CS 356: Computer Network Architectures

Lecture 13: Border Gateway Protocol and switching hardware

[PD] chapter 4.1.2

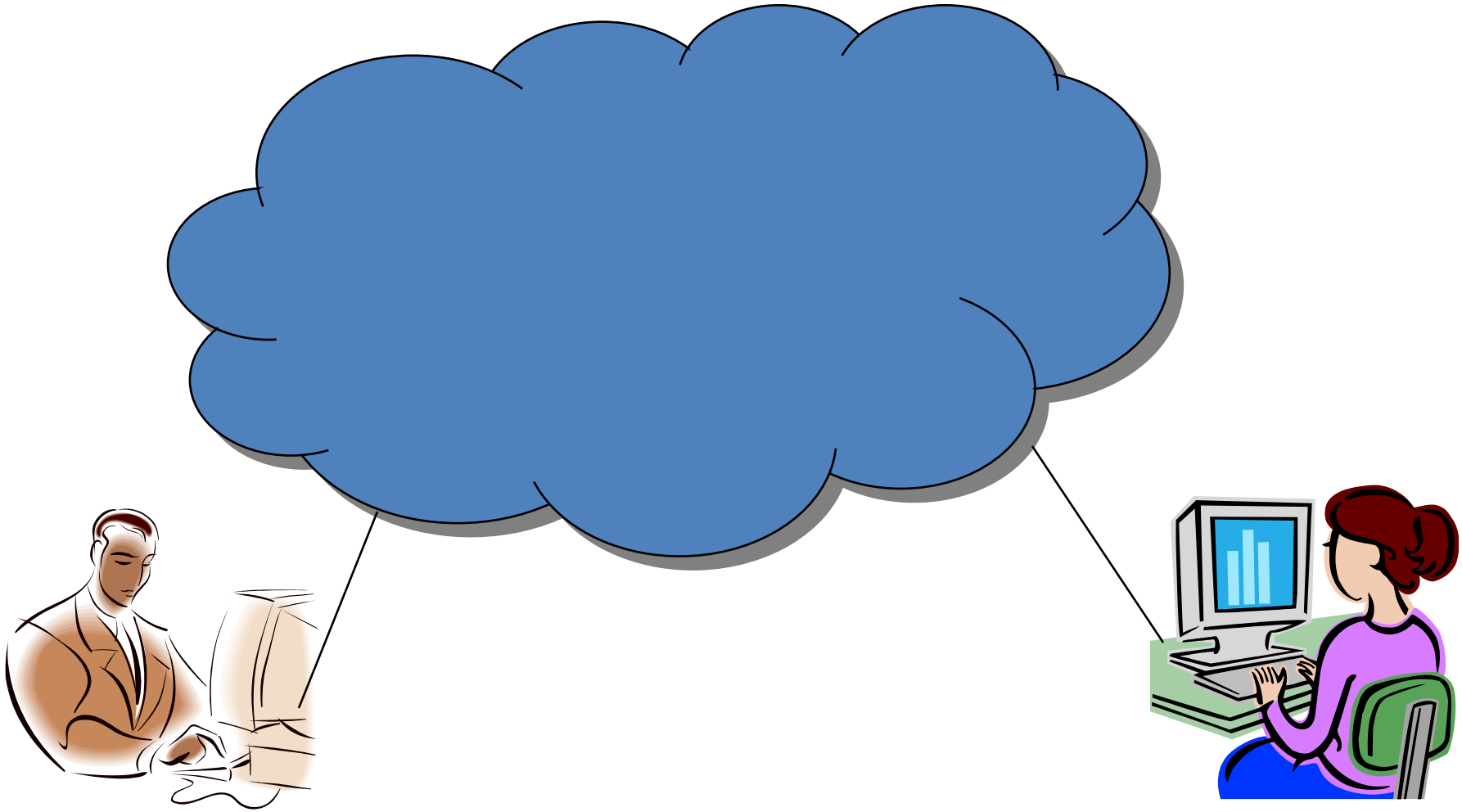
Xiaowei Yang

xwy@cs.duke.edu

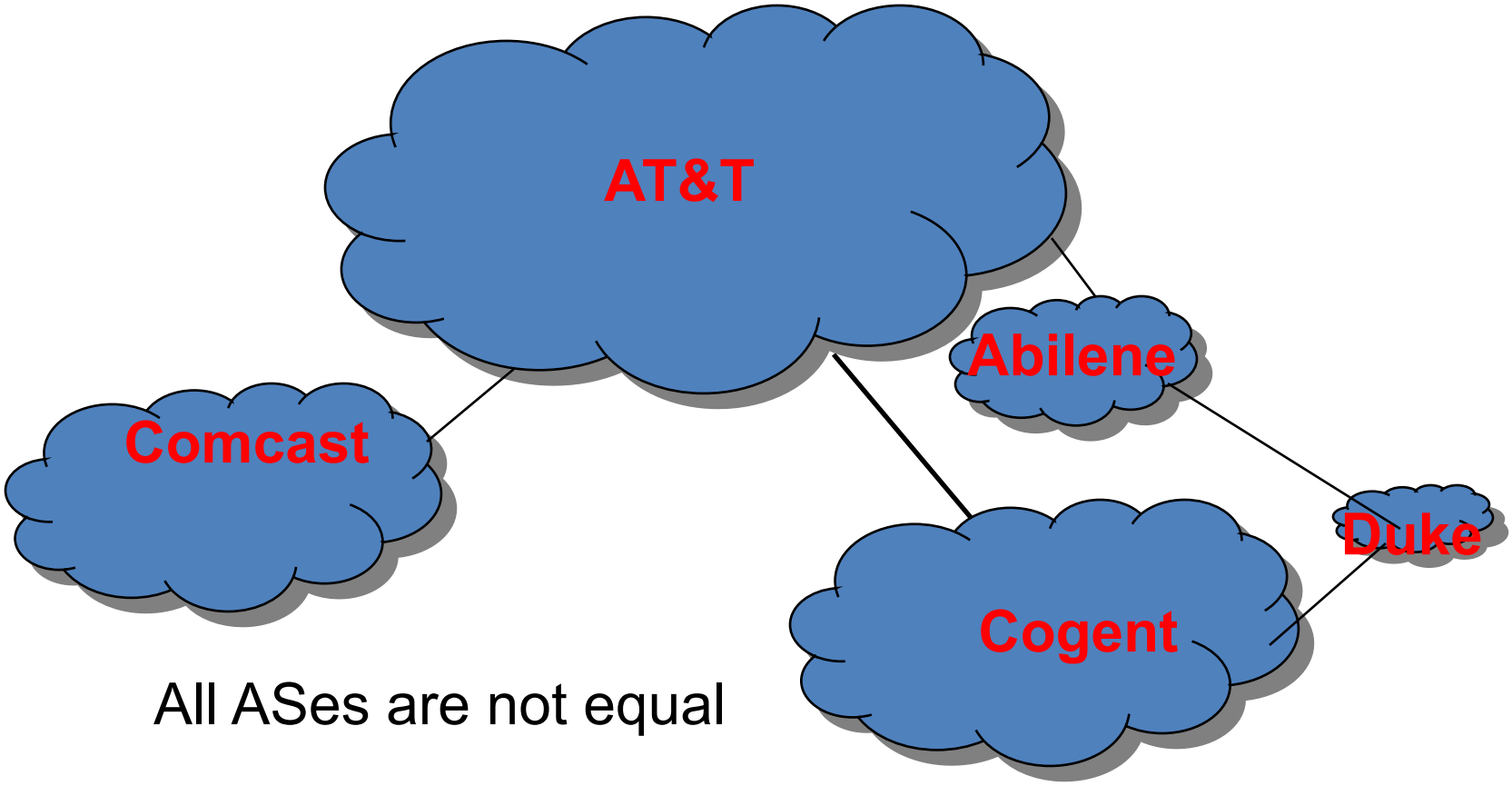
Today

- Border Gateway Protocol (BGP)
- Lab 2

The Internet

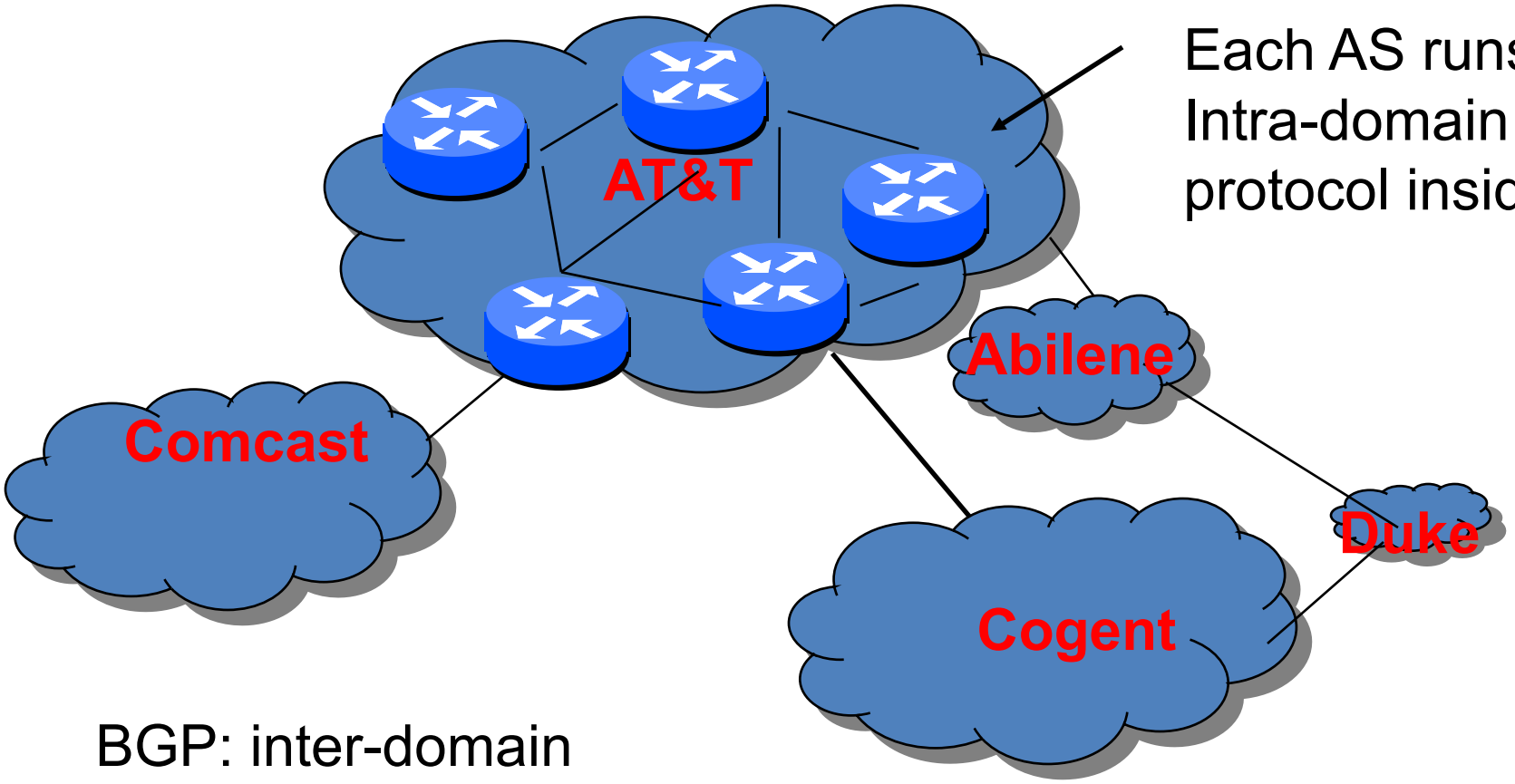


The Internet: Zooming In 2x



All ASes are not equal

Intra-domain vs. inter-domain routing



BGP: inter-domain

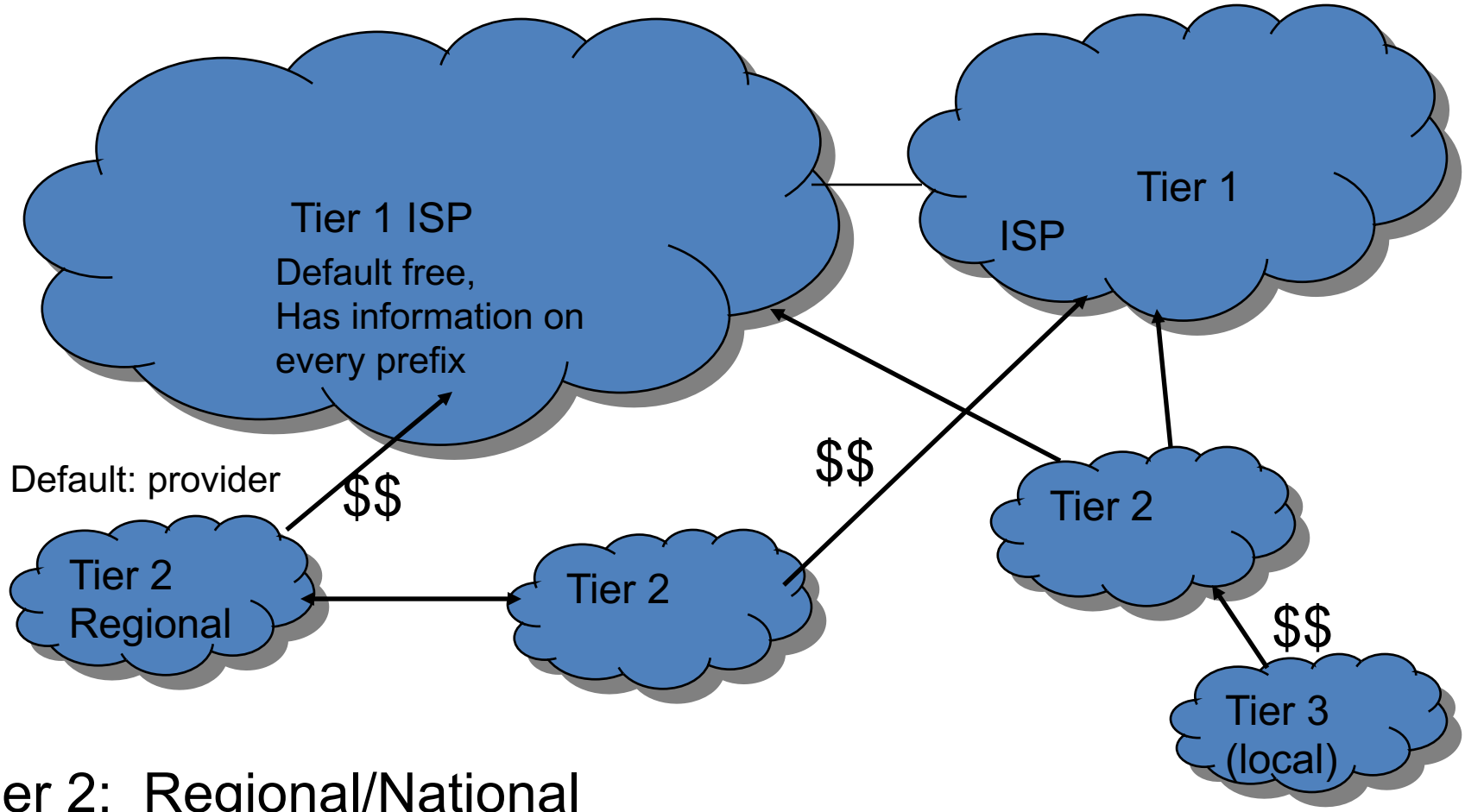
BGP is a policy routing protocol

- BGP helps an AS choose a next-hop AS
- Decision made based on AS policies
- Polices are largely determined by AS relationships

AS relationships

- Very complex economic landscape
- Simplifying a bit:
  - Transit: "I pay you to carry my packets to everywhere" (provider-customer)
  - Peering: "For free, I carry your packets to my customers only." (peer-peer)
- Technical definition of tier-1 ISP: In the "default-free" zone. No transit.
  - Note that other "tiers" are marketing, but convenient. "Tier 3" may connect to tier-1.
- ASes keep them as secret

Zooming in 4x



Tier 2: Regional/National

Tier 3: Local

Who pays whom?

- Transit: Customer pays the provider
  - Who is who? Usually, the one who can "live without" the other. AT&T does not need Duke, but Duke needs *some* ISP.

- What if both need each other? Free Peering.
  - Instead of sending packets over $$ transit, set up a direct connection and exchange traffic for free!
  - http://vijaygill.wordpress.com/2009/09/08/peering-policy-analysis/

- Tier 1s must all peer with each other by definition
  - Tier 1s form a full mesh Internet core
- Peering *can* give:
  - Better performance
  - Lower cost
- But negotiating can be very tricky!

Business and peering

- Cooperative competition (coopetition)
- Much more desirable to have your peer's customers
  - Much nicer to get paid for transit
- Peering "tiffs" are relatively common in early days

31 Jul 2005: Level 3 Notifies Cogent of intent to disconnect.
16 Aug 2005: Cogent begins massive sales effort and mentions a 15 Sept. expected depeering date.
31 Aug 2005: Level 3 Notifies Cogent again of intent to disconnect (according to Level 3)
5 Oct 2005 9:50 UTC: Level 3 disconnects Cogent. Mass hysteria ensues up to, and including policymakers in Washington, D.C.
7 Oct 2005: Level 3 reconnects Cogent

During the "outage", Level 3 and Cogent's singly homed customers could not reach each other. (~ 4% of the Internet's prefixes were isolated from each other)

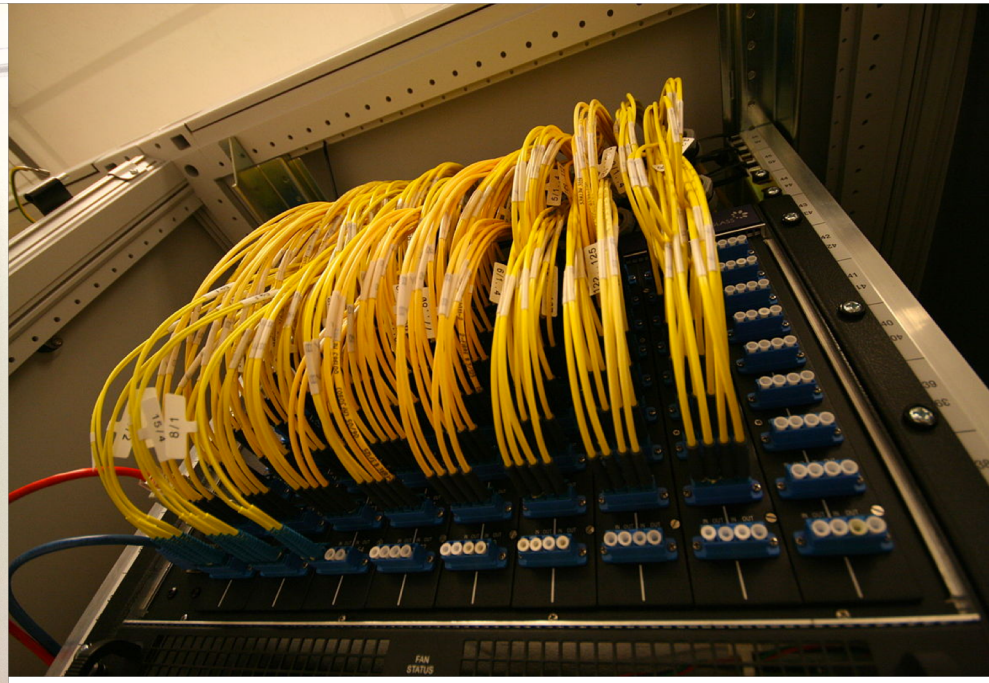
Internet exchange point

- https://www.internetexchangemap.com/
- Places where ISPs interconnect and exchange traffic
- https://www.internetexchangemap.com/

London Internet Exchange (LINX)

- Telehouse Docklands, July 2005. Photo by John Arundel.

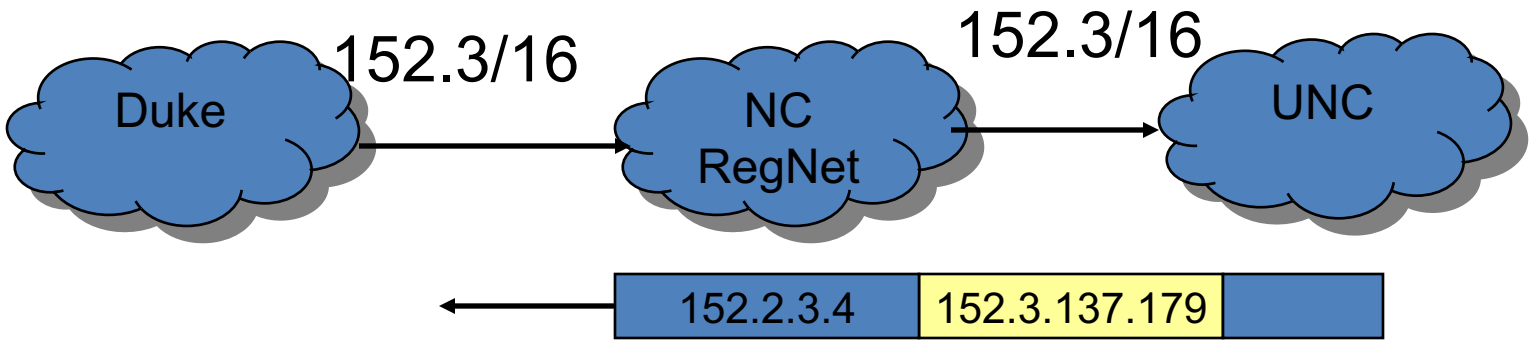
Inside an Internet Exchange Point



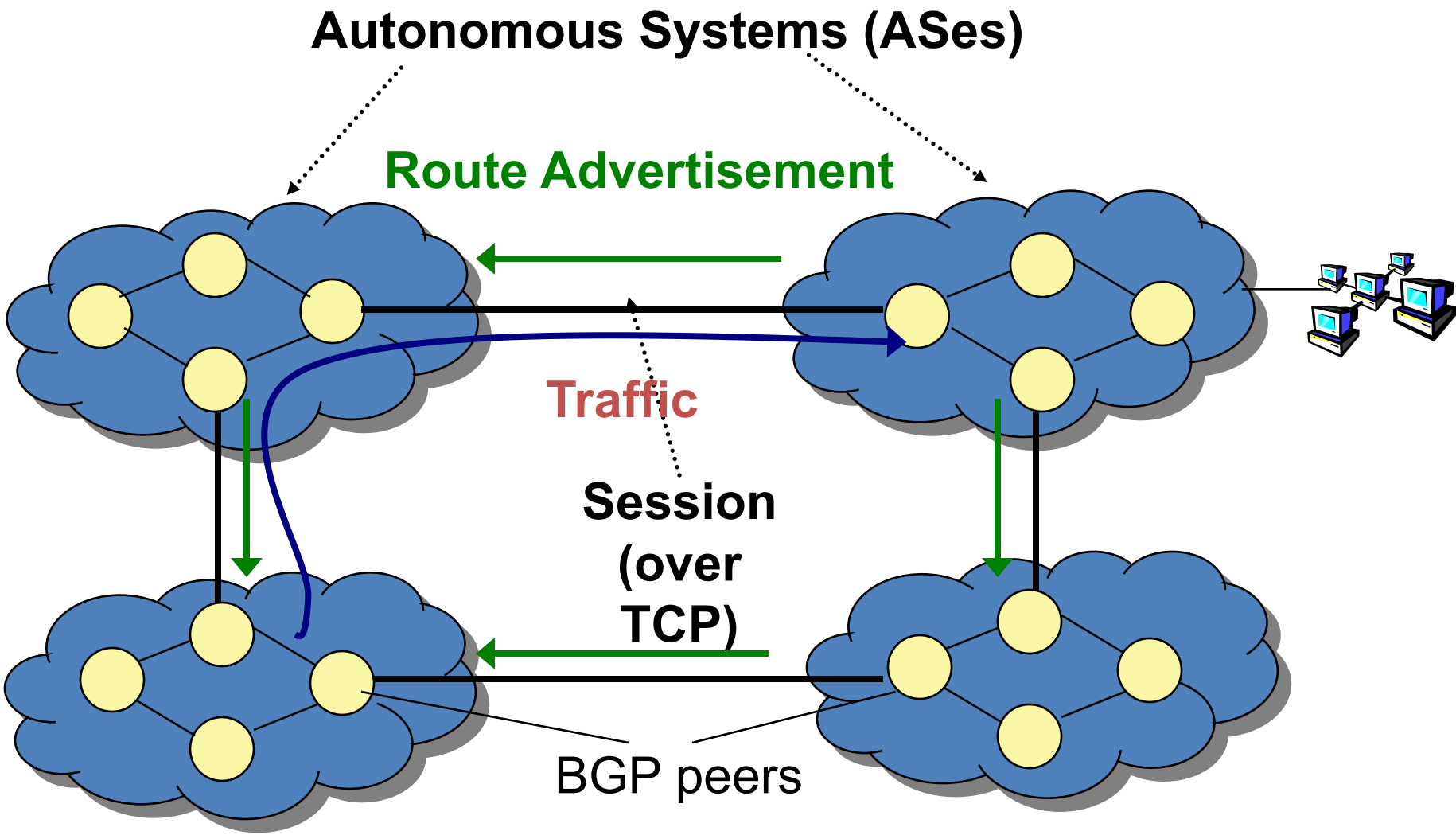
- By Fabienne Serriere - http://fbz.smugmug.com/gallery/4650061_iuZVn/5/282300855_hV8xq#282337724_tZqT2, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=4092825
- By Stefan Funke from Frankfurt, Germany - Switch RackUploaded by MainFrame, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=26260389

Terms

- Route: a network prefix plus path attributes
- Customer/provider/peer routes: route advertisements heard from customers/providers/peers
- Transit service: If A advertises a route to B, it implies that A will forward packets coming from B to any destination in the advertised prefix



BGP


Autonomous Systems (ASes)
Route Advertisement
Traffic
Session
(over
TCP)
BGP peers

Enforcing relationships

- Two mechanisms
  - Route export filters
    - Control what routes you send to neighbors
  - Route import ranking
    - Controls which route you prefer of those you hear.
    - “LOCALPREF” – Local Preference. More later.

Export Policies

- Provider → Customer
  - All routes so as to provide transit service
- Customer → Provider
  - Only customer routes
  - Why?
  - Only transit for those that pay
- Peer → Peer
  - Only customer routes

Import policies

- Same routes heard from providers, customers, and peers, whom to choose?
  - customer > peer > provider
  - Why?
  - Choose the most economic routes!
    - Customer route: charge $$ ☺
    - Peer route: free
    - Provider route: pay $$ ☹

Now the nitty-gritty details!

BGP

- BGP = Border Gateway Protocol
  - Currently in version 4, specified in RFC 1771. (~ 60 pages)
- Inter-domain routing protocol for routing between autonomous systems
- Uses TCP to establish a BGP session and to send routing messages over the BGP session
- BGP is a path vector protocol
  - Similar to distance vector routing, but routing messages in BGP contain complete paths
- Network administrators can specify routing policies

BGP policy routing

- BGP's goal is to find any path (not an optimal one)
  - Since the internals of the AS are never revealed, finding an optimal path is not feasible
- Network administrator sets BGP's policies to determine the best path to reach a destination network

BGP messages

- OPEN
- UPDATE
  - Announcements
    - Dest Next-hop AS Path … other attributes …
    - 128.2.0.0/16 196.7.106.245 2905 701 1239 5050 9
  - Withdrawals
- KEEPALIVE
  - Keepalive timer / hold timer

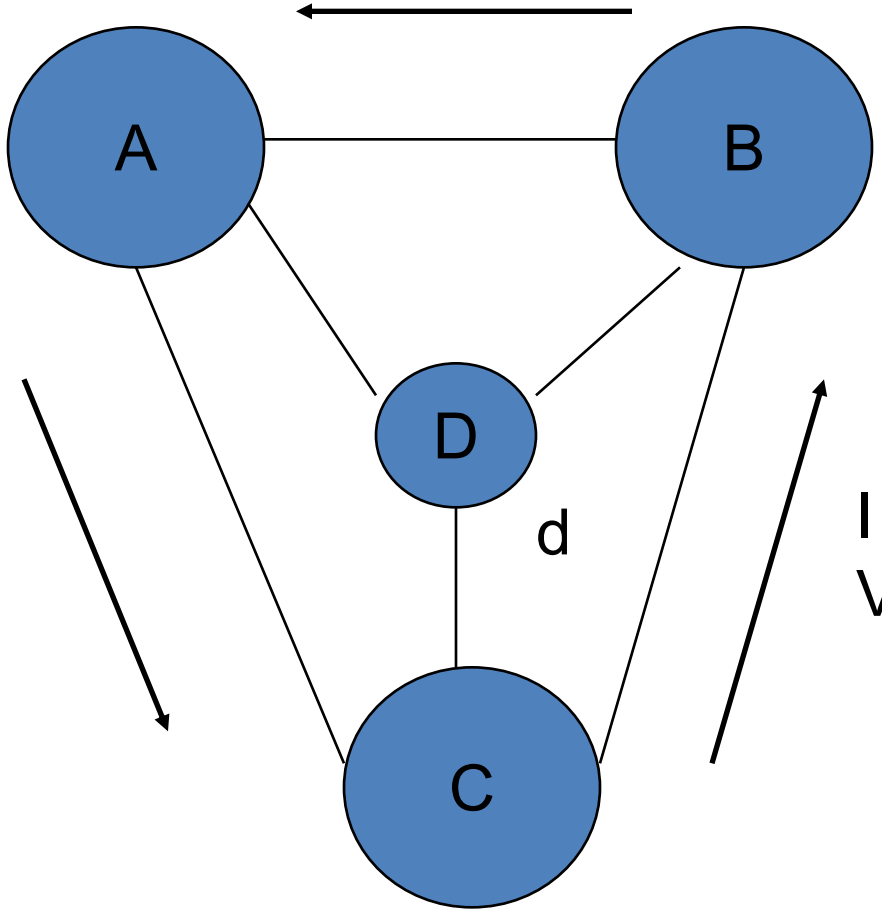
- Key thing: The Next Hop attribute

Path Vector

- ASPATH Attribute
  - Records what ASes a route goes through
  - Loop avoidance: Immediately discard
  - Shortest path heuristics

- Like distance vector, but fixes the count-to-infinity problem

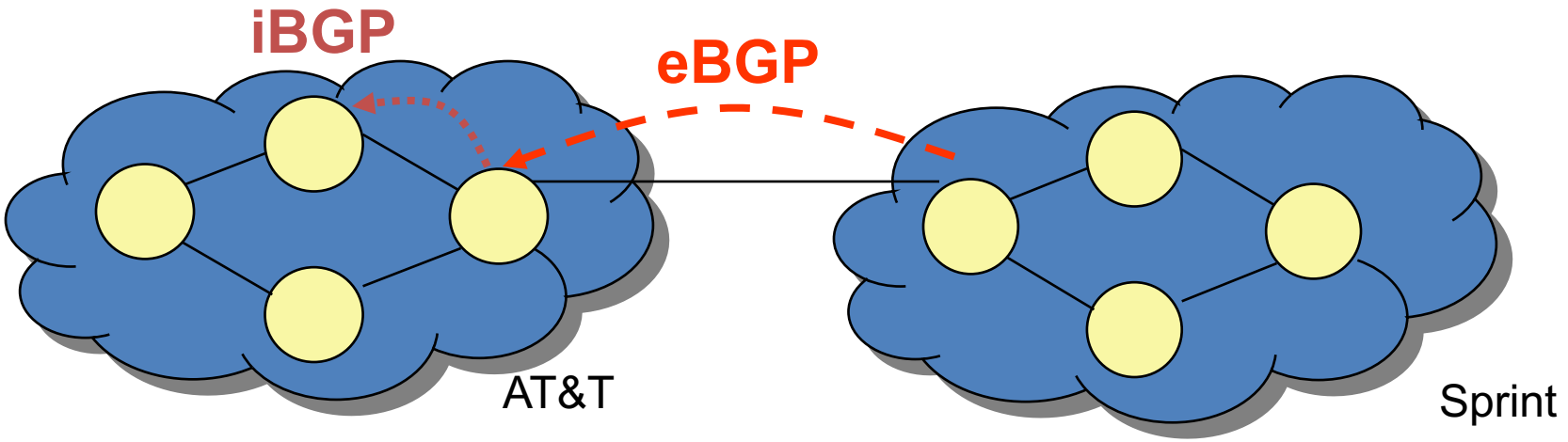
I can reach d via B,D



I can reach d
Via A,B,D

I can reach d
Via C,A,B,D

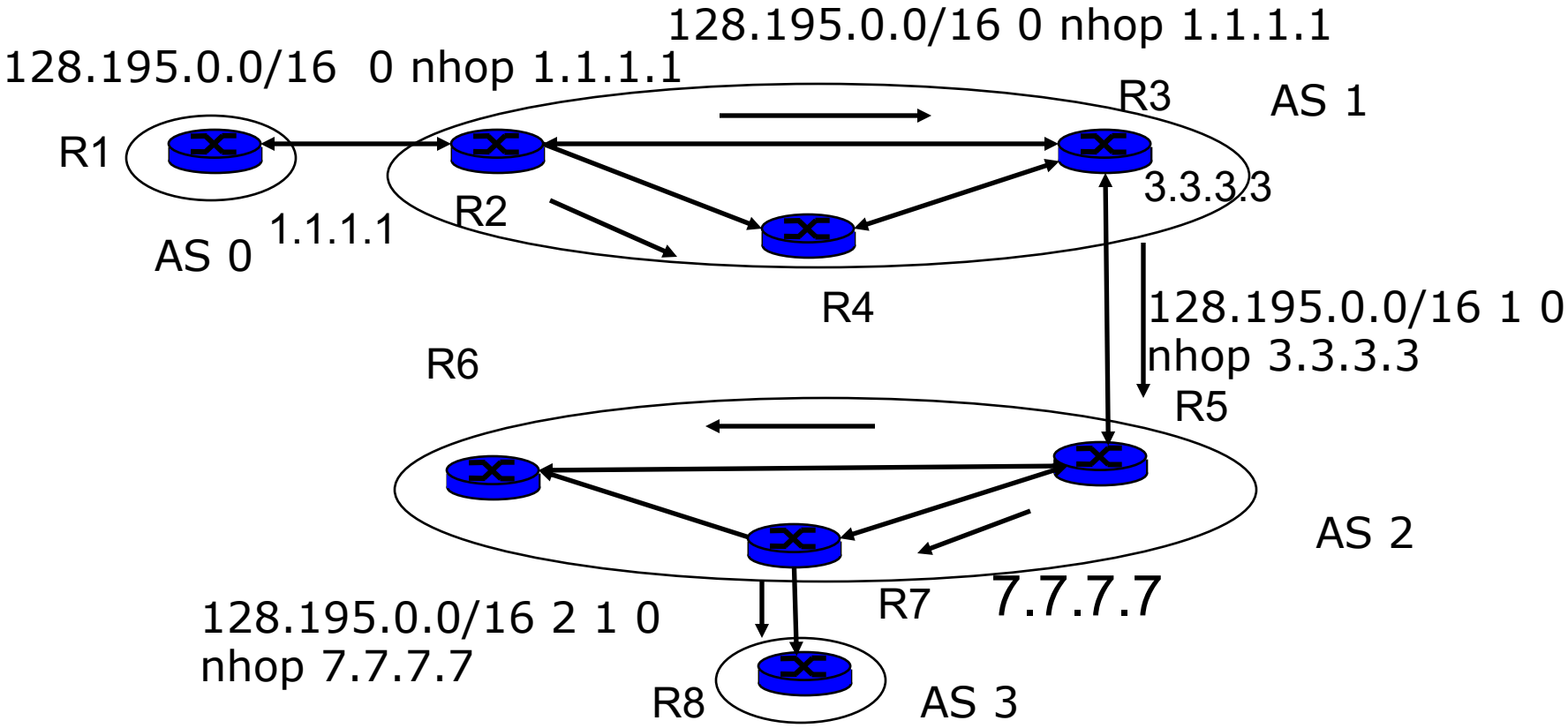
Two types of BGP sessions



- eBGP session is a BGP session between two routers in different ASes

- iBGP session is a BGP session between internal routers of an AS.

Route propagation via eBGP and iBGP



- iBGP is organized into a full mesh topology, or iBGP sessions are relayed using a route reflector.

Common BGP path attributes

- **Origin**: indicates how BGP learned about a particular route
  - IGP (internal gateway protocol)
  - EGP (external gateway protocol)
  - Incomplete

- **AS path** :
  - When a route advertisement passes through an autonomous system, the AS number is added to an ordered list of AS numbers that the route advertisement has traversed

- **Next hop**

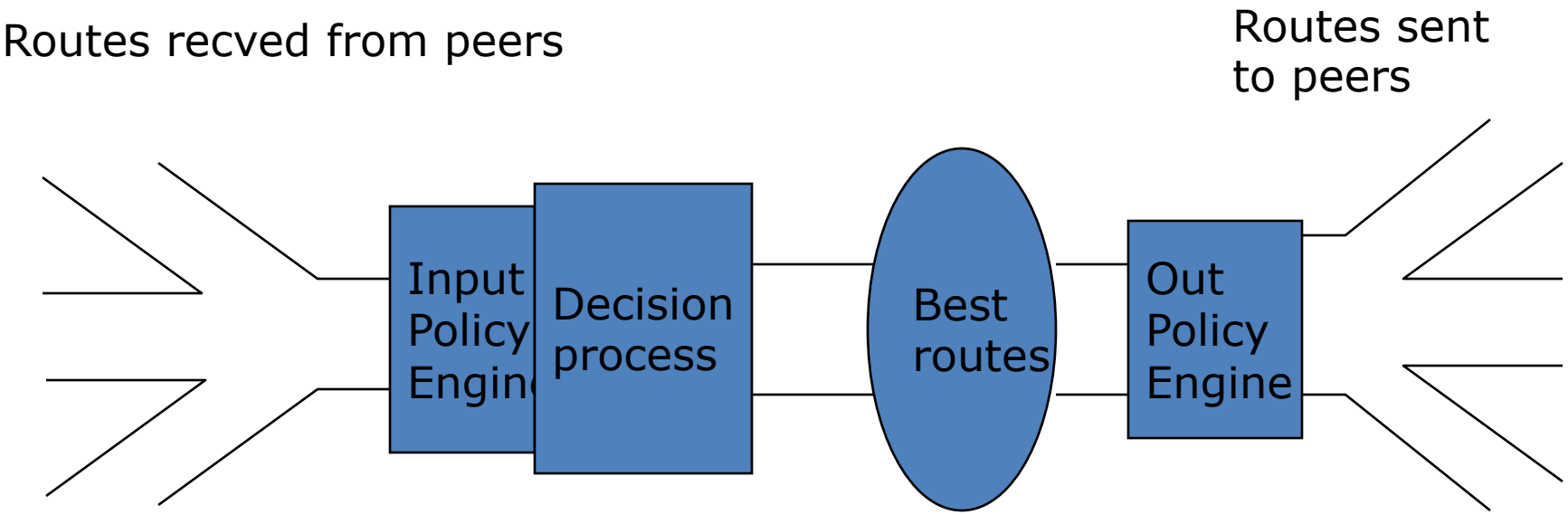
- **Multi_Exit_Disc** (MED, multiple exit discriminator):

-used as a suggestion to an external AS regarding the preferred route into the AS

- **Local_pref**: is used to prefer an exit point from the local autonomous system

- **Community**: apply routing decisions to a group of destinations

BGP route selection process

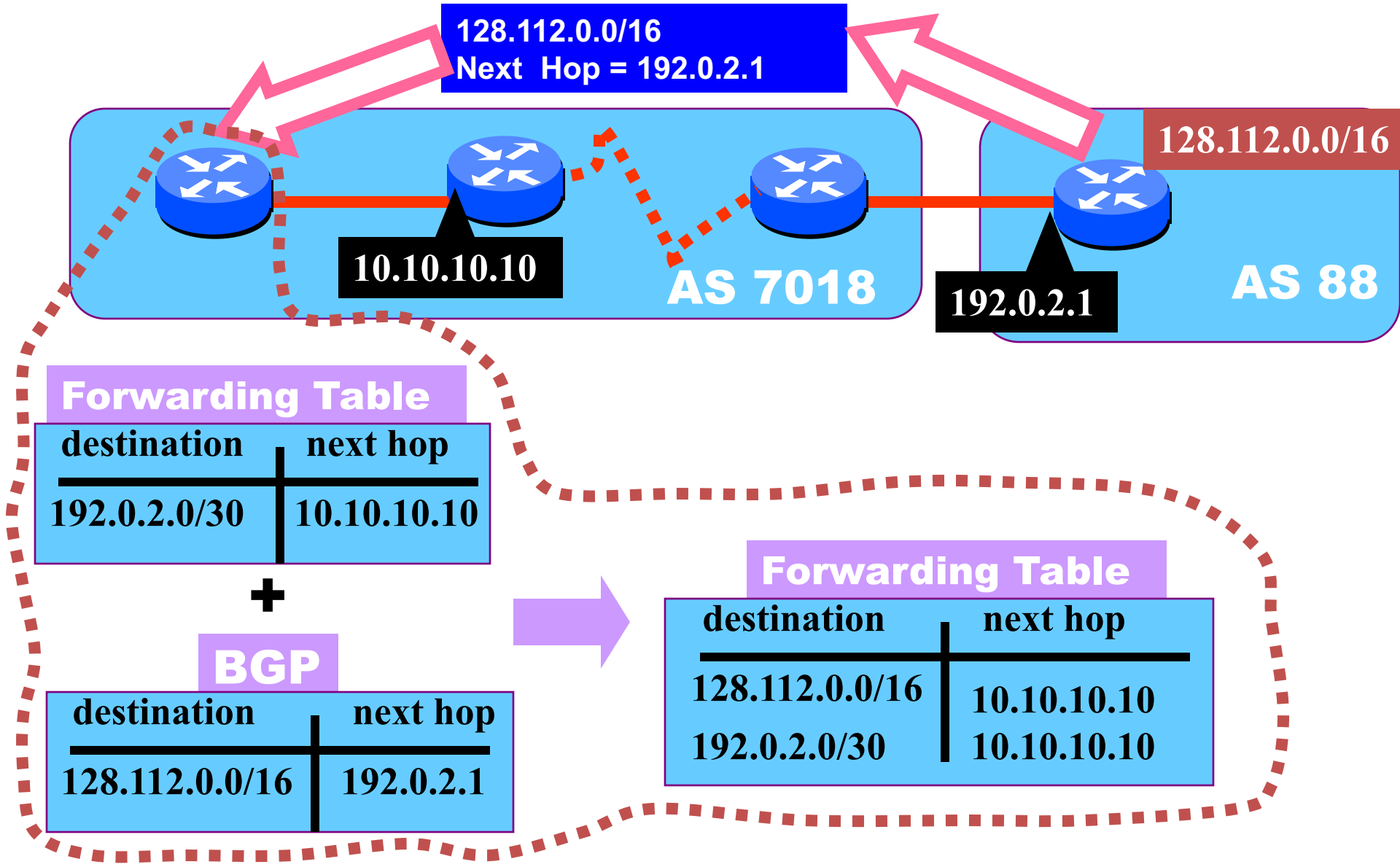


- Input/output engine may filter routes or manipulate their attributes

Best path selection algorithm

1. If next hop is inaccessible, ignore routes
2. Prefer the route with the largest local preference value
3. If local prefs are the same, prefer route with the shortest AS path
4. If AS_path is the same, prefer route with lowest origin (IGP < EGP < incomplete)
5. If origin is the same, prefer the route with lowest MED
6. IF MEDs are the same, prefer eBGP paths to iBGP paths
7. If all the above are the same, prefer the route that can be reached via the closest IGP neighbor
8. If the IGP costs are the same, prefer the router with lowest router id

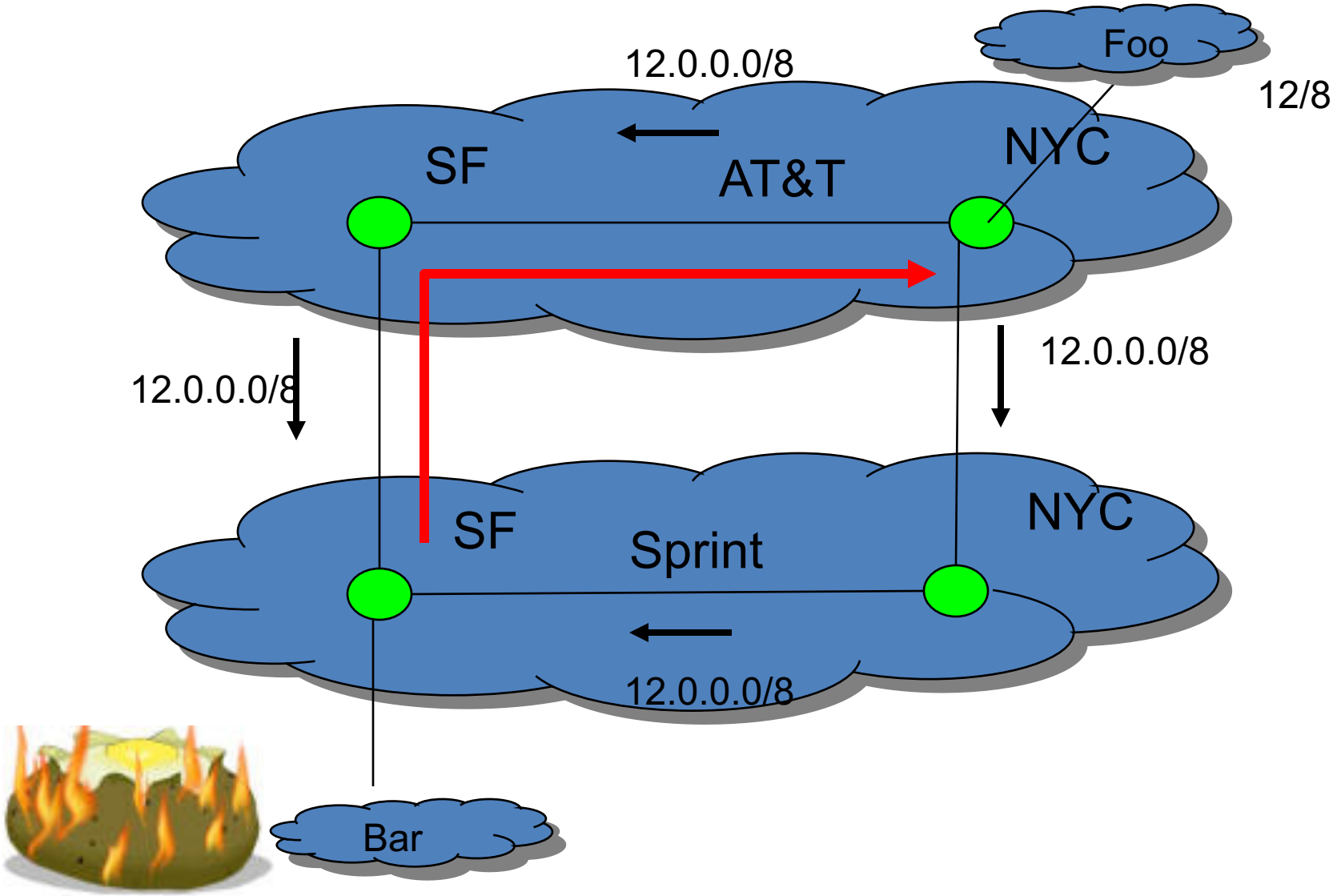
Joining BGP with IGP Information



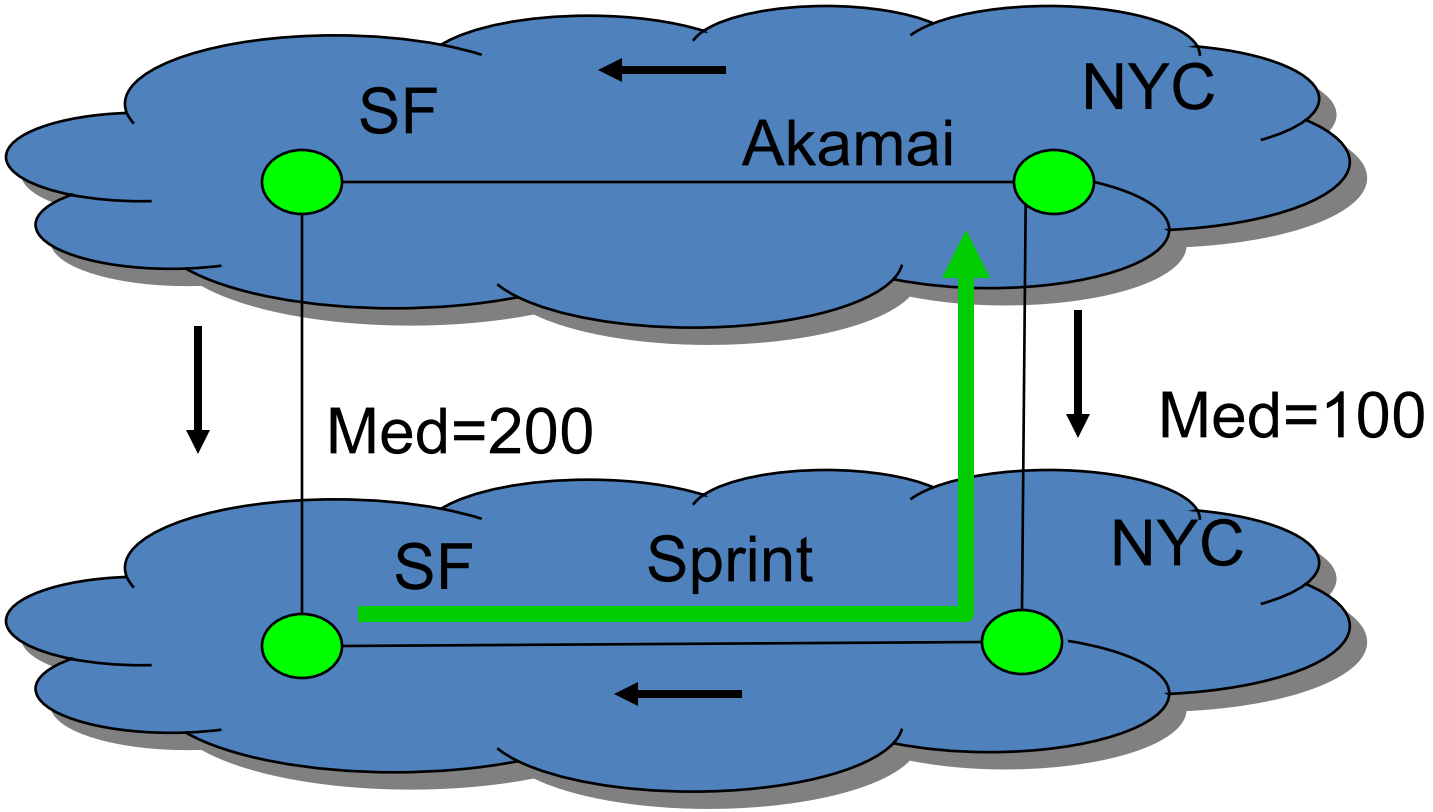
Load balancing

- Same route from two providers
- Outbound is “easy” (you have control)
  - Set localpref according to goals
- Inbound is tough (nobody has to listen)
  - AS path prepending
  - MEDs
    - Hot and Cold Potato Routing (picture)
    - Often ignored unless contracts involved
    - Practical use: tier-1 peering with a content provider

Hot-Potato Routing (early exit)

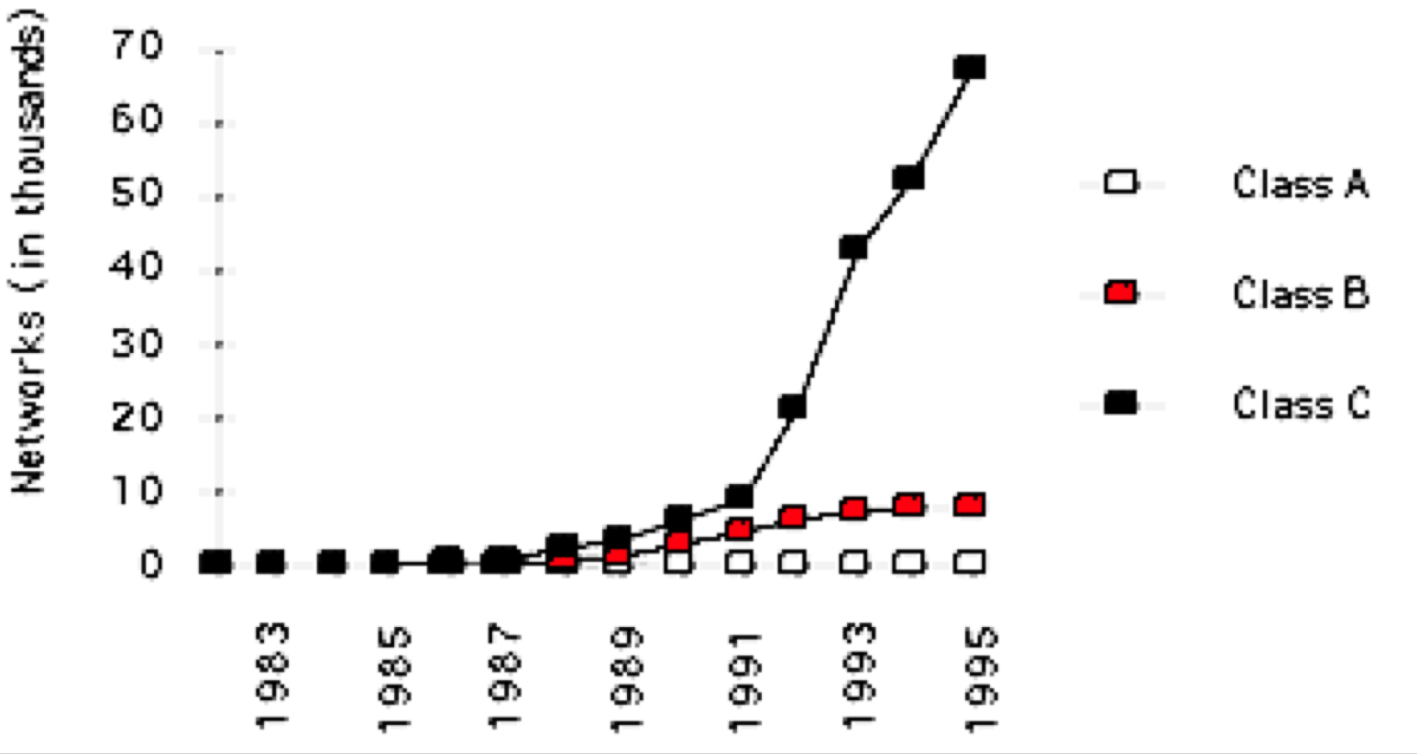


Cold-Potato Routing (MED)



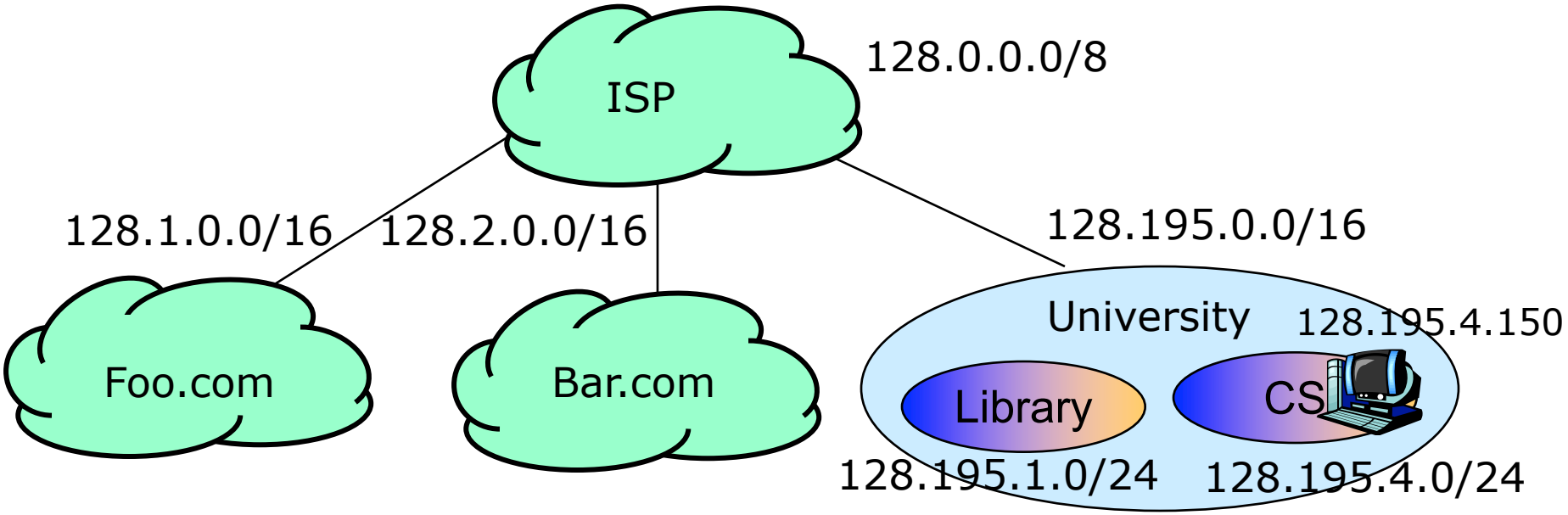
BGP Scalability

Routing table scalability with Classful IP Addresses



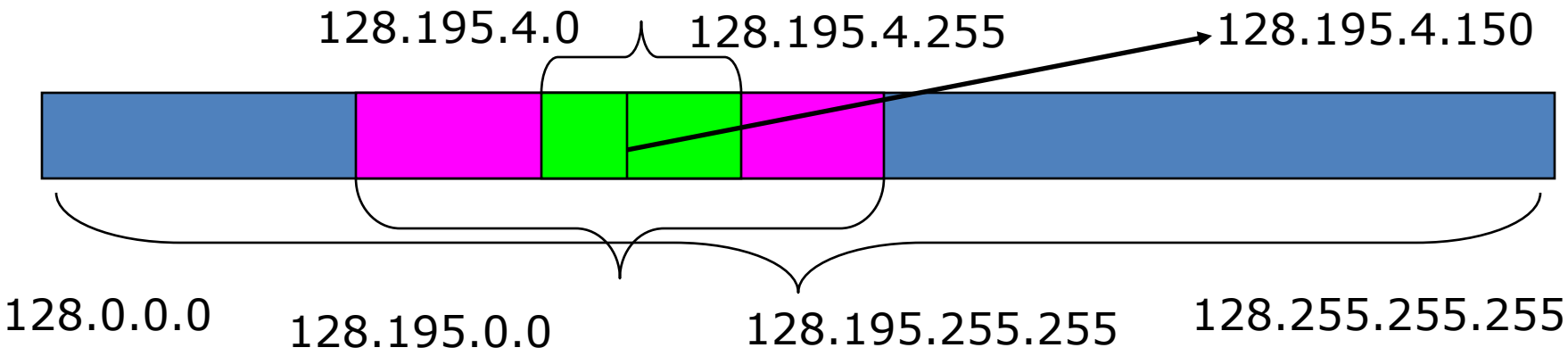
- Fast growing routing table size
- Classless inter-domain routing aims to address this issue

CIDR hierarchical address allocation



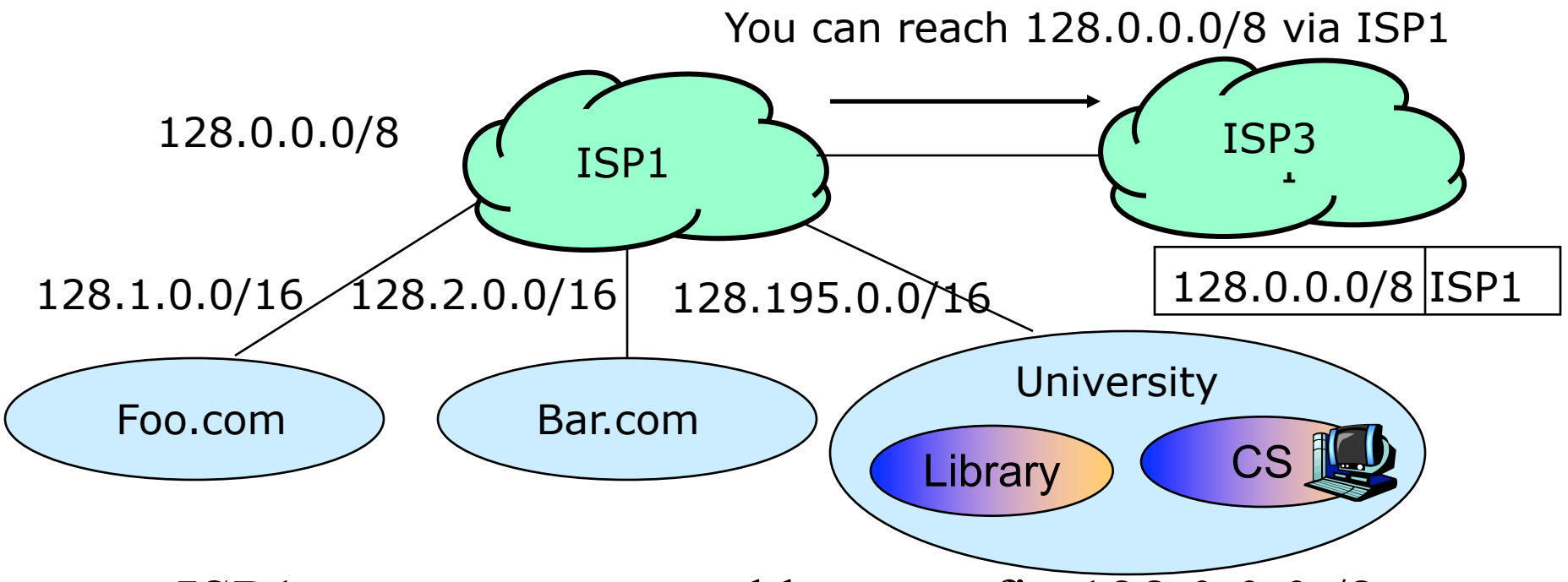
- IP addresses are hierarchically allocated.
- An ISP obtains an address block from a Regional Internet Registry
- An ISP allocates a subdivision of the address block to an organization
- An organization recursively allocates subdivision of its address block to its networks
- A host in a network obtains an address within the address block assigned to the network

Hierarchical address allocation



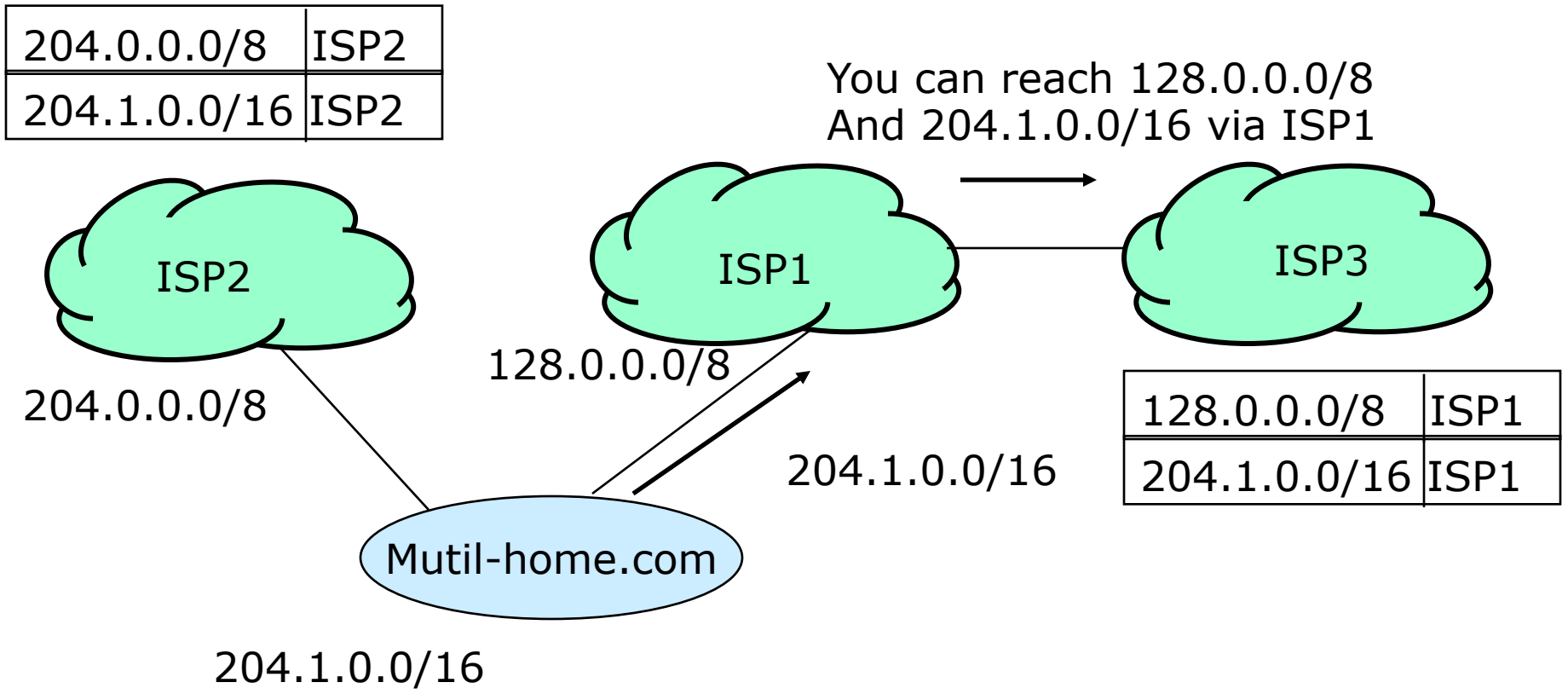
- ISP obtains an address block 128.0.0.0/8 → [128.0.0.0, 128.255.255.255]
- ISP allocates 128.195.0.0/16 ([128.195.0.0, 128.195.255.255]) to the university.
- University allocates 128.195.4.0/24 ([128.195.4.0, 128.195.4.255]) to the CS department's network
- A host on the CS department's network gets one IP address 128.195.4.150

CIDR allows route aggregation

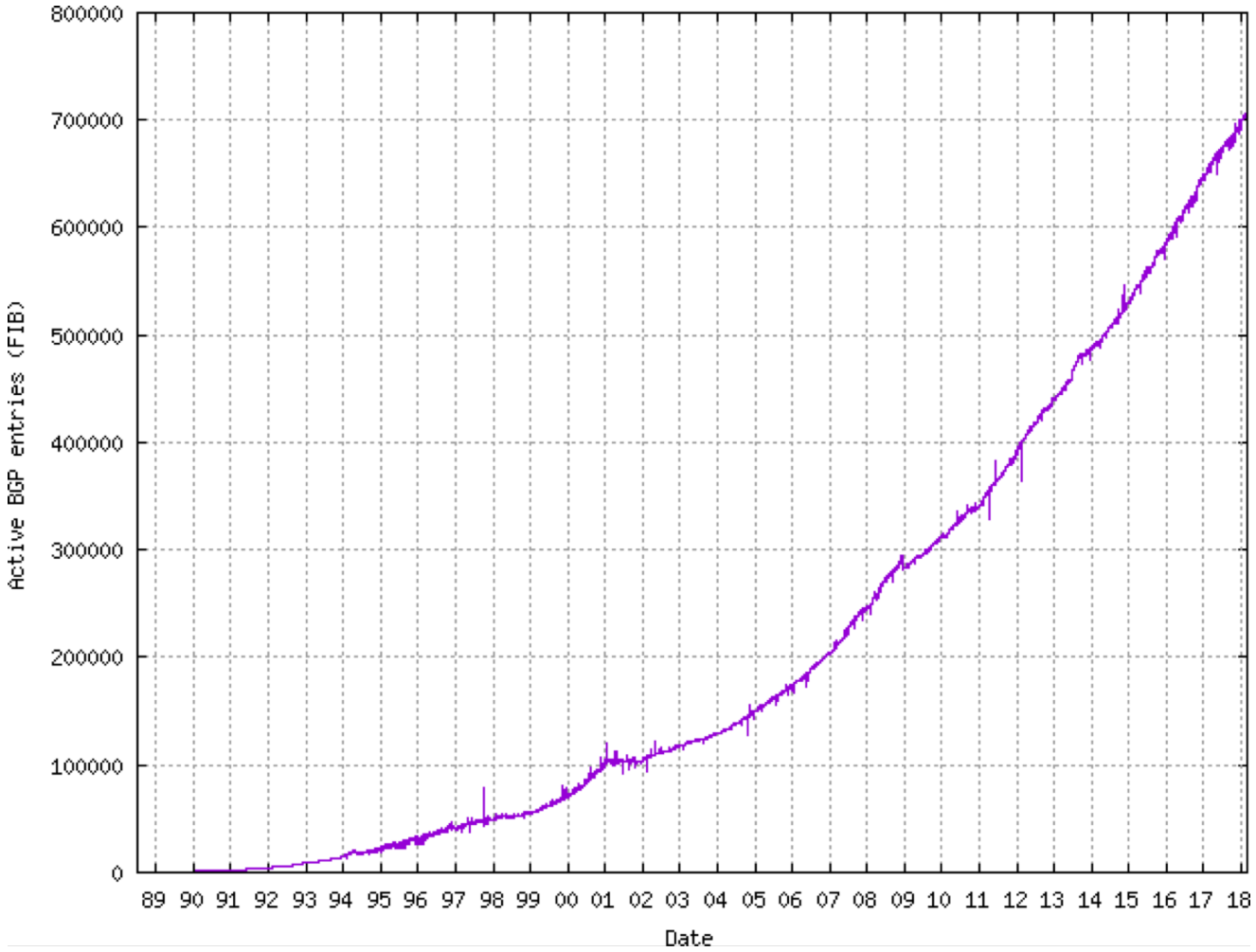


- ISP1 announces one address prefix 128.0.0.0./8 to ISP2
- ISP2 can use one routing entry to reach all networks connected to ISP1

Multi-homing increases routing table size



Global routing tables continue to grow (1989-now)



Source: https://www.cidr-report.org

BGP Summary

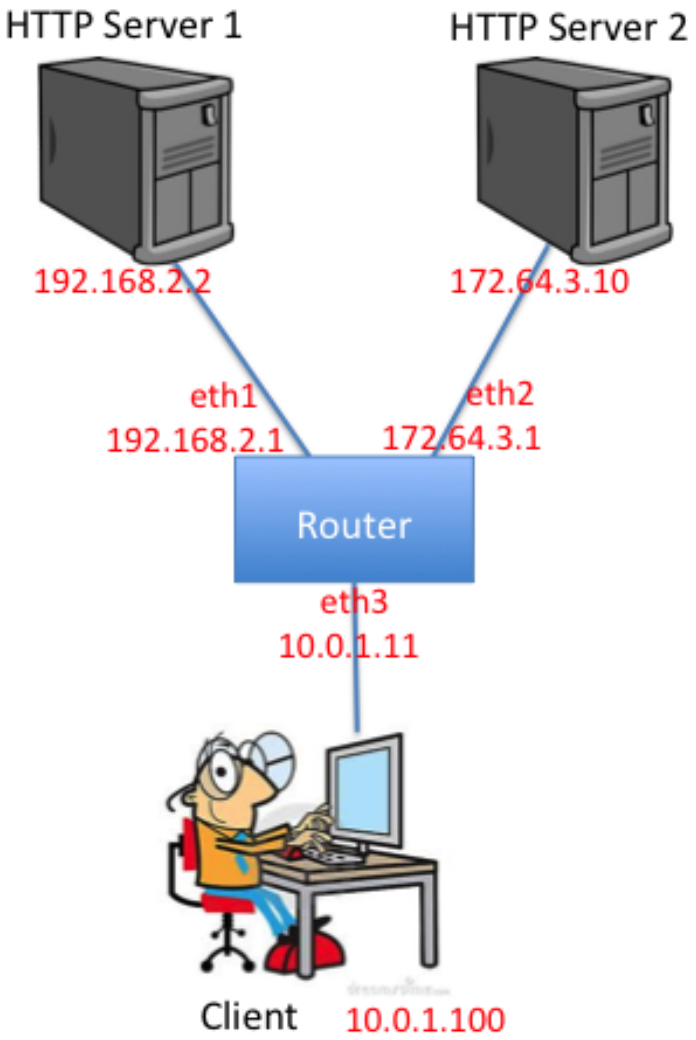
- BGP uses the path vector algorithm
- Its path selection algorithm is complicated
- Policy is mostly determined by economic considerations

Lab2 – Simple Router

COMPSCI 356

2019sp

Topology


HTTP Server 1
HTTP Server 2
192.168.2.2
172.64.3.10
eth1
192.168.2.1
eth2
172.64.3.1
Router
eth3
10.0.1.11
Client
10.0.1.100

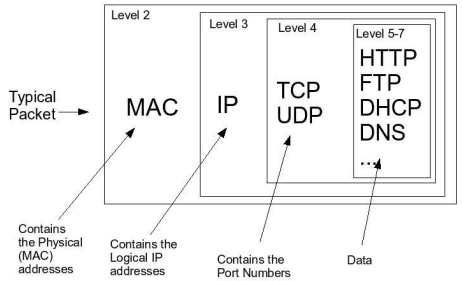
Overview

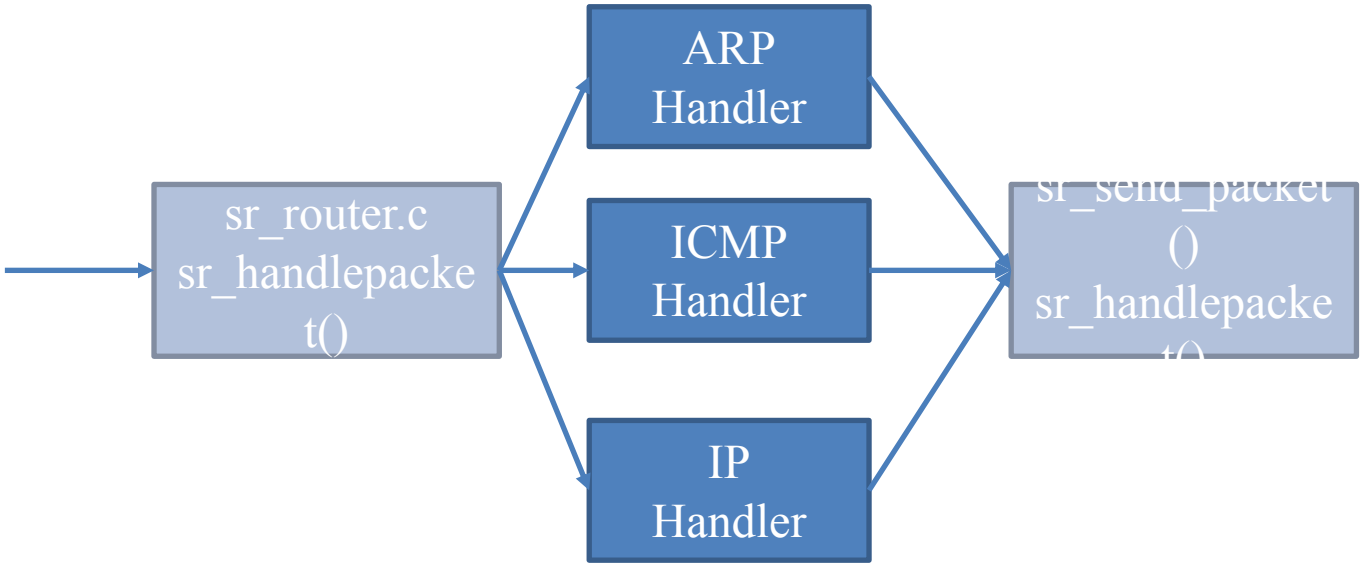
Your task is to implement a simple router with a static routing table. It will be able to do the following:

- The router will handle raw Ethernet frames;
- It will process the packets just like a real router;
- then forward them to the correct outgoing interface.

Packet Handling Procedure

- In the aspect of Router





https://tournasdimitrios1.wordpress.com/2011/01/19/the-basics-of-network-packets/

What you need to implement

- sr_arpcache.c sr_arpcache_sweepreqs(struct sr_instance *sr)
  - The assignment requires you to send an ARP request about once a second until a reply comes back or we have sent five requests. This function is defined in sr_arpcache.c and called every second, and you should add code that iterates through the ARP request queue and re-sends any outstanding ARP requests that haven't been sent in the past second. If an ARP request has been sent 5 times with no response, a destination host unreachable should go back to all the sender of packets that were waiting on a reply to this ARP request.

- sr_router.c sr_handlepacket(struct sr_instance* sr, …)
  - This method, located in sr_router.c, is called by the router each time a packet is received. The "packet" argument points to the packet buffer which contains the full packet including the ethernet header. The name of the receiving interface is passed into the method as well.

Helper Functions-1

arpcache.c

sr_arpcache_lookup(struct sr_arpcache *cache, uint32_t ip)

look for the MAC address in cache based on ip

sr_arpcache_dump(struct sr_arpcache *cache)
print the list of current ARP cache

sr_if.c

sr_get_interface(struct sr_instance* sr, const char* name)

get the property of specific interface by its name

sr_print_if_list(struct sr_instance* sr)
print the list of interfaces in current router

sr_protocol.h

header definition
define the header information

Helper Functions-2

sr_rt.c **sr_print_routing_table(struct sr_instance* sr)**

print out the content of routing table

sr_print_routing_entry(struct sr_rt* entry)

print out the verbose information of a specific routing entry

sr_utils.c **cksum (const void *_data, int len)**

calculate the checksum of a range of packet content

print_hdrs(uint8_t *buf, uint32_t length)

print the content of a network packet header

ICMP packet format

- http://www.networksorcery.com/enp/protocol/icmp.htm

MAC header	IP header	ICMP header	Data :::

ICMP header:

00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Type								Code								ICMP header checksum															
Data :::																															

Type. 8 bits.
Specifies the format of the ICMP message.

Type	Description	References
0	Echo reply.	RFC 792
1		
2		
3	Destination unreachable.	RFC 792
4	Source quench.	RFC 792
5	Redirect.	RFC 792
6	Alternate host address.	
7		
8	Echo request.	RFC 792
9	Router advertisement.	RFC 1256
10	Router solicitation.	RFC 1256
11	Time exceeded.	RFC 792
12	Parameter problem.	RFC 792
13	Timestamp request	RFC 792