15-441 Computer Networks

Inter-Domain Routing

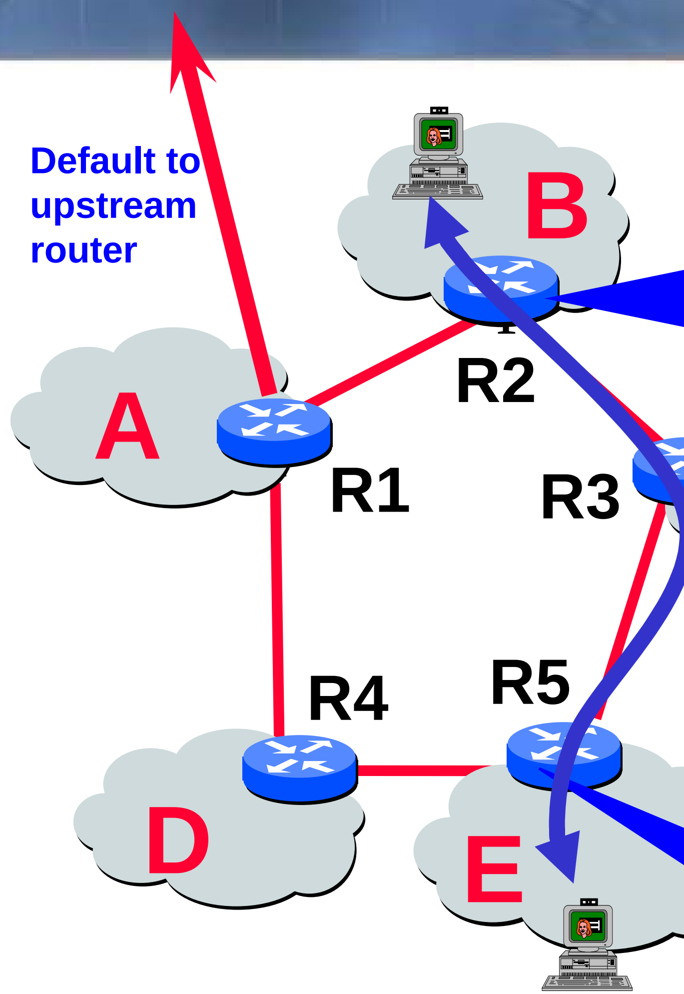
Professor Hui Zhang

hzhang@cs.cmu.edu

To Learn More About BGP

- http://www.cambridge.intel-research.net/~tgriffin/talk_tutoriasl/

Routing vs. Forwarding



Net	Nxt Hop
A	R1
B	Direct
C	R3
D	R1
E	R3
default	R1

Forwarding always works

Routing can be badly broken

Net	Nxt Hop
A	R2
B	R2
C	Direct
D	R5
E	R5
default	R2

Net	Nxt Hop
A	R4
B	R3
C	R3
D	R4
E	Direct
default	R4

Forwarding: determine next hop

Routing: establish end-to-end paths

Hui Zhang

Two Classes of Routing Protocols

- **Distance vector (RIP)**
  - Distribute path computation
  - Keep only local link data
  - Bellman-Ford algotrithm
- **Link state (OSPF, IS-IS)**
  - Local path computation
  - Distribute all link data
  - Dijkstra's algorithm

Link State vs. Distance Vector

Link State

- Topology information is flooded within the routing domain
- Best end-to-end paths are computed locally at each router.
- Best end-to-end paths determine next-hops.
- Based on minimizing some notion of distance
- Works only if policy is shared and uniform
- Examples: OSPF, IS-IS

Vectoring

- Each router knows little about network topology
- Only best next-hops are chosen by each router for each destination network.
- Best end-to-end paths result from composition of all next-hop choices
- Does not require any notion of distance
- Does not require uniform policies at all routers
- Examples: RIP, BGP

IP Addressing and Forwarding

- **Routing Table Requirement**
  - For every possible destination IP address, give next hop
  - Nearly 2^{32} (4.3×10^9) possibilities!
- **Hierarchical Addressing Scheme**

	x	y
`pfx`	`network`	`host`

- **Address split into network ID and host ID**
  - E.g., CMU has one network ID shared by all hosts within CMU
  - All packets to given network follow same route, until they reach destination network
- **Fields**
  - `pfx` Prefix to specify split between network & host IDs
  - `network` 2^x possibilities
  - `host` 2^y possibilities

IP Address Classes

- **Class A**

First digit: 1–126

	7	24
0	network	host

  - mit.edu: 18.7.22.69

- **Class B**

First digit: 128–191

	14	16
10	network	host

  - cmu.edu: 128.2.11.43

- **Class C**

First digit: 192–223

	21	8
110	network	host

- **Classes D, E, F**
  - Not commonly used

IP Address Classes

Class	Count	Hosts
A	$2^7-2 = 126$	$2^{24}-2 = 16,777,214$
B	$2^{14} = 16,398$	$2^{16}-2 = 65,534$
C	$2^{21} = 2,097,512$	$2^8-2 = 254$
Total	2,114,036	

- **Partitioning too Coarse**
  - Not enough big (class A) addresses
  - No organization needs 16.7 million hosts
    - Large organization likely to be geographically distributed
  - Many organizations must make do with multiple class C's
- **Too many different Network IDs**
  - Routing tables must still have 2.1 million entries

Improving the Hierarchy

- **Basic Idea of Hierarchy is Good**
  - Organizations of different sizes can be assigned different numbers of IP addresses
- **Shortcomings of Class-Based Addressing**
  - Class A too coarse; Class C too fine; not enough Class B's
  - When fully deployed would have too many entries in routing table (2.1 million)
- **Solution**
  - Hierarchy with finer gradation of network/host ID split

Classless Interdomain Routing

- CIDR, pronounced “cider”

❖ Arbitrary Split Between Network & Host IDs

- Specify either by mask or prefix length

`1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0`

network	host

- E.g., CMU can be specified as
  - – 128.2.0.0 with netmask 255.255.0.0
  - – 128.2.0.0/16

Aggregation with CIDR

- Original Use: Aggregate Class C Addresses
- One organization assigned contiguous range of class C's
  - e.g., Microsoft given all addresses 207.46.192.X -- 207.46.255.X
  - Specify as CIDR address 207.46.192.0/18

0	8	16	24 31	
207	**46**	**192**	**0**	**Decimal**
cf	**2e**	**c0**	**00**	**Hexadecimal**
1100 1111	**0010 1110**	**11xx xxxx**	**xxxx xxxx**	**Binary**

$\underbrace{\text{0–17}}_{\textbf{Upper 18 bits frozen}}$ $\underbrace{\text{18–31}}_{\textbf{Lower 14 bits arbitrary}}$

Routing Table Entry Examples

- Snapshot From MAE-West Routing Table

Address	Prefix Length	Third Byte	Byte Range
207.46.0.0	19	$000xxxxx_2$	0 – 31
207.46.32.0	19	$001xxxxx_2$	32 – 63
207.46.64.0	19	$010xxxxx_2$	64 – 95
207.46.128.0	18	$10xxxxxx_2$	128 – 191
207.46.192.0	18	$11xxxxxx_2$	192 – 255

microsoft.com: 207.46.245.214 & 207.46.245.222

- Note hole in table: Nothing covers bytes 96 – 127

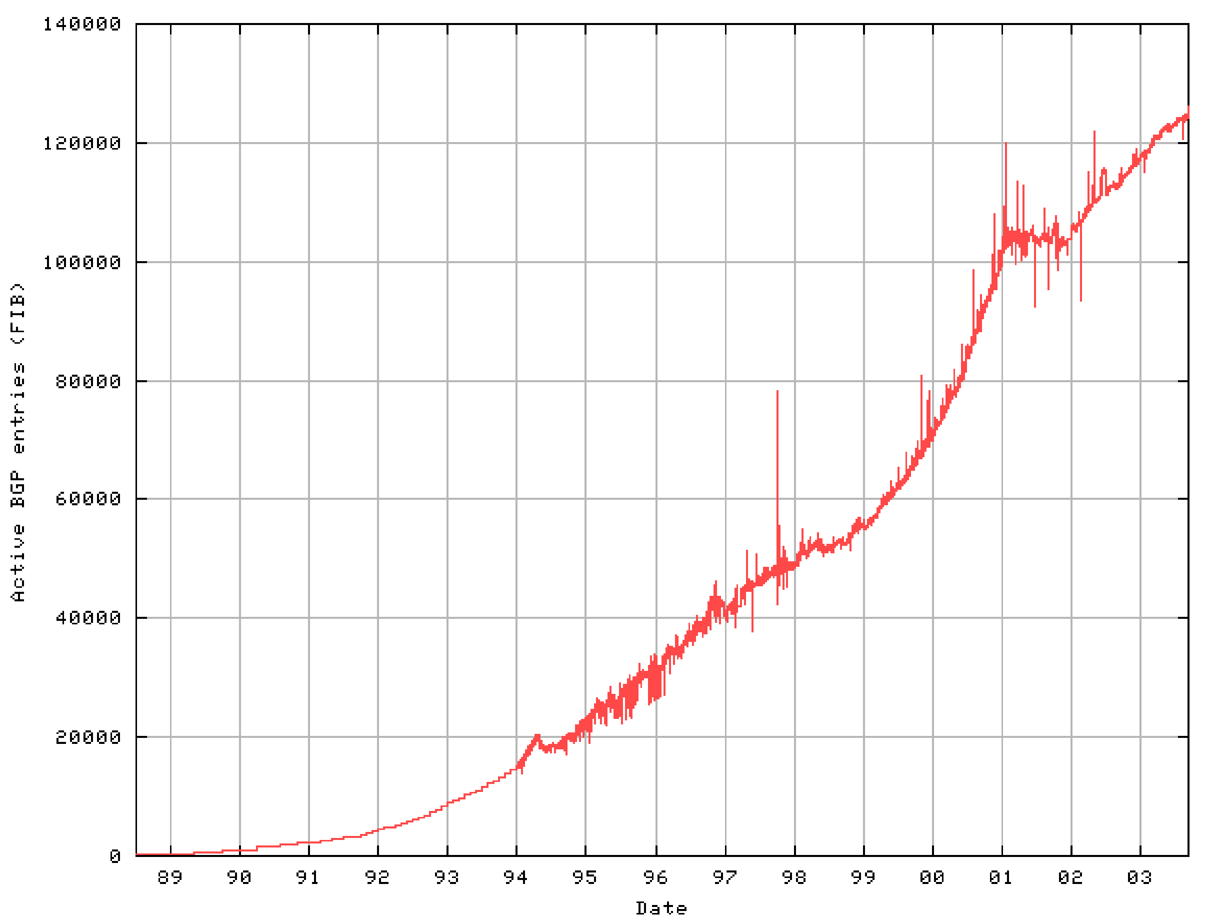
Splitting with CIDR

- Expose subnetting structure to external routers

Example

- Class A address 12.X.X.X has 413 entries in MAE-WEST table
- Prefix lengths 8--24
- attbi.com
  - Backbone services of AT&T
- Geographically distributed
  - Don't want all packets to concentrate to single region

Size of Complete Routing Table



- Shows that CIDR has kept # table entries in check
  - Currently require 124,894 entries for a complete table
  - Only required by backbone routers

IPv6 Addressing

- Main motivation for switch from IPv4
- Getting hard to manage 32-bit address allocation

128-Bit Addresses

- Standard unicast addresses 125 bits long (3-bit prefix)
- 4.2×10^{37} nodes
- Earth radius is 6371 km
- Metric: $4.2 \times 10^{37} / [4 \pi (6.371 \times 10^{8})^{2}] = 8 \times 10^{18}$ nodes / cm^2

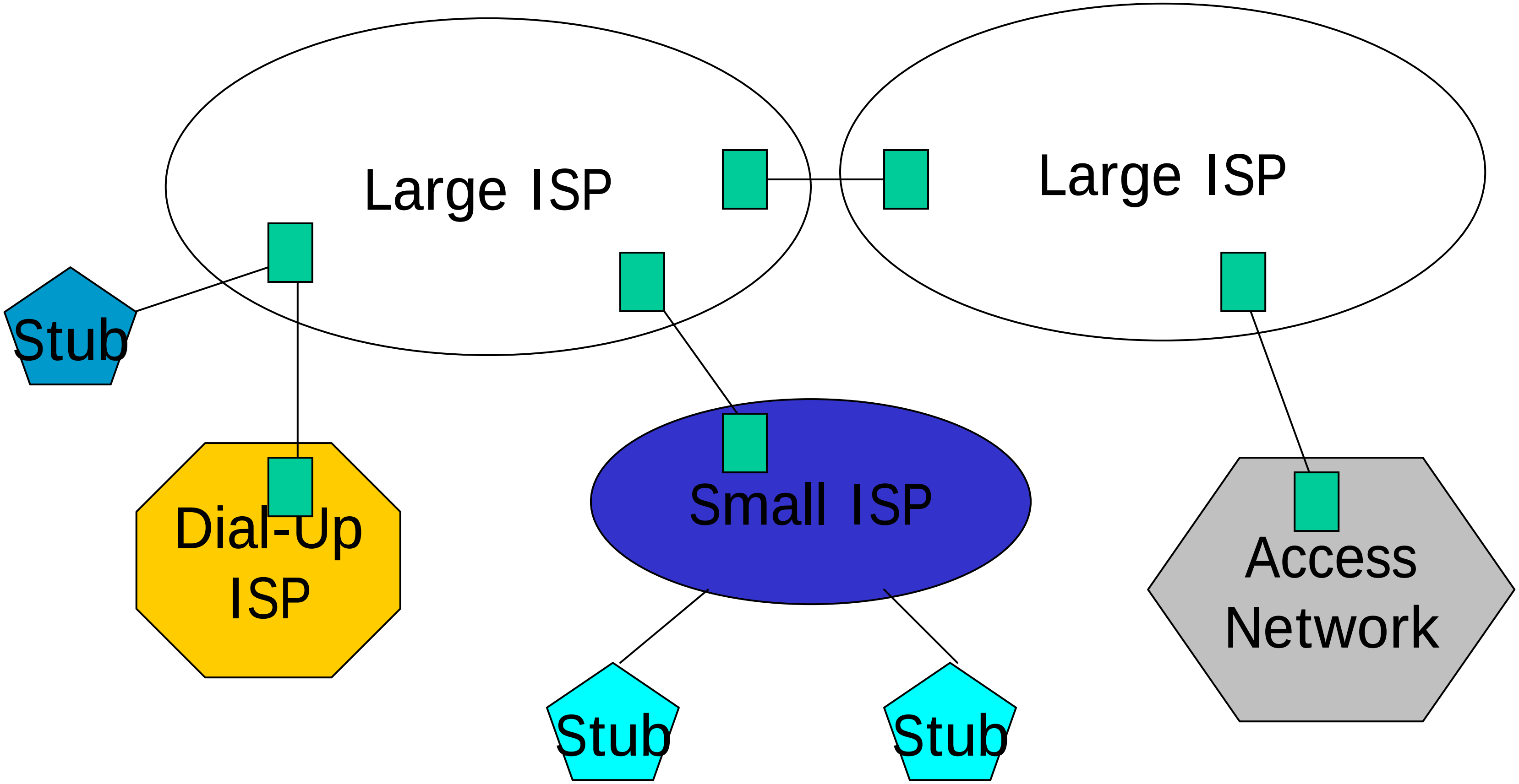
Name, Address, Route

- **First order approximation**
  - Name tells who you are
    - www.cnn.com
  - Address tells where you are
    - 64.236.16.20
  - Route tells how to get there
    - <prefix, nextHop>

Name, Address, Route

- How many names each address can have?
- How many addresses each name can have?
- How many addresses each “node” can have?
- How many names each “node” can have?
- How many routes to each prefix?

Internet Structure



The Internet contains a large number of diverse networks

Routing between ISPs

- **Routing protocol (BGP) contains reachability information (no metrics)**
  - Not about optimizing anything
  - All about policy (business and politics)
- **Why?**
  - Metrics optimize for a particular criteria
  - AT&T's idea of a good route is not the same as UUnet's
  - Scale

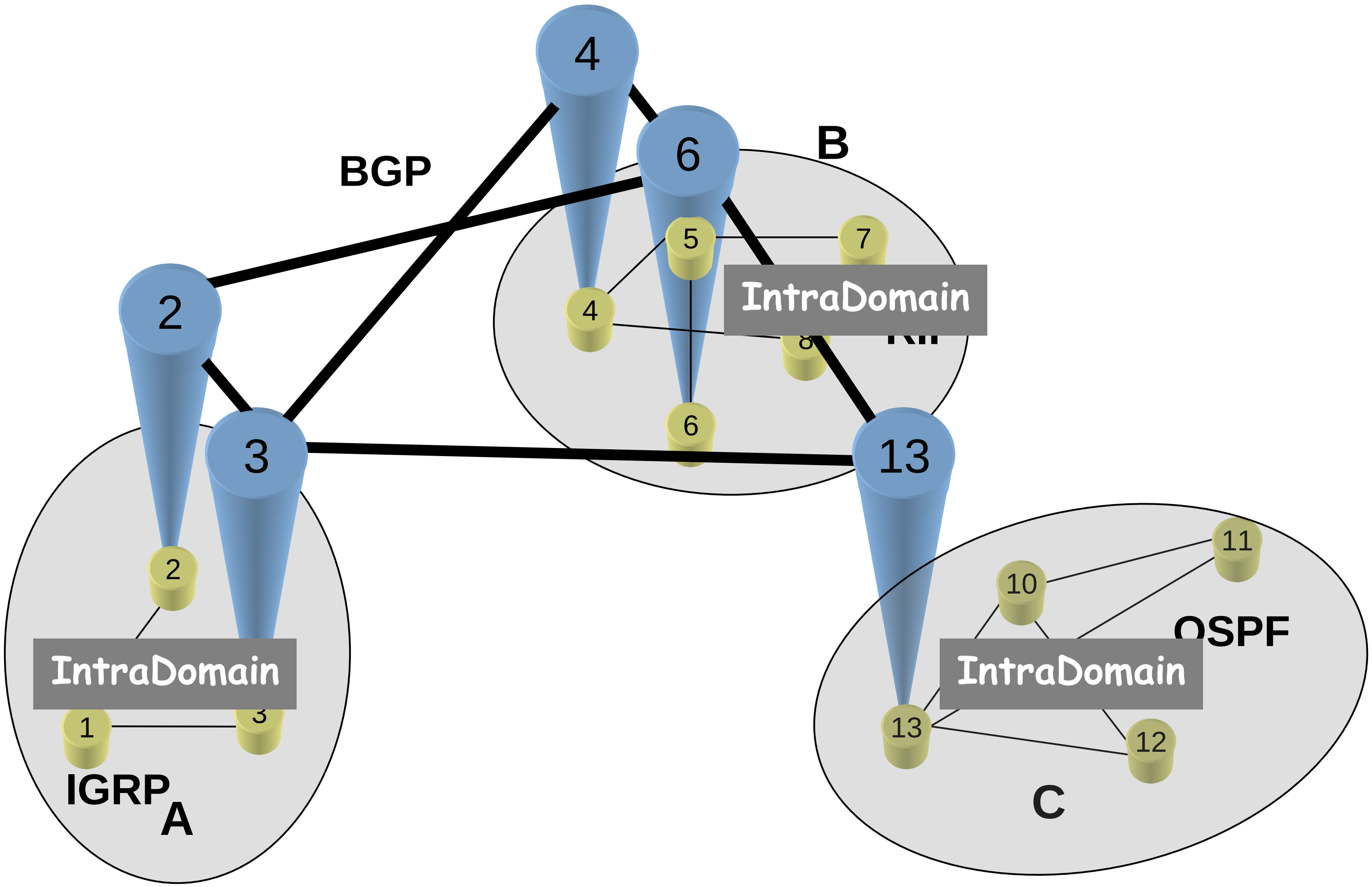
Autonomous Systems (AS)

- Internet is not a single network!
- The Internet is a collection of networks, each controlled by different administrations
- An autonomous system (AS) is a network under a single administrative control

Implications

- **ASs want to choose own local routing algorithm**
  - AS takes care of getting packets to/from their own hosts
  - Interdomain routing and Intradomain routing
- **ASs want to choose own nonlocal routing policy**
  - Interdomain routing must accommodate this
  - BGP is the current interdomain routing protocol

Intradomain And Interdomain



Hui Zhang

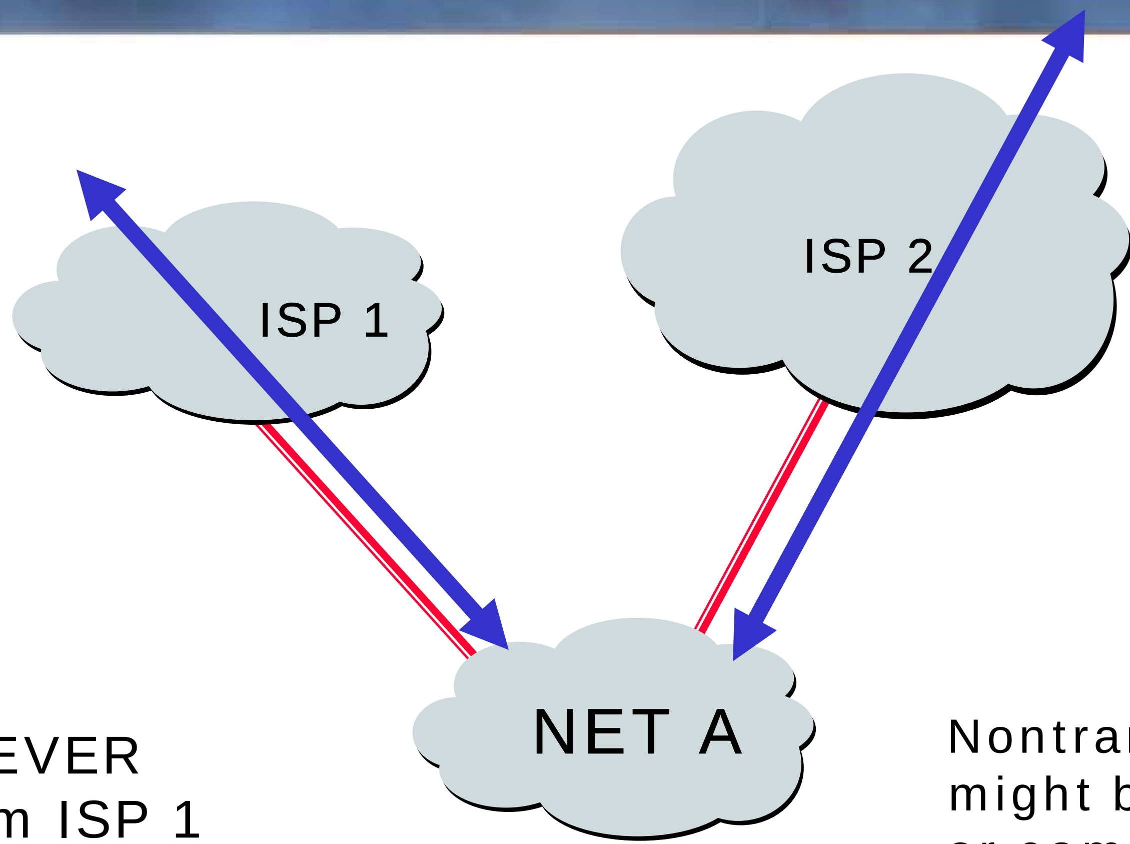
AS Numbers (ASNs)

ASNs are 16 bit values.
64512 through 65535 are "private"

Currently over 11,000 in use.

- Genuity: 1
- CMU: 9
- Harvard: 11
- UC San Diego: 7377
- AT&T: 7018, 6341, 5074, ...
- UUNET: 701, 702, 284, 12199, ...
- Sprint: 1239, 1240, 6211, 6242, ...
- ...

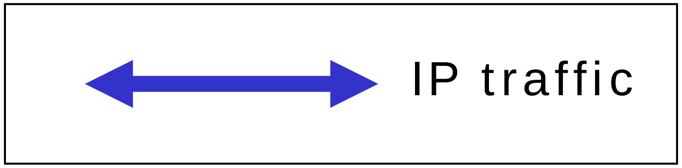
Nontransit vs. Transit ASes



Internet Service providers (often) have transit networks

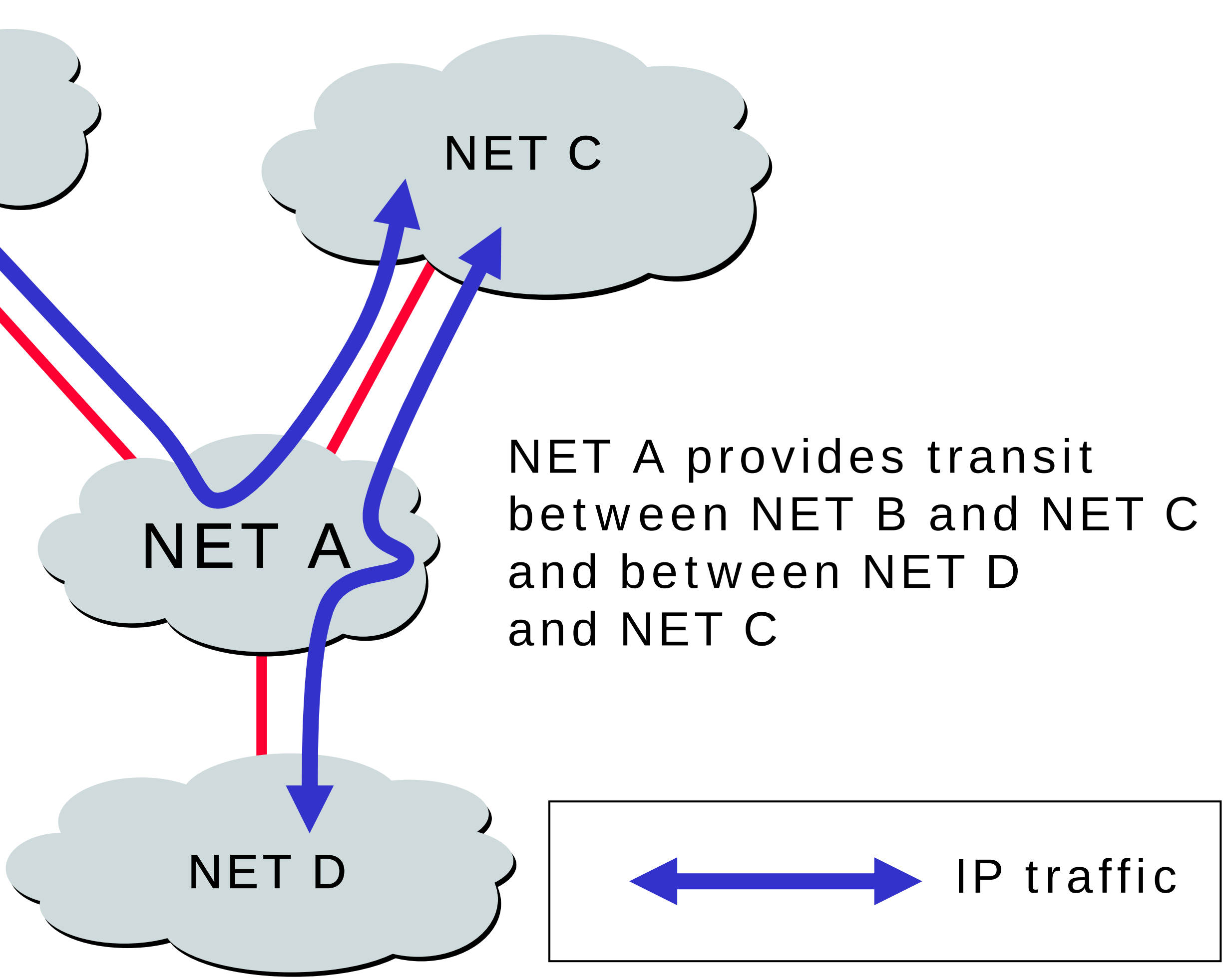
Traffic NEVER flows from ISP 1 through NET A to ISP 2 (At least not intentionally!)

Nontransit AS might be a corporate or campus network. Could be a “content provider”



Hui Zhang

Selective Transit



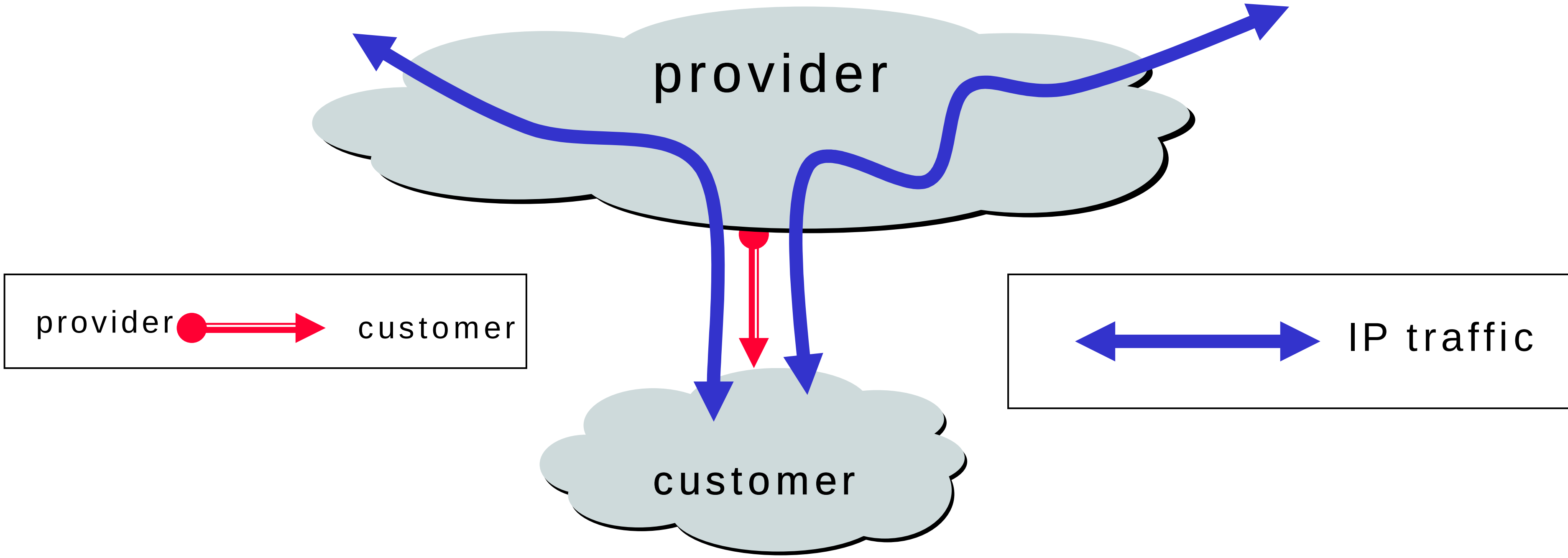
NET A DOES NOT provide transit Between NET D and NET B

NET A provides transit between NET B and NET C and between NET D and NET C

Most transit networks transit in a selective manner...

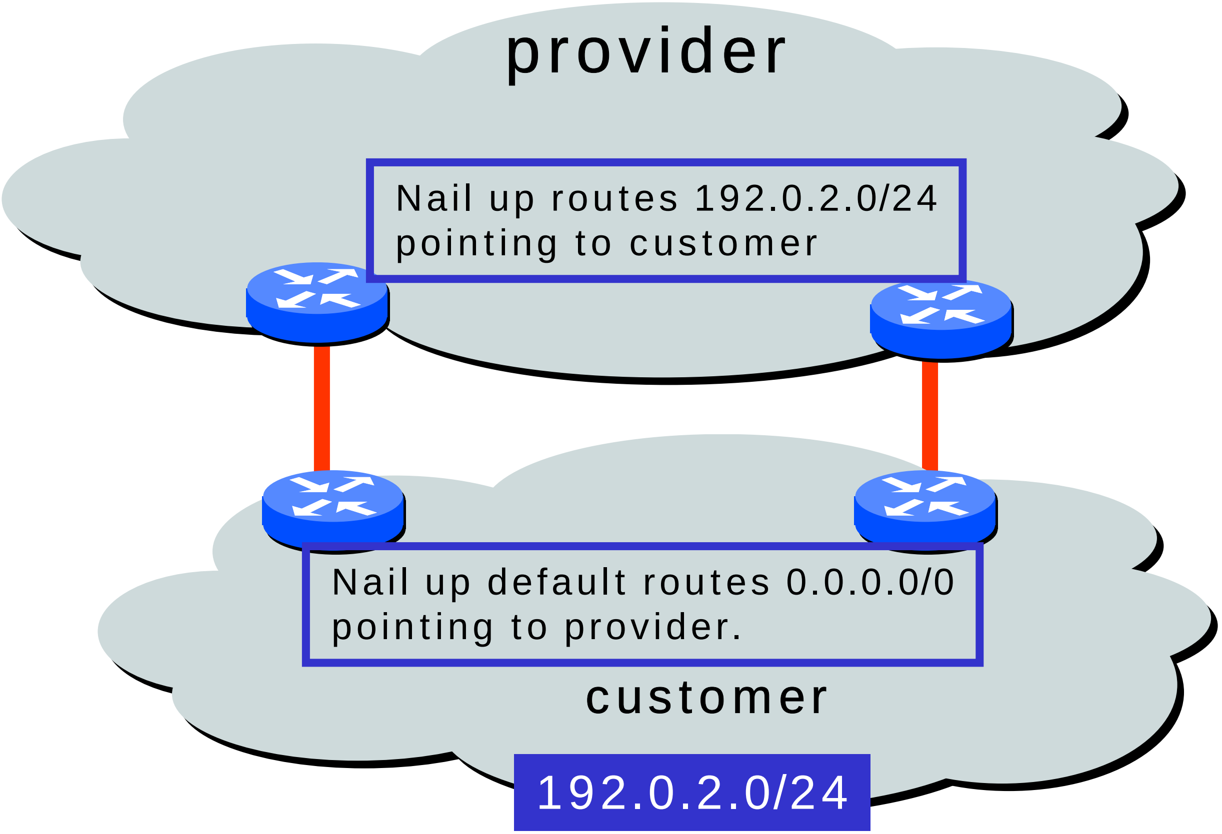
Hui Zhang

Customers and Providers



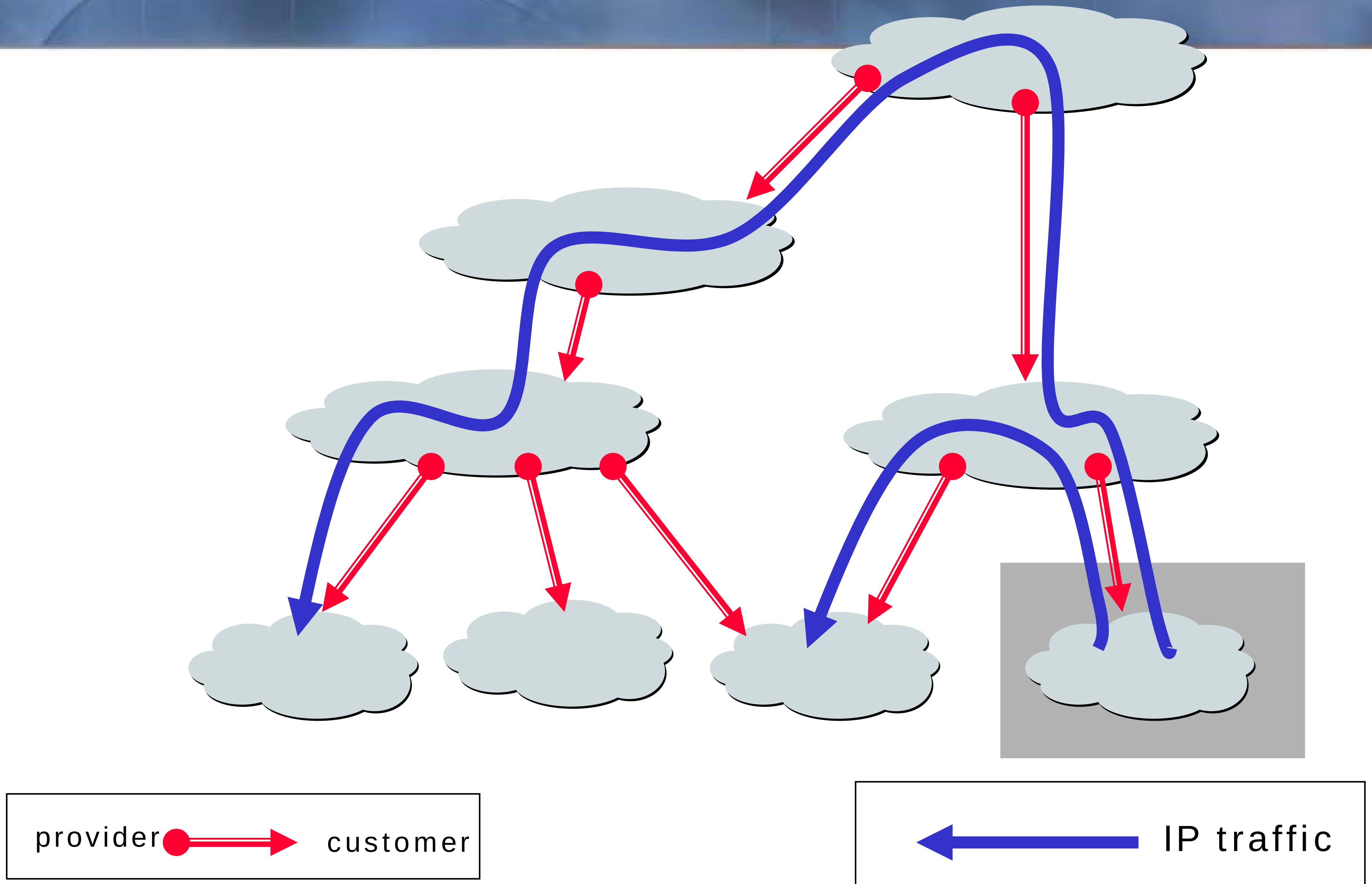
Customer pays provider for access to the Internet

Customers Don't Always Need BGP

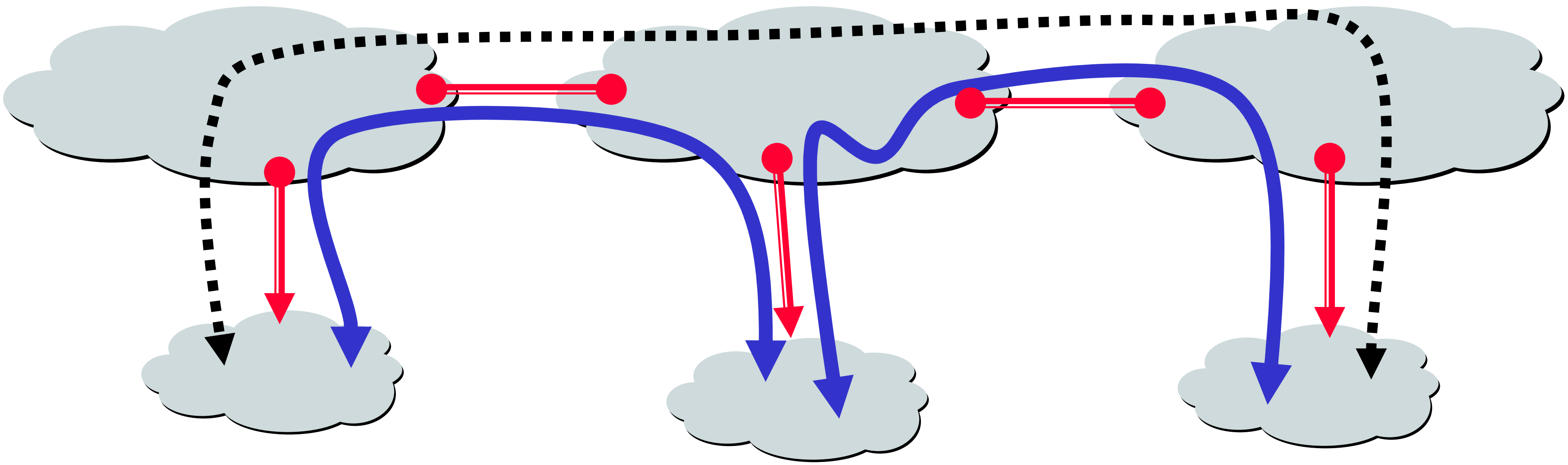


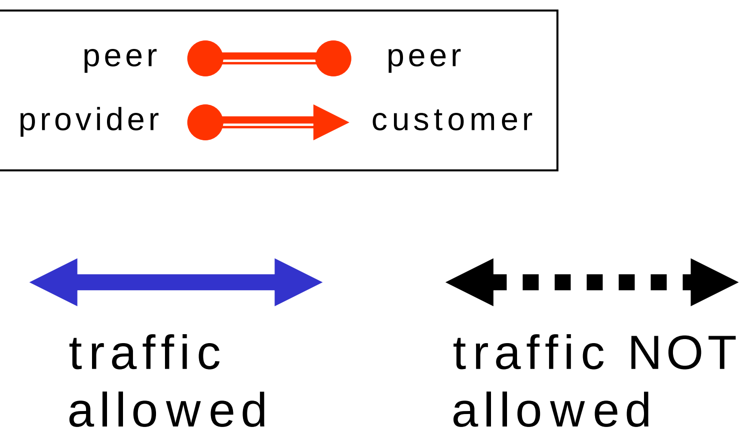
Static routing is the most common way of connecting an autonomous routing domain to the Internet.
This helps explain why BGP is a mystery to many ...

Customer-Provider Hierarchy



The Peering Relationship





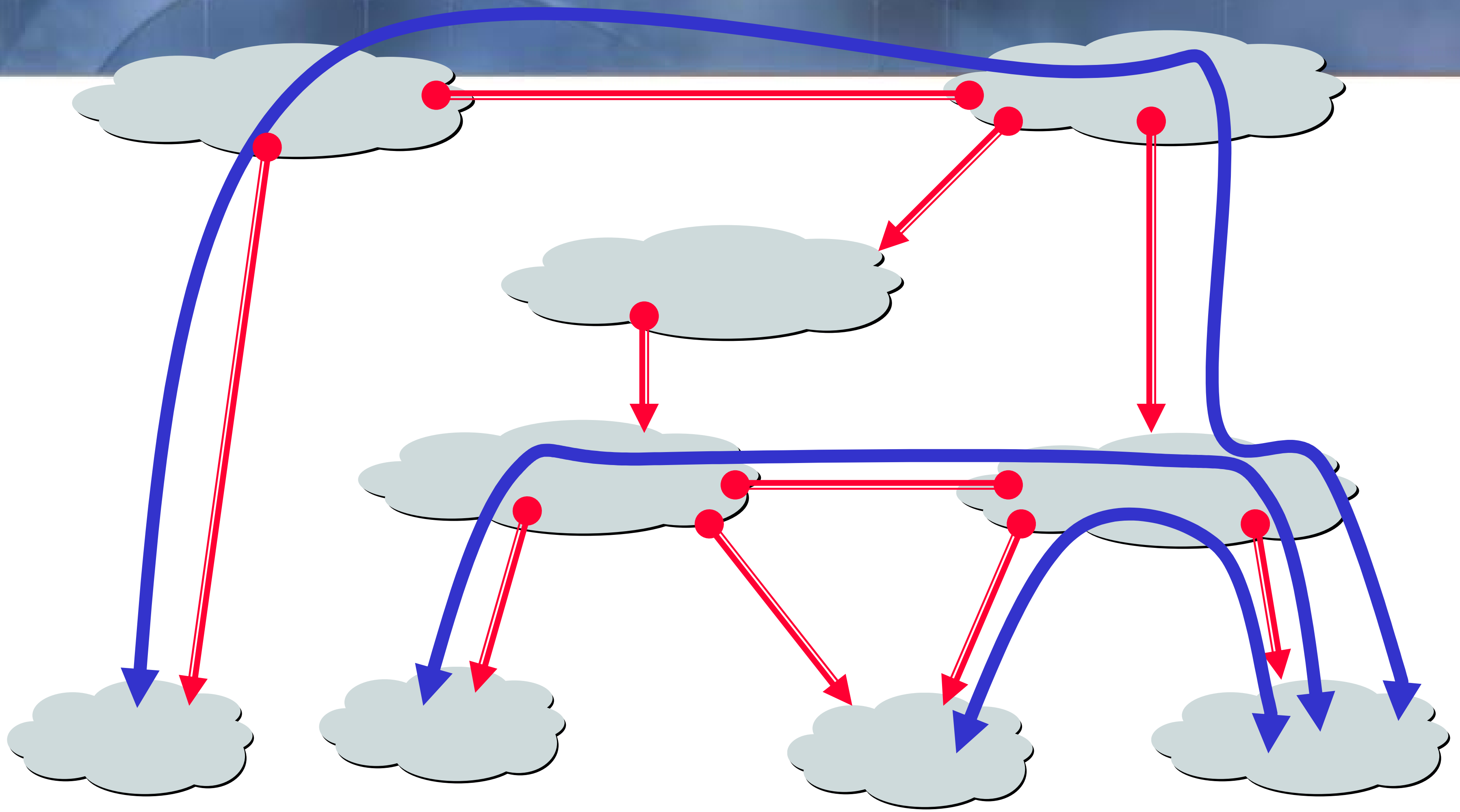
Peers provide transit between their respective customers

<u>Peers do not provide transit between peers</u>

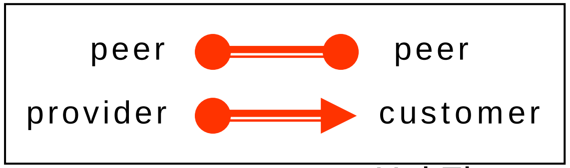
Peers (often) do not exchange $$$

Hui Zhang

Peering Provides Shortcuts



Peering also allows connectivity between the customers of “Tier 1” providers.



Peering Wars

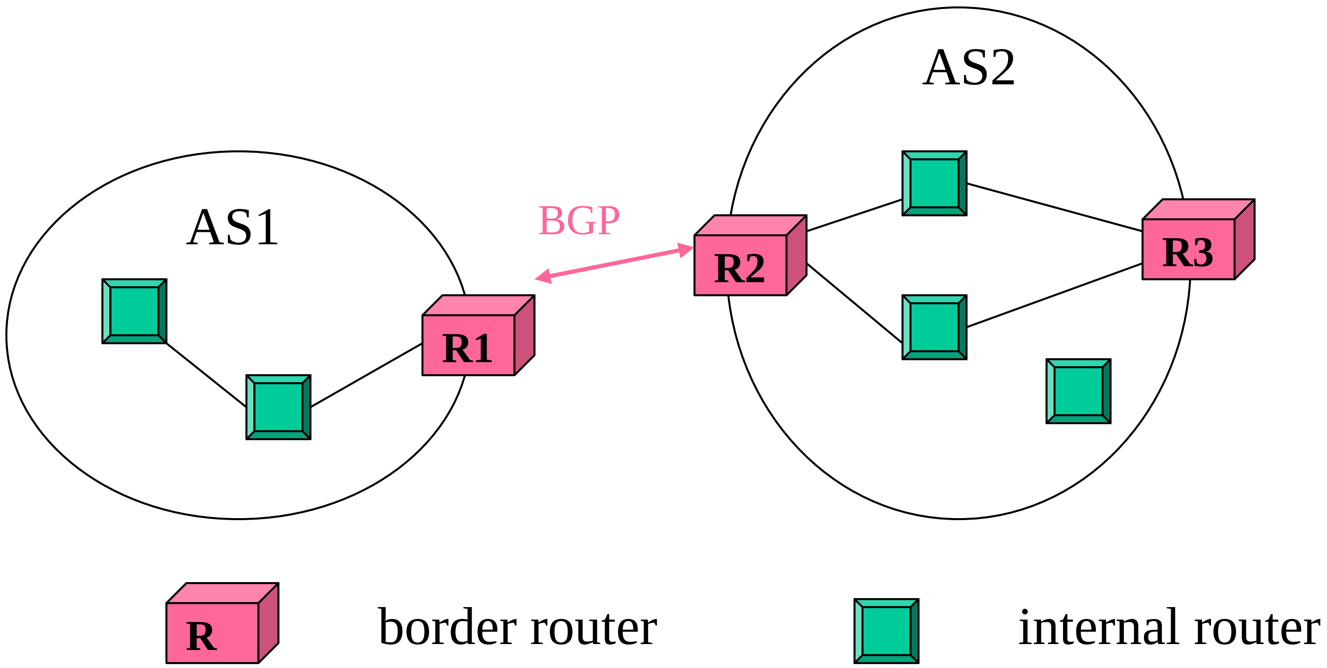
Peer	Don't Peer
Reduces upstream transit costs	You would rather have customers
Can increase end-to-end performance	Peers are usually your competition
May be the only way to connect your customers to some part of the Internet ("Tier 1")	Peering relationships may require periodic renegotiation

Peering struggles are by far the most contentious issues in the ISP world!

Peering agreements are often confidential.

Hui Zhang

Border Gateway Protocol



- Two types of routers
  - Border router, Internal router

Purpose of BGP

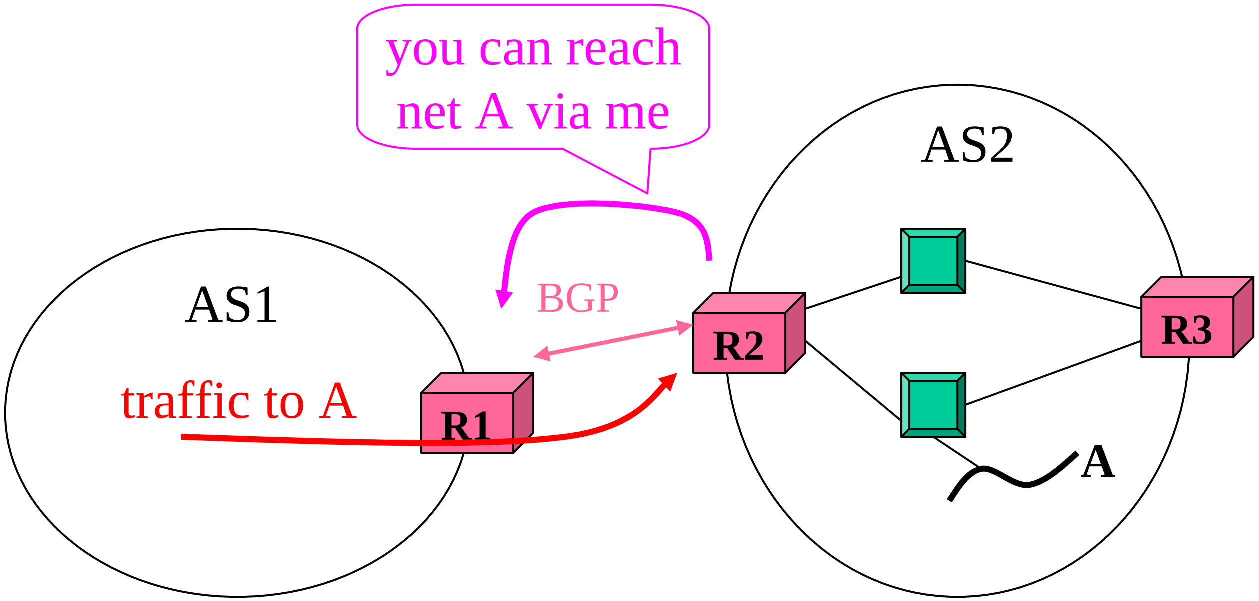
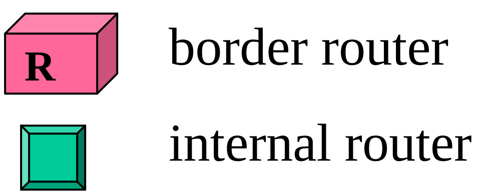


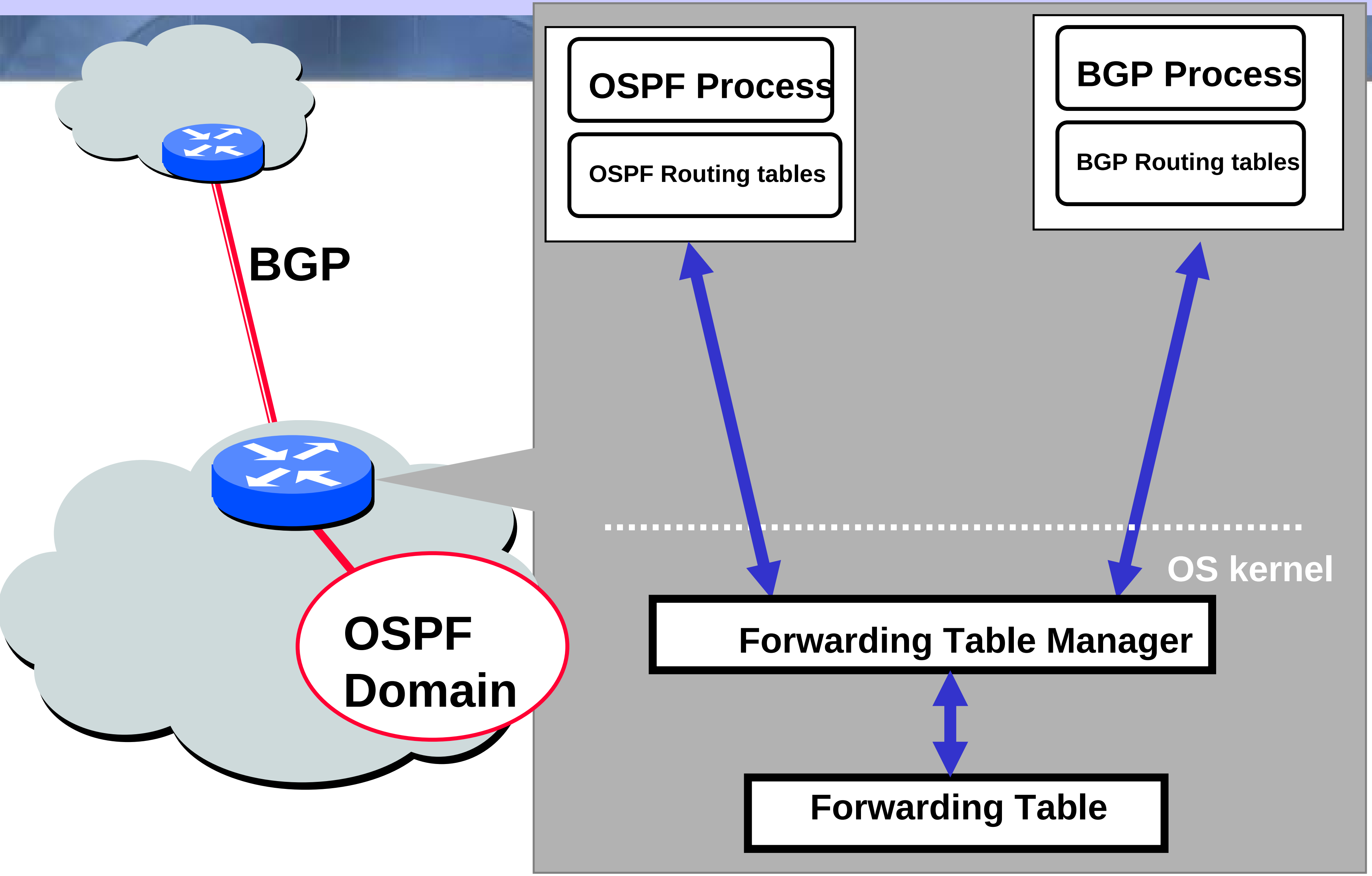
table at R1:

dest	next hop
A	R2

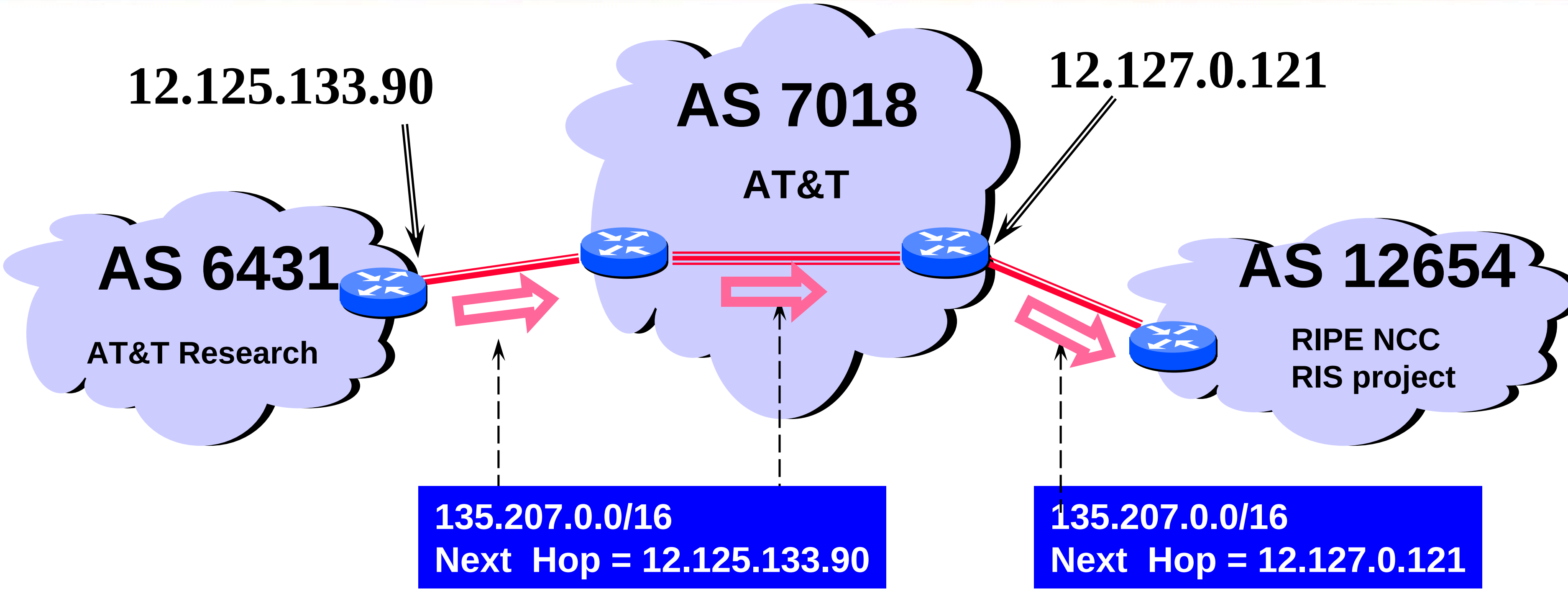


Share connectivity information across ASes

Multiple Routing Processes on a Single Router

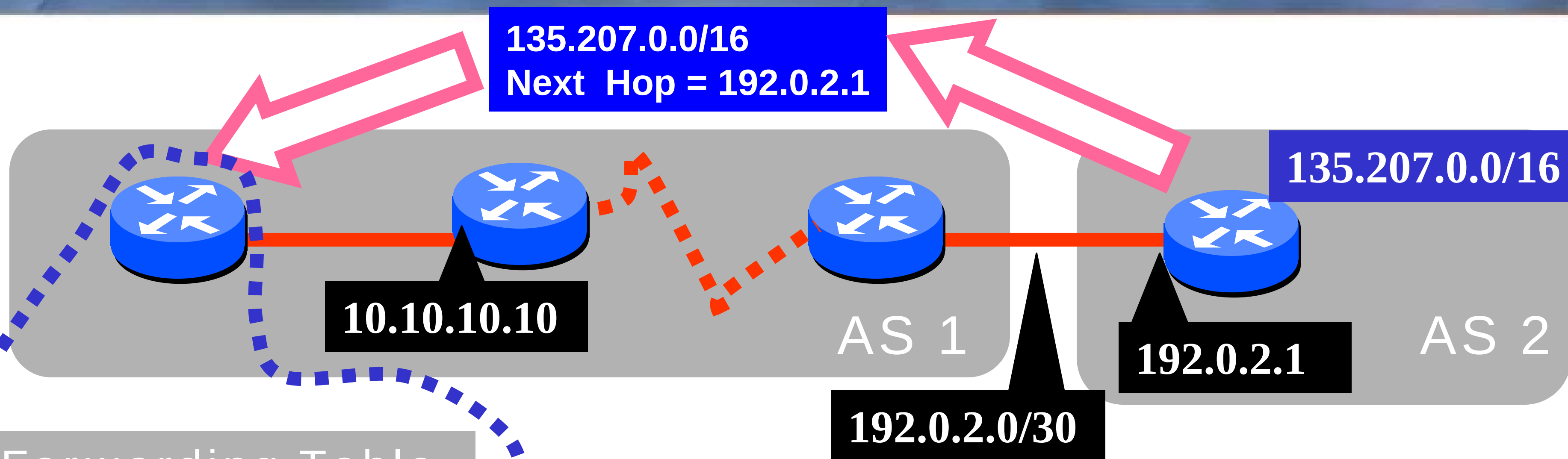


Example of Advertising Route



Every time a route announcement crosses an AS boundary, the Next Hop attribute is changed to the IP address of the border router that announced the route.

Join EGP with IGP For Connectivity



Forwarding Table

destination	next hop
192.0.2.0/30	10.10.10.10

+

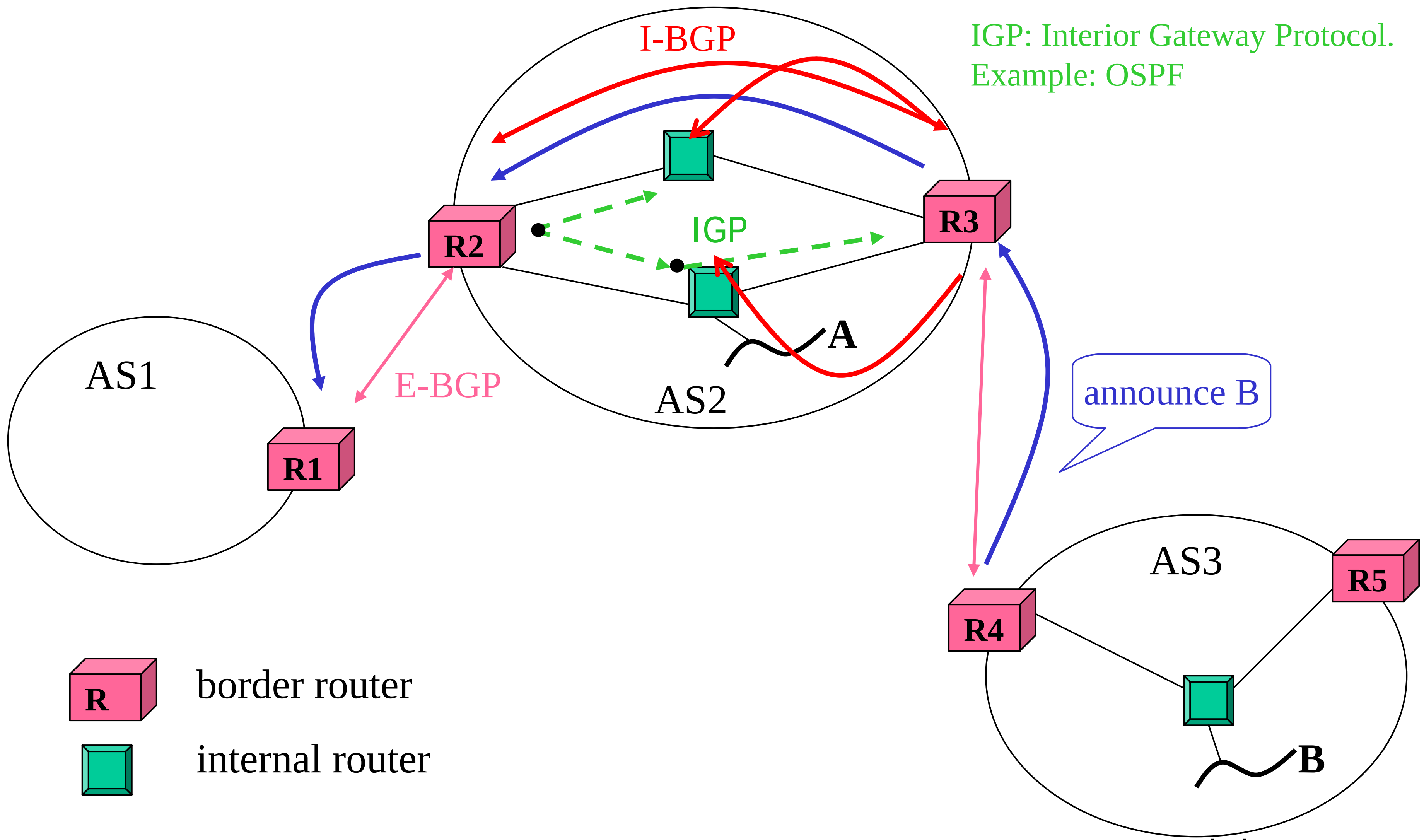
EGP

destination	next hop
135.207.0.0/16	192.0.2.1

Forwarding Table

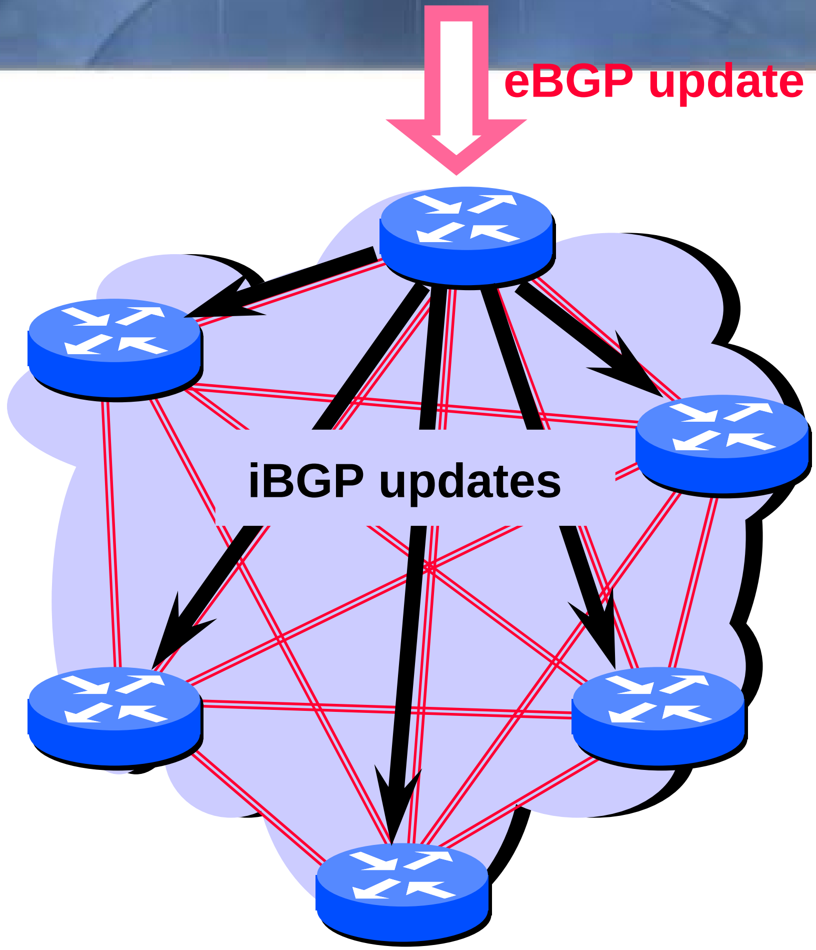
destination	next hop
135.207.0.0/16	10.10.10.10
192.0.2.0/30	10.10.10.10

I-BGP and E-BGP



Hui Zhang

iBGP Peers Must be Fully Meshed



iBGP neighbors do not announce routes received via iBGP to other iBGP neighbors.

- Injecting external routes into IGP does not scale and causes BGP policy information to be lost
- BGP does not provide "shortest path" routing
- Is iBGP an IGP? NO!

Path Vector Protocol

- **Distance vector algorithm with extra information**
  - For each route, store the complete path (ASs)
  - No extra computation, just extra storage

- **Advantages:**
  - can make policy choices based on set of ASs in path
  - can easily avoid loops

BGP Operations (Simplified)