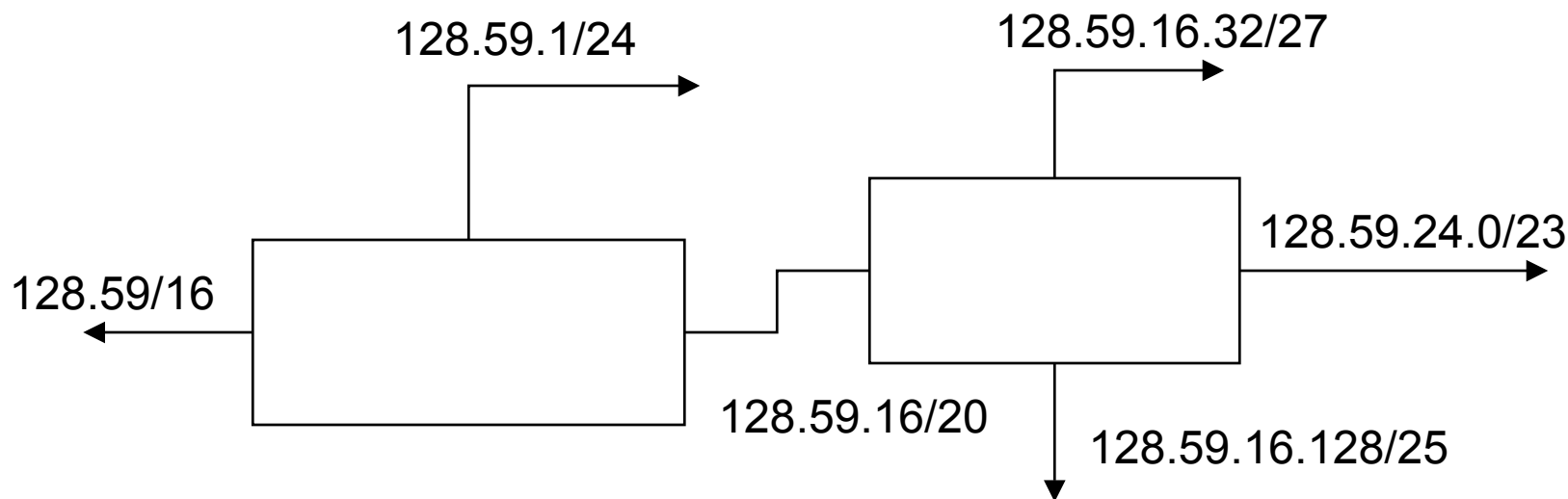
- Breaking up the “host” part of the address into subnet-host.
 - 2 levels of hierarchy good, more are better!
- Bridging too expensive even within a site.
- Originally, along 8-bit boundaries.
- Netmask: shows which part of address is “net”:
 - 128.59.16.0/255.255.240.0
 - 10000000 00111011 00010000 00000000
 - 11111111 11111111 11110000 00000000
 - 128.59.16.0/20
- Allows for address aggregation within a site:
 - Access router handles 128.59.0.0/16.
 - Various routers around campus handle subnets.
 - Originally, all subnetted interfaces had to have the same netmask.

Subnetting, Cont'd

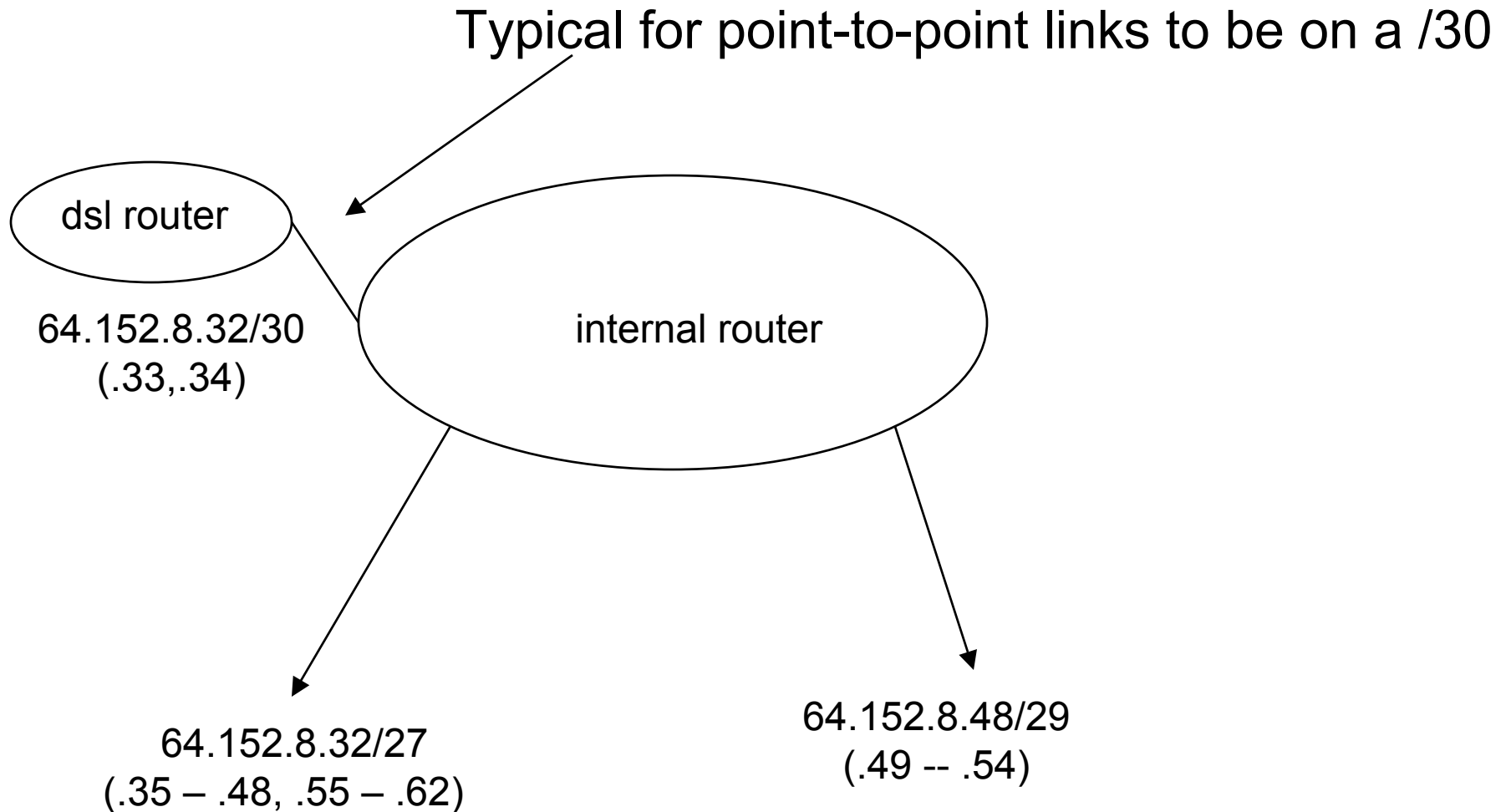
- Host-within-subnet part all zeros → anycast.
 - 135.207.4.64/26
- Host-within-subnet part all ones → directed broadcast.
 - 192.4.13.127/25
- Whether an address is anycast/broadcast or not depends on the receiving router.
 - Router for 135.207.0.0/16 just routes 135.207.4.64.
 - As does router for 135.207.0.0/20.
 - Router for 135.207.4.64/26 considers it an anycast.

VLSM

- Obvious thing to do.
- Different subnets have different size requirements.
- No need to subnet at the longest prefix.
- Forward according to “longest-match”.



VLSM can have overlapping allocations



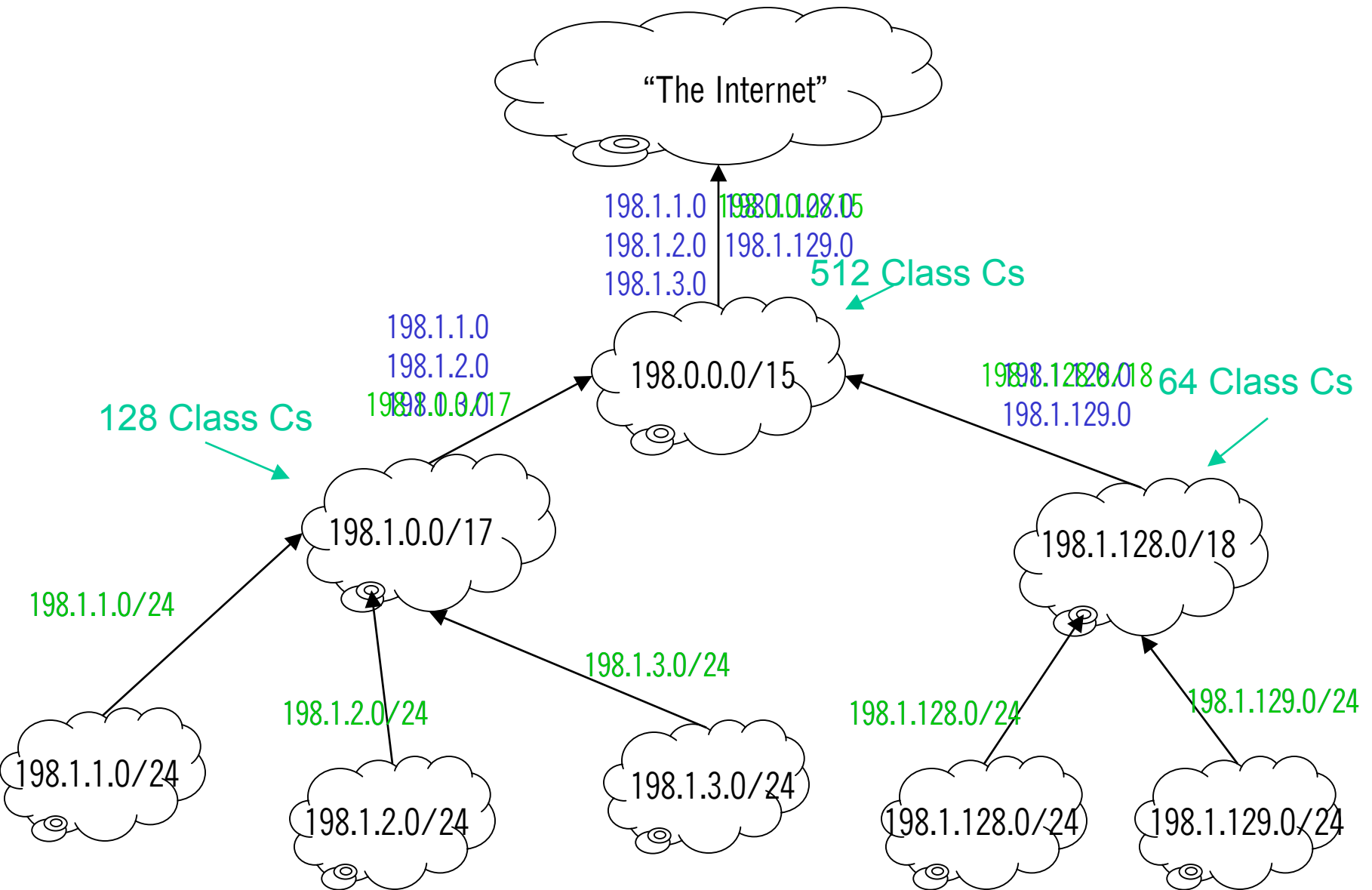
Classless Interdomain Routing (CIDR)

- “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”.

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-05.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 $((src \ \& \ netmask) == (dst \ \& \ netmask)) \{ \dots$
 - Find MAC address of destination (if not on p2p link).
 - Send packet.
- How to send an IP packet to a host on different subnet?
 - $\dots \} \text{ else } \{ \dots$
 - 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.

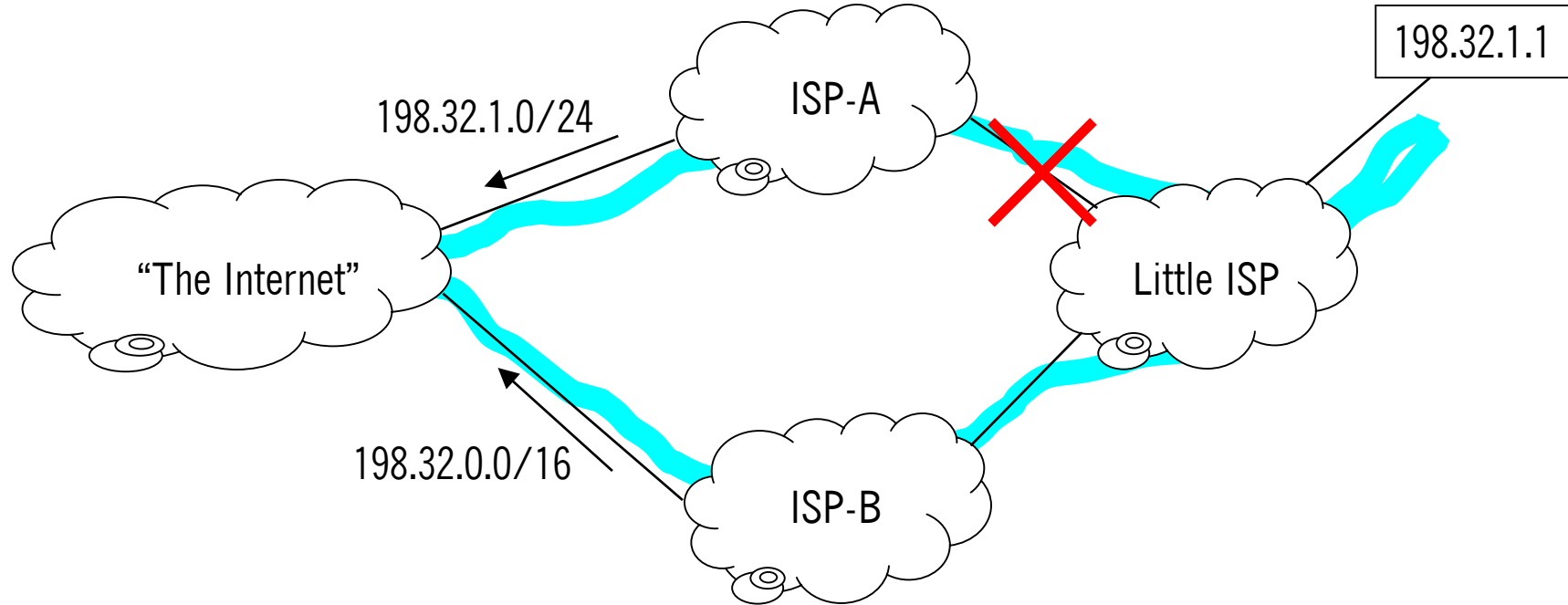
Forwarding for Routers

- ***Forwarding vs. Routing***
 - Forwarding is selecting the next-hop machine for each outgoing packet.
 - Forwarding table.
 - Routing is the process of deciding the path from a source to a destination.
 - Routing table.
- 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).

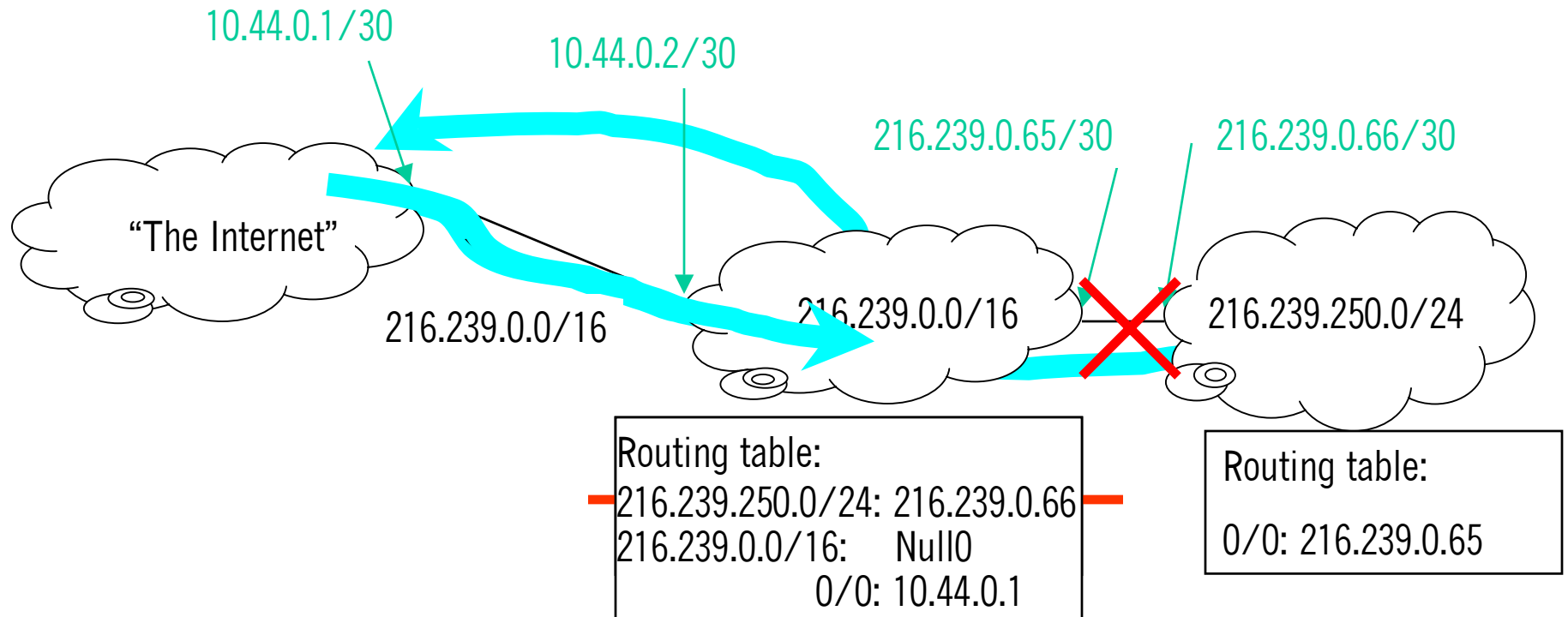
Forwarding beyond the LAN

- IP routing is destination-addressed 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

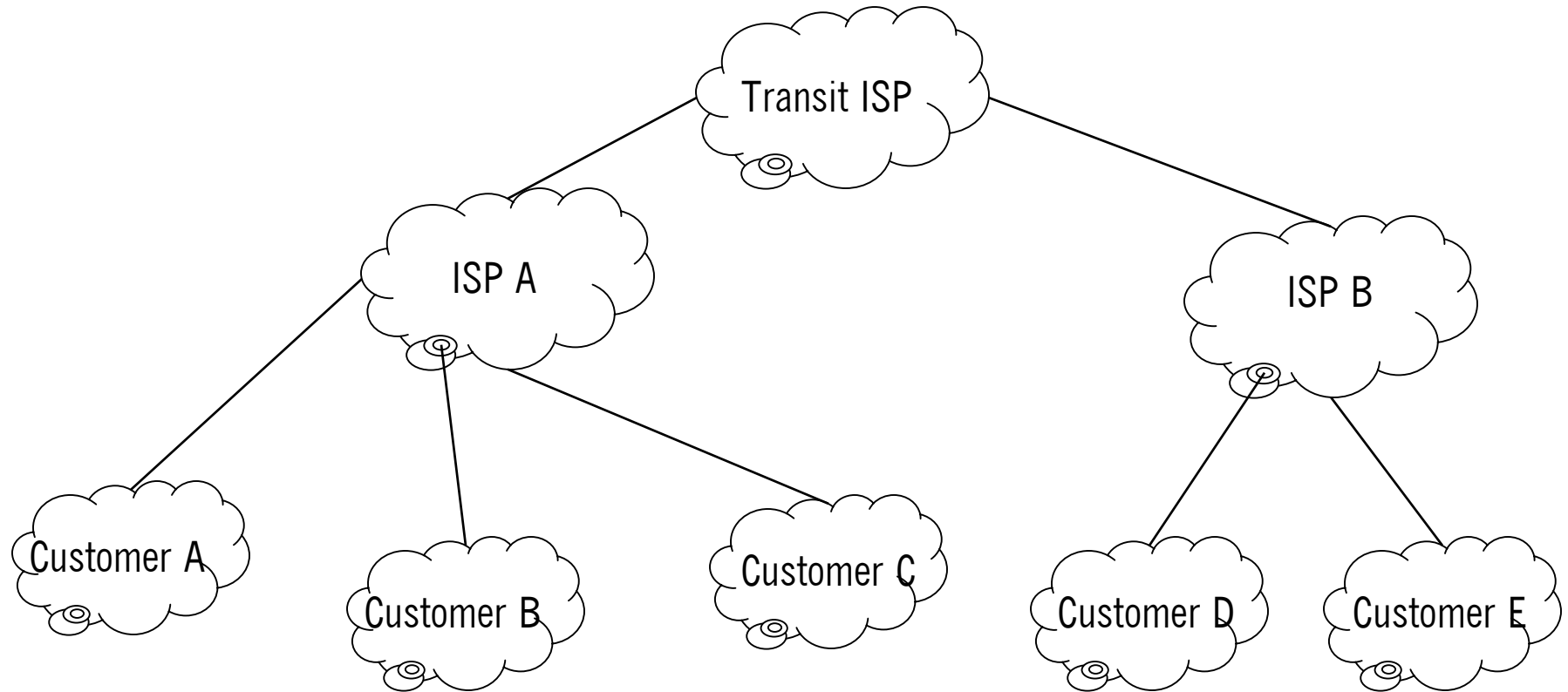


Default Routing



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

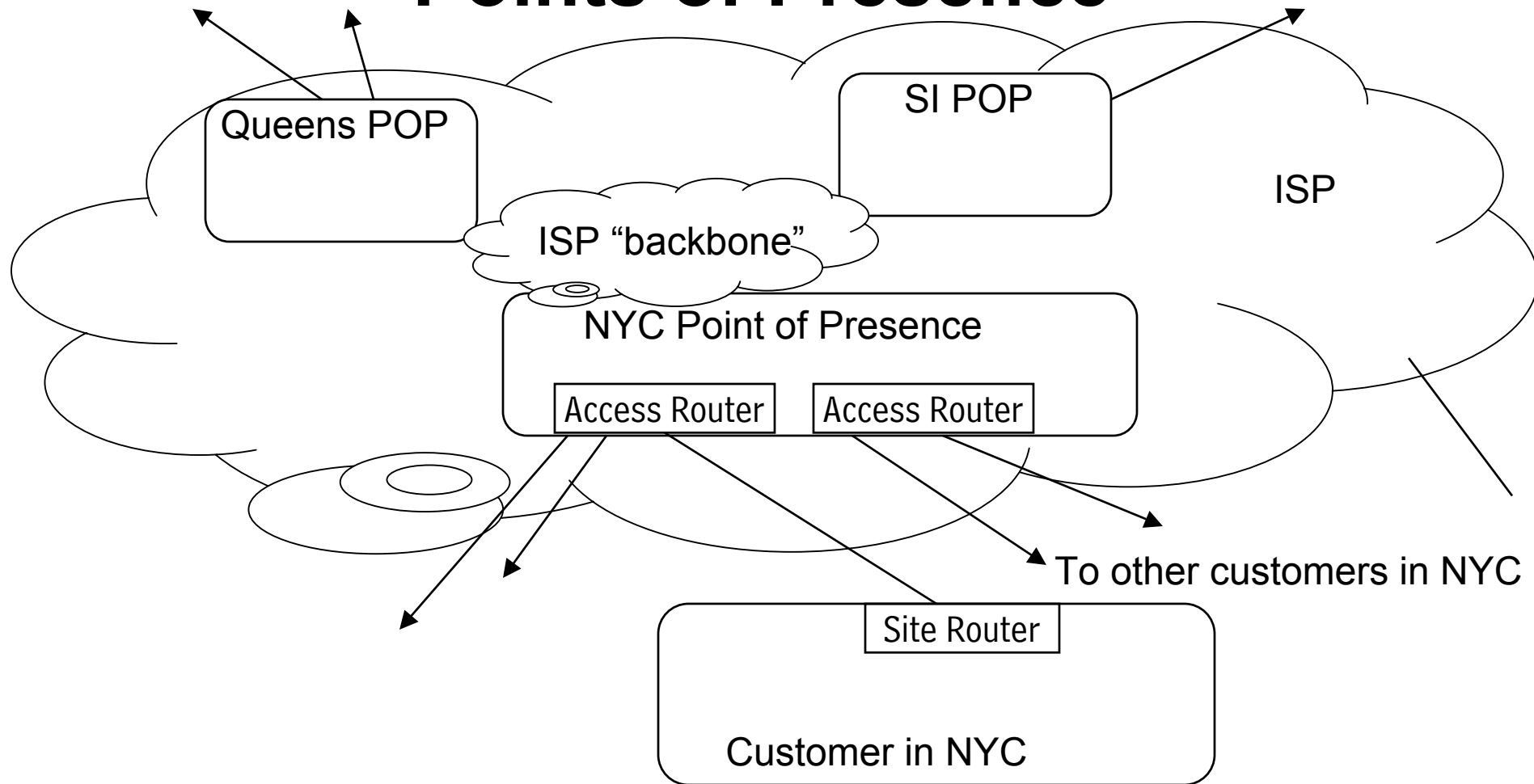
Transit Networks



Peering

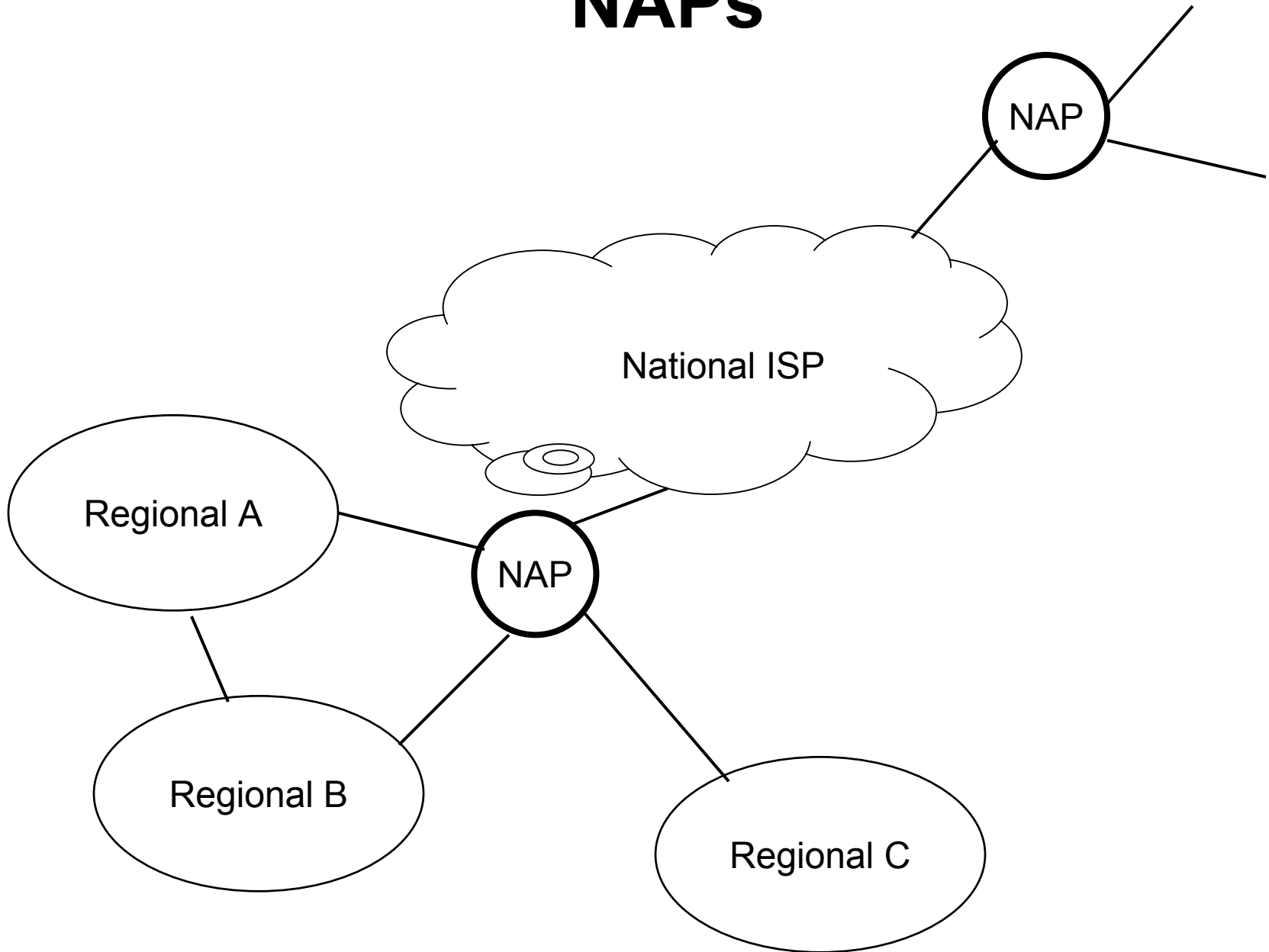
- The Internet is not a tree!
 - Unlike bridging, all links are active.
 - Routing determines which links are used for each pair of hosts.
- Network Providers exchange traffic at peering points.
- Regional Networks.
- Tier-1 networks.
- NAPs.
- Private peering.
 - Peering agreements.
 - Often very confidential.
- 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.

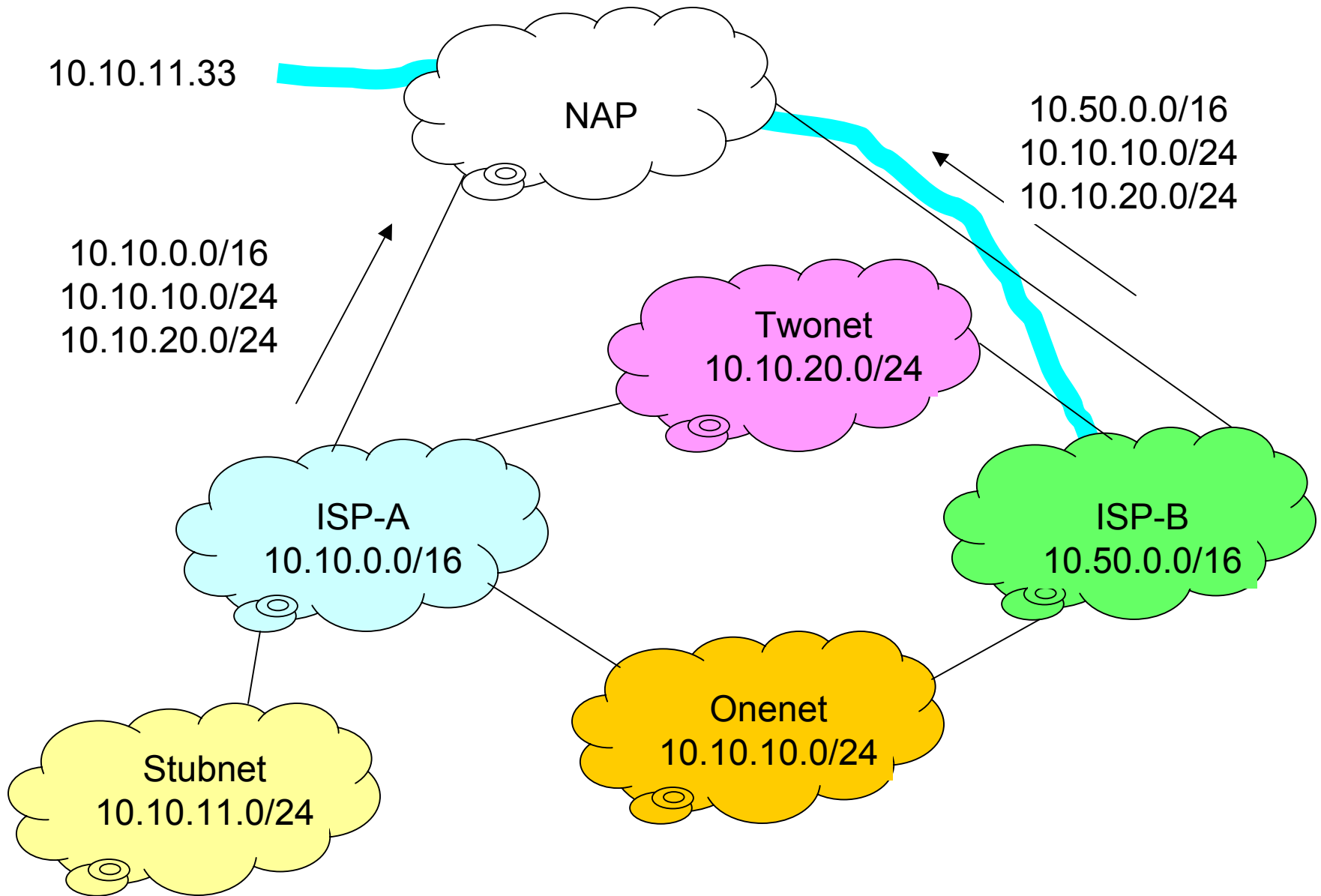
Single-homed networks.

- 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.

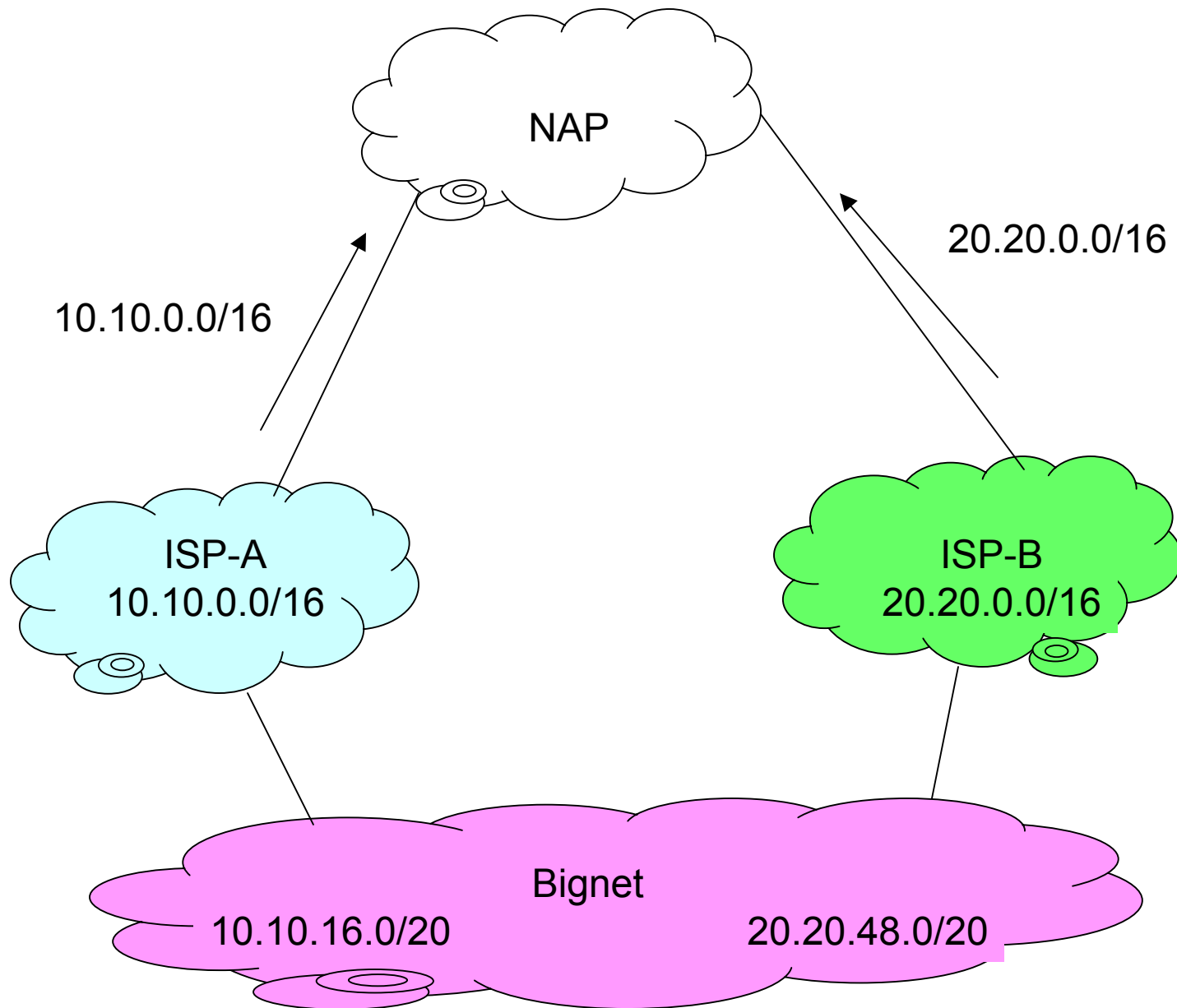
Multihoming

- 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.

Advertising wrong aggregate



Multihoming II



Multihoming II

