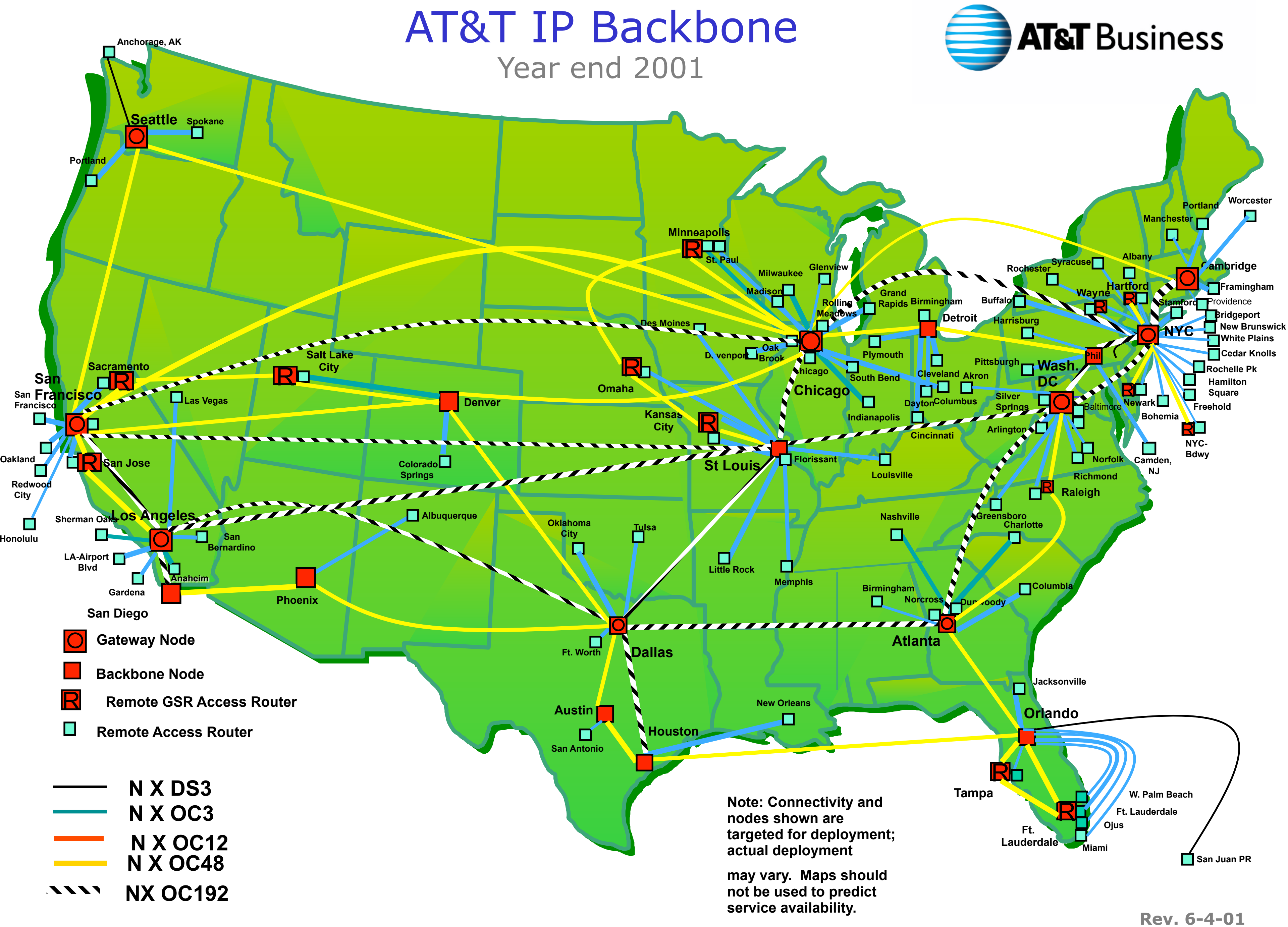
A more detailed look at BGP

PART I

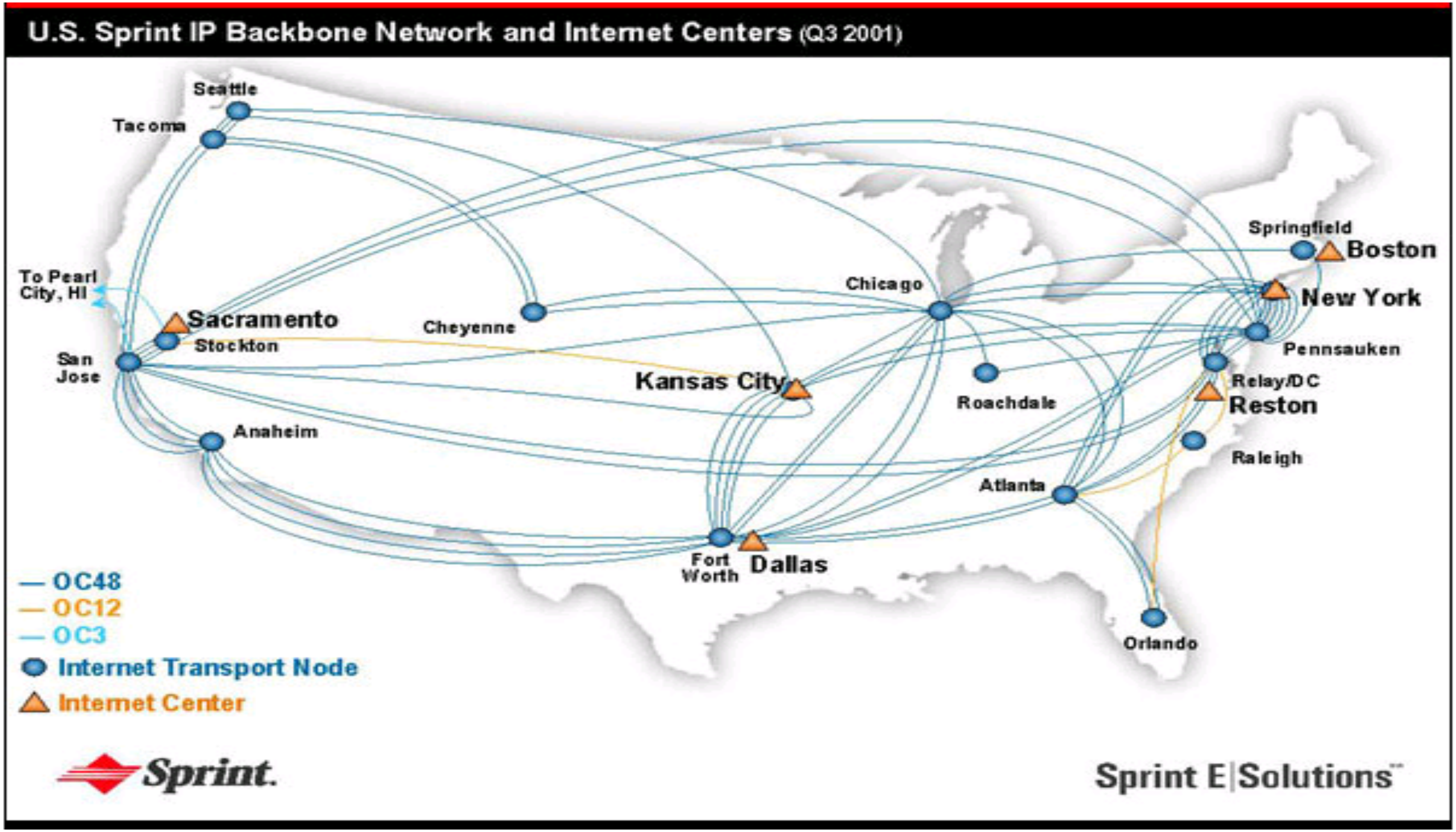
Physical Connectivity

AT&T IP Backbone

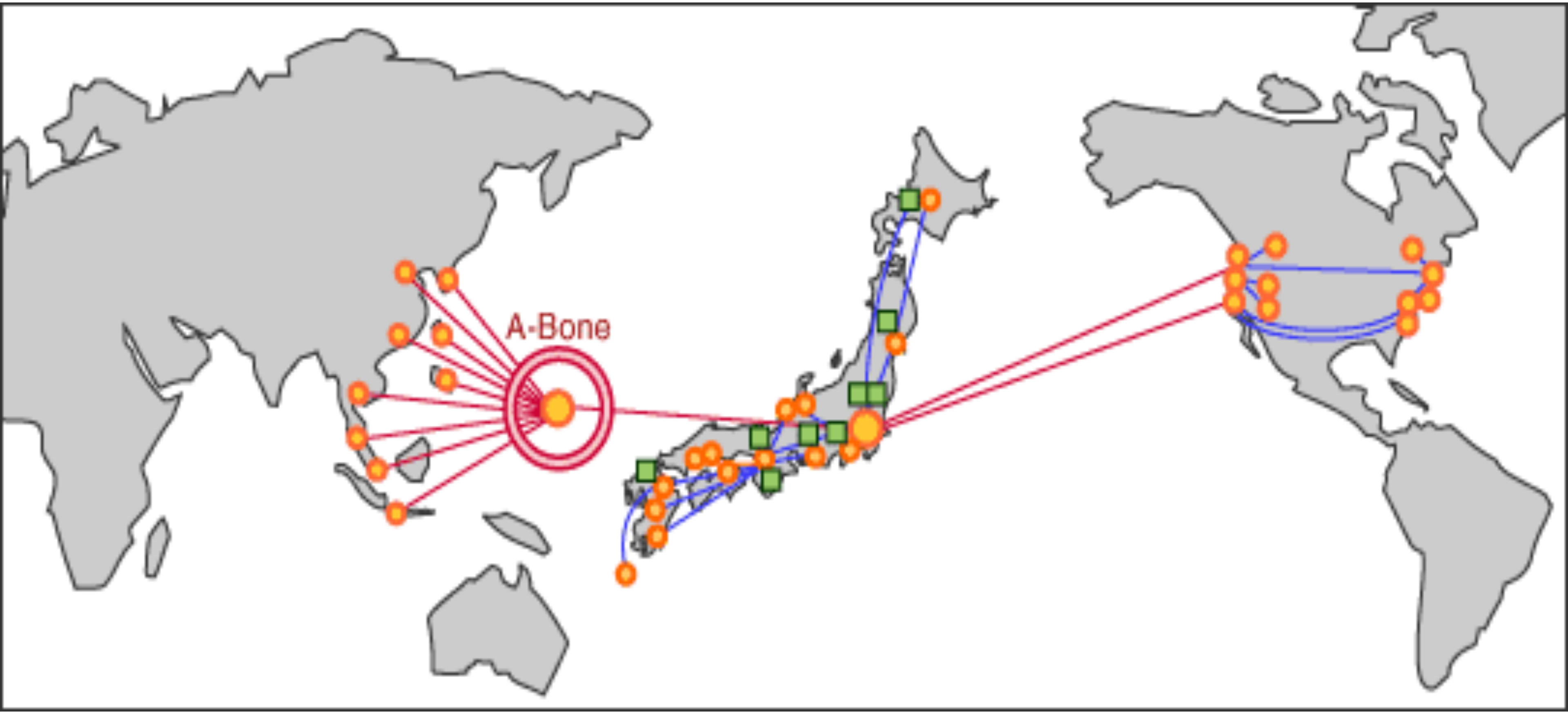
Year end 2001



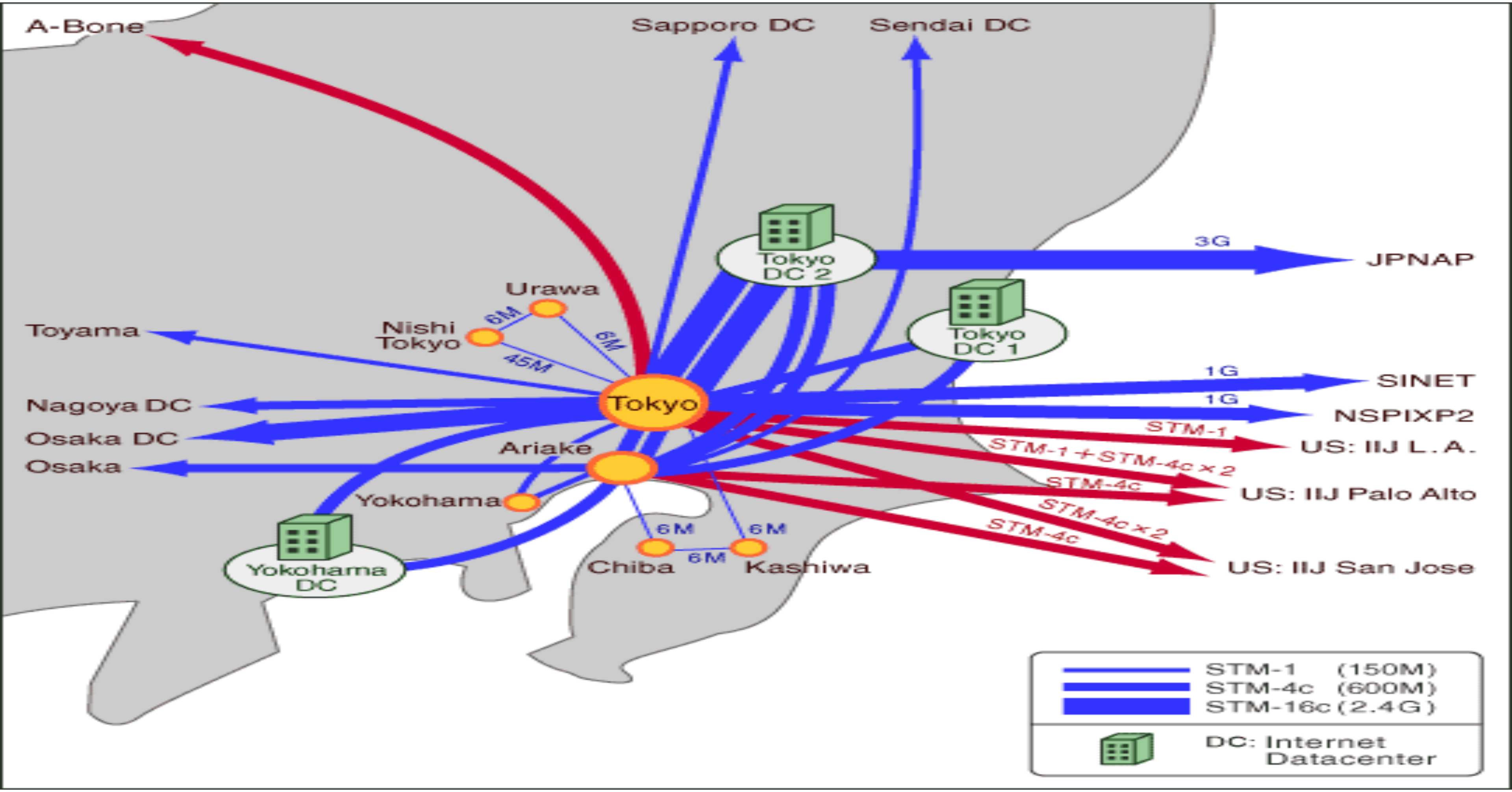
Sprint, USA


U.S. Sprint IP Backbone Network and Internet Centers (Q3 2001)
Seattle
Tacoma
To Pearl City, HI
Sacramento
Stockton
San Jose
Anaheim
Cheyenne
Kansas City
Chicago
Roachdale
Springfield
Boston
New York
Pennsauken
Relay/DC
Reston
Raleigh
Atlanta
Orlando
Fort Worth
Dallas
— OC48
— OC12
— OC3
Internet Transport Node
Internet Center
Sprint.
Sprint E|Solutions

Internet Initiative Japan (IIJ)

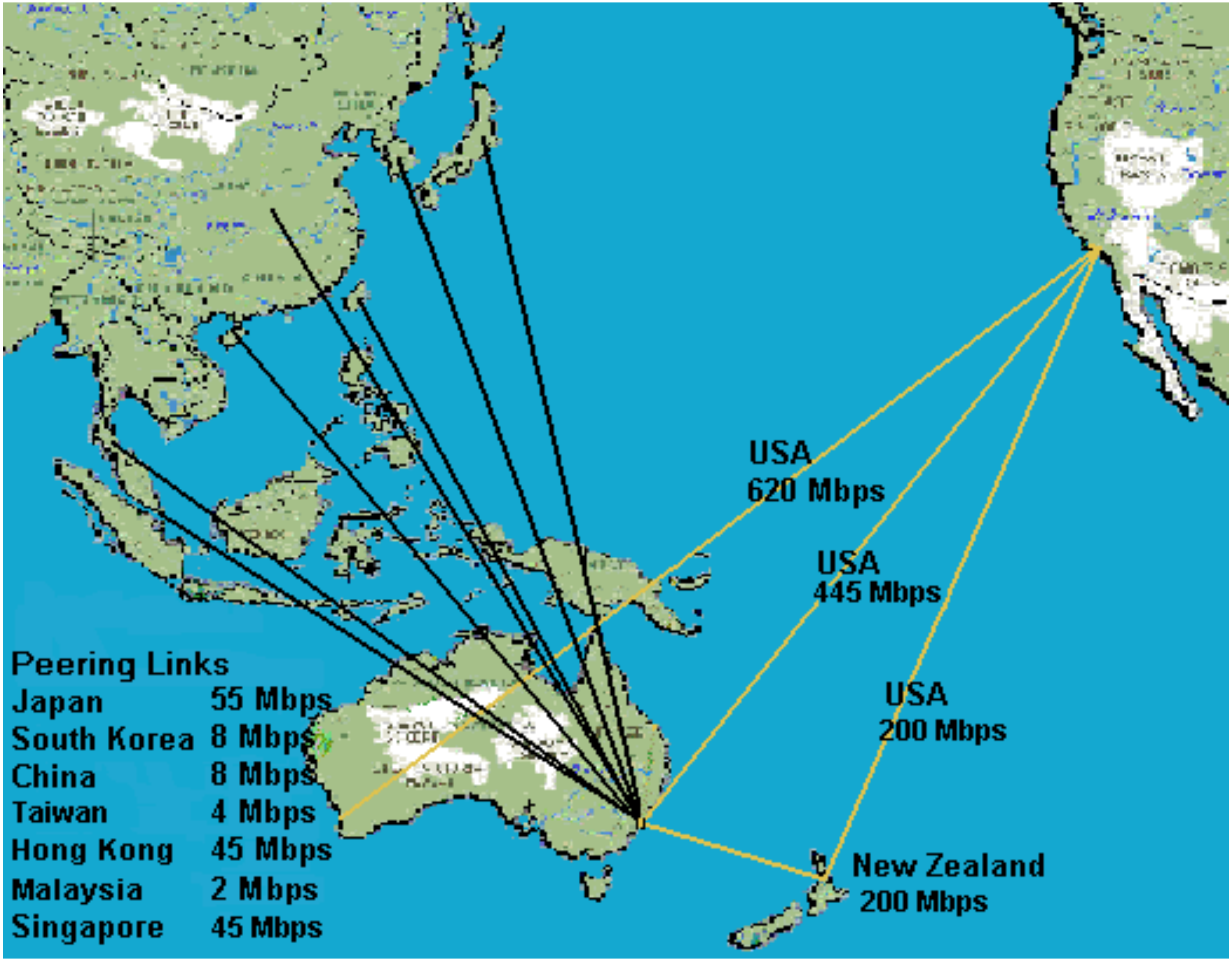


IIJ, Tokyo

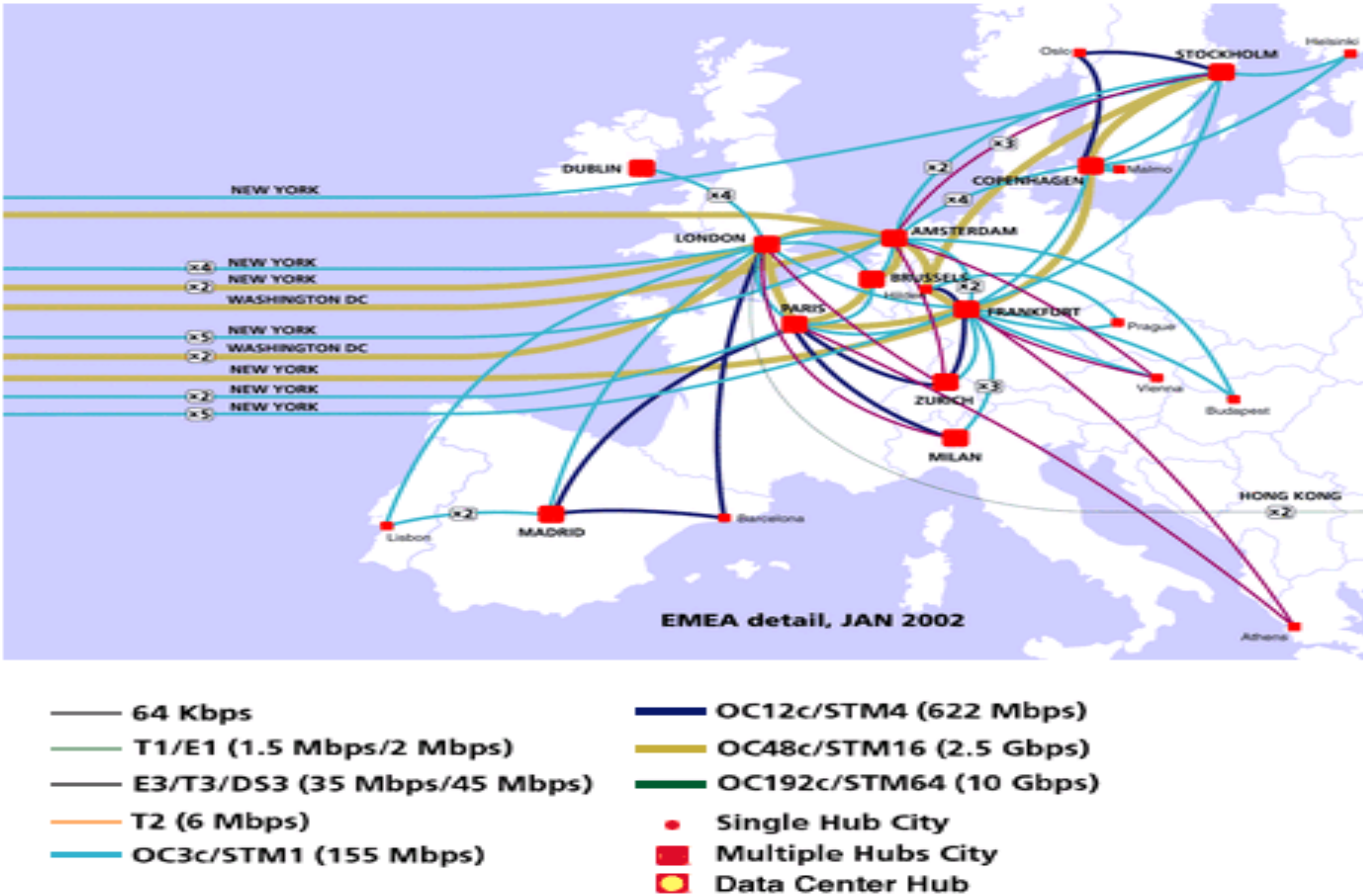


A-Bone
Sapporo DC
Sendai DC
Tokyo DC 2
3G
JPNAP
Tokyo DC 1
Urawa
6M
6M
Nishi Tokyo
45M
Toyama
Tokyo
1G
SINET
1G
NSPIXP2
Nagoya DC
Osaka DC
Osaka
Ariake
STM-1
US: IIJ L.A.
STM-1 + STM-4c x 2
STM-4c
US: IIJ Palo Alto
STM-4c x 2
STM-4c
US: IIJ San Jose
Yokohama
6M
6M
6M
Chiba
Kashiwa
Yokohama DC
STM-1 (150M)
STM-4c (600M)
STM-16c (2.4G)
DC: Internet Datacenter

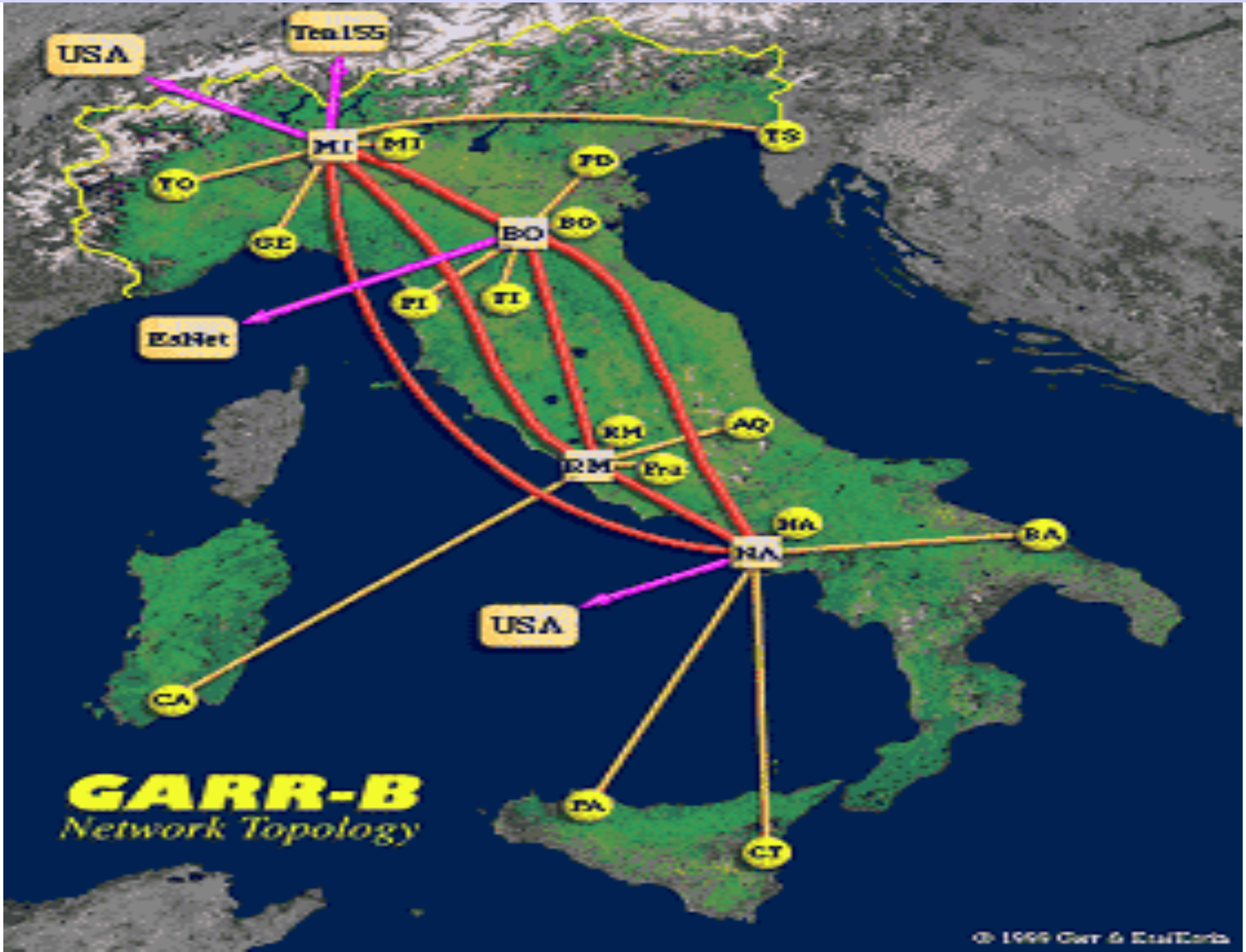
Telstra international



UUNet, Europe



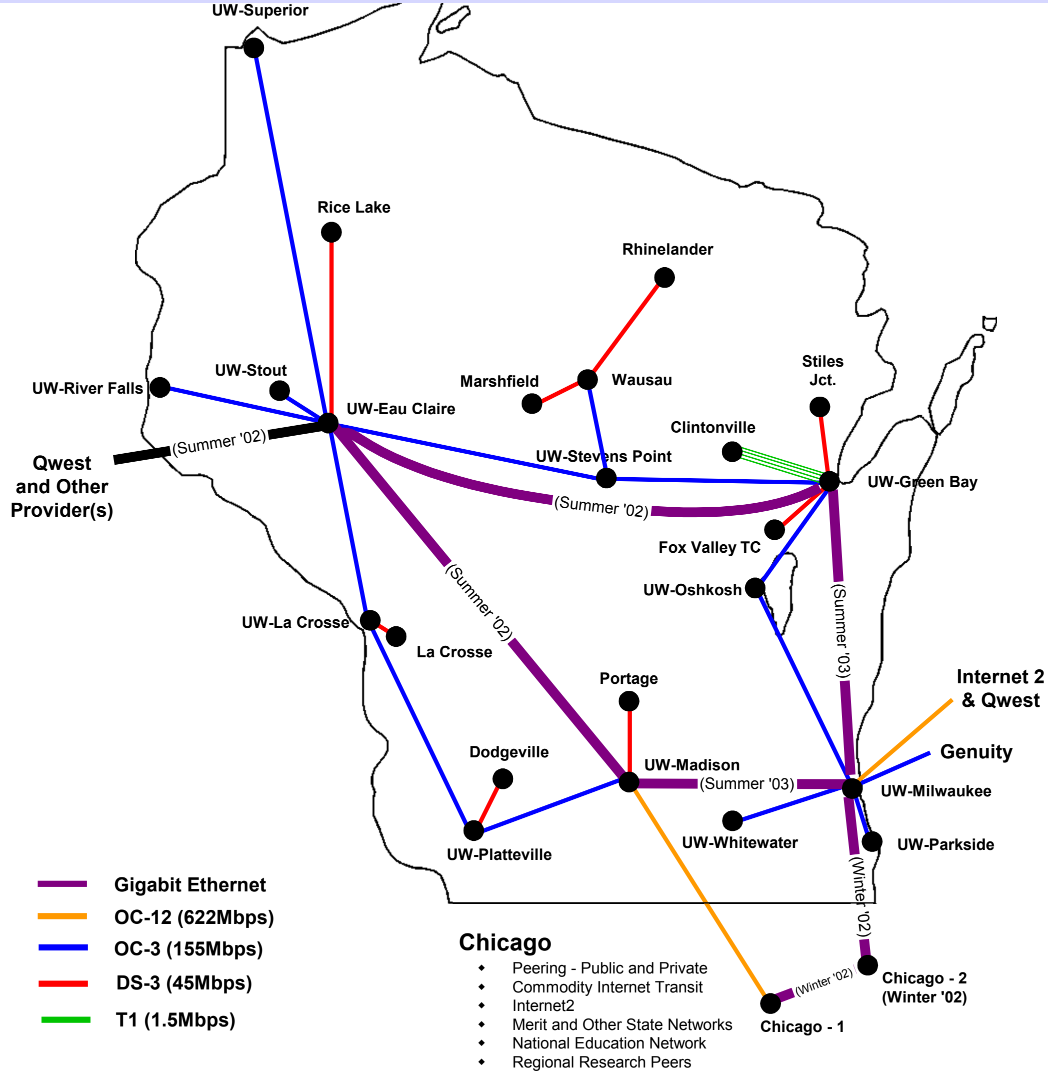
GARR-B



USA
Ten155
MI
MI
TS
PD
TO
BO
BO
GE
PI
FI
EsNet
RM
AQ
RM
Fra
NA
NA
BA
USA
CA
PA
CT
GARR-B
Network Topology
© 1999 Garr & EuroEarth

wiscnet.net

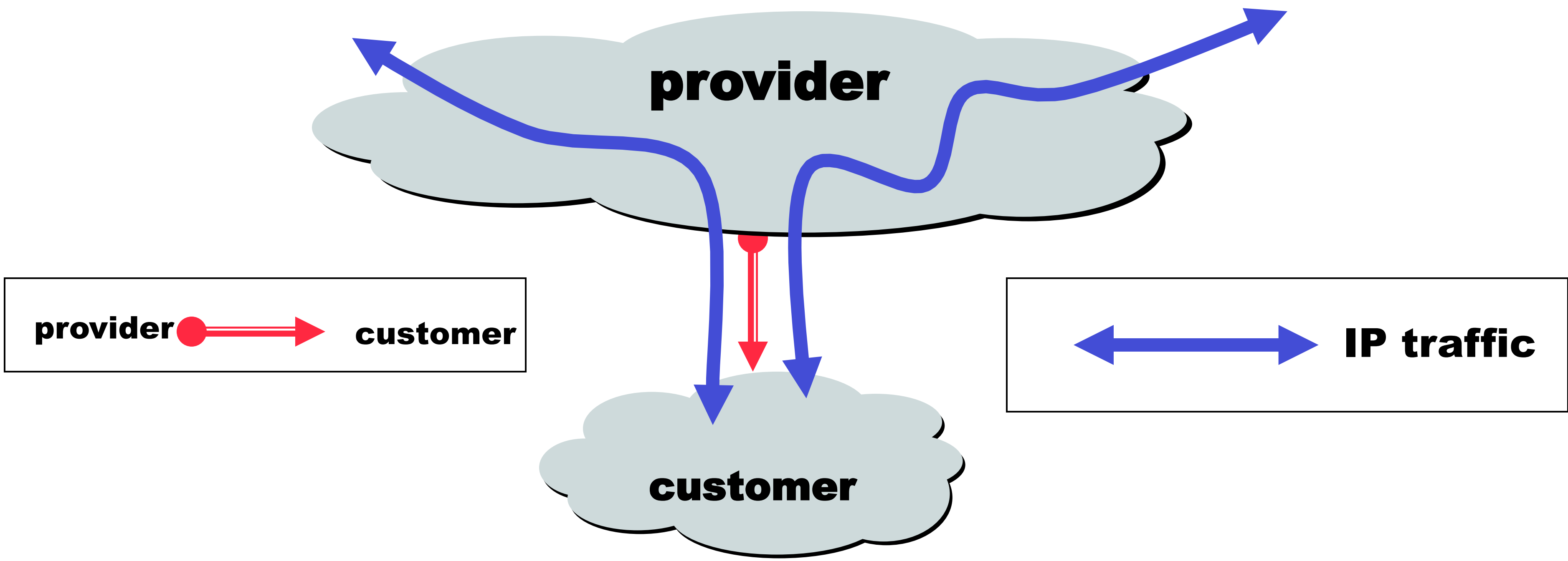
GO BUCKY!



PART II

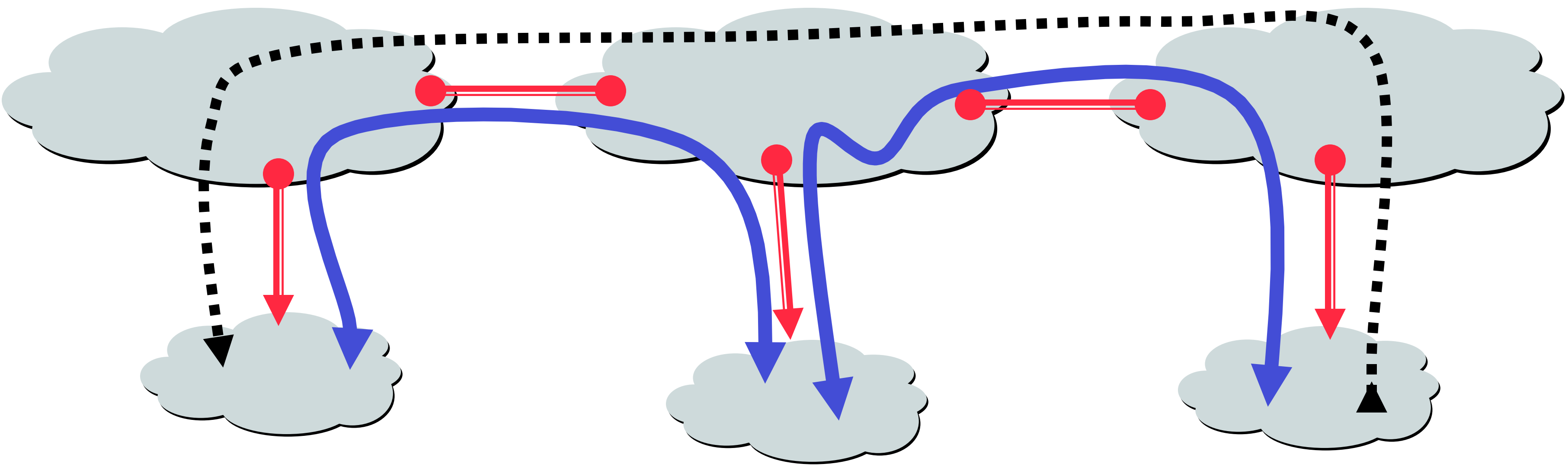
Relationships Between Networks

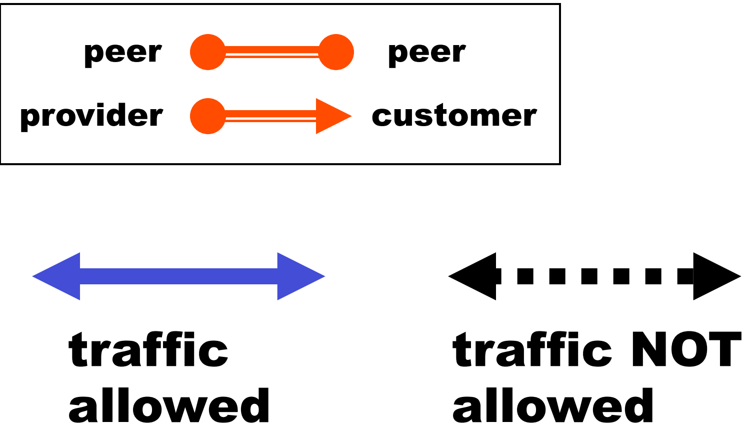
Customers and Providers



Customer pays provider for access to the Internet

The "Peering" Relationship





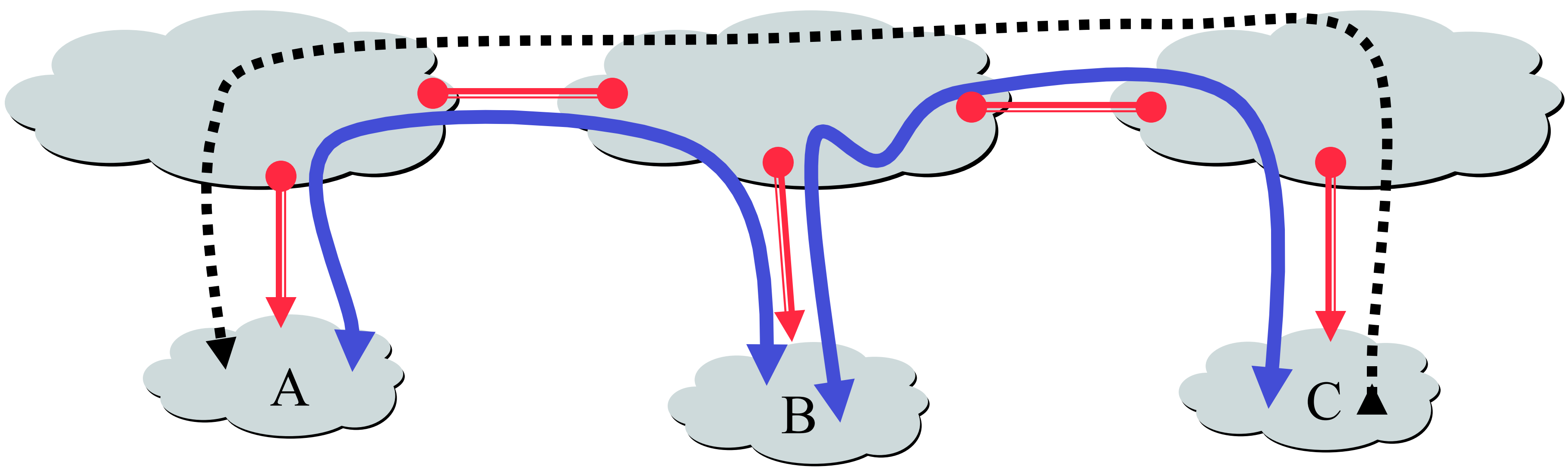
Peers provide transit between their respective customers

Peers do not provide transit between peers

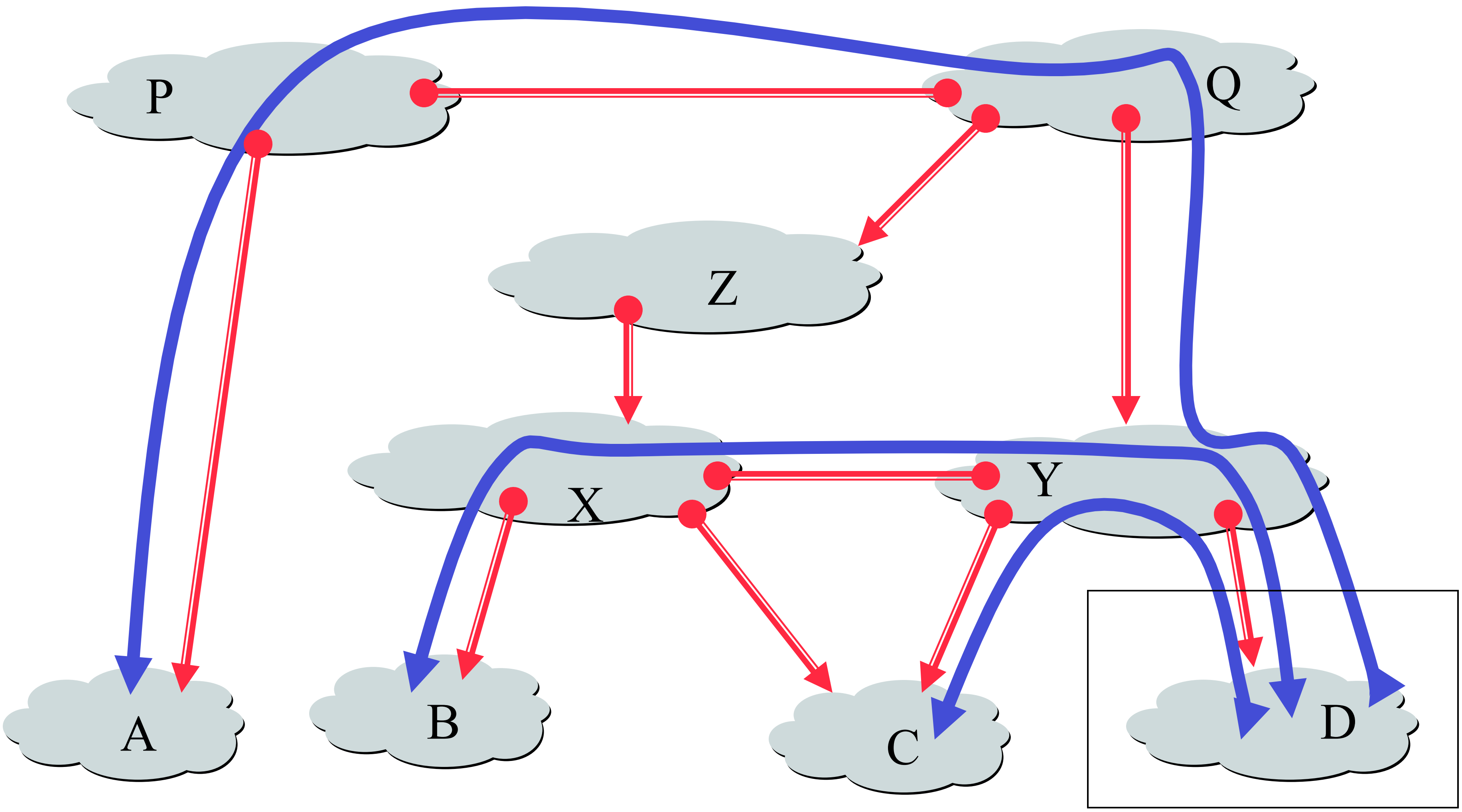
Peers (often) do not exchange $$$

Connectivity vs Reachibility

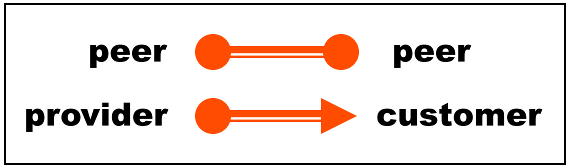
Connectivity does not imply reachability
(A and C may not be able to reach each other)



Peering Provides Shortcuts



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



Peering Wars

Peer

- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet ("Tier 1")

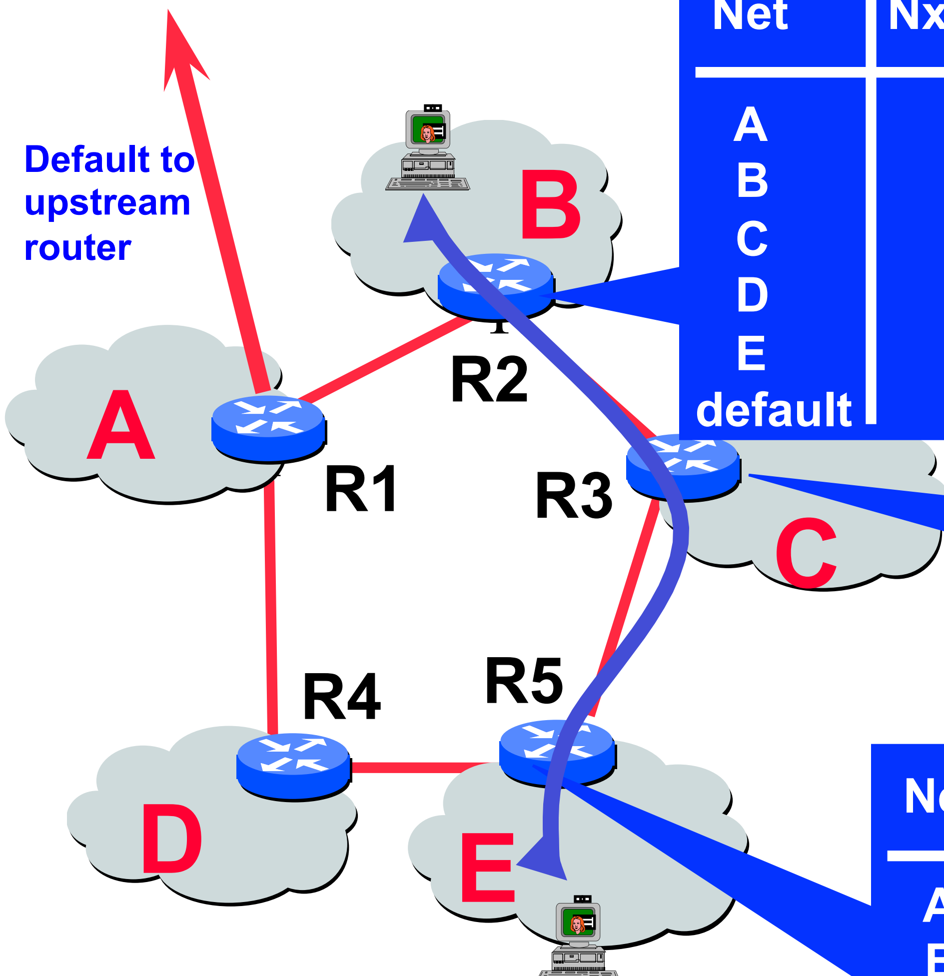
Don't Peer

- You would rather have customers
- Peers are usually your competition
- Peering relationships may require periodic renegotiation

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

Peering agreements are often confidential.

Routing vs. Forwarding



Forwarding always works

Routing can be badly broken

R2 table:

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

R3 table:

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

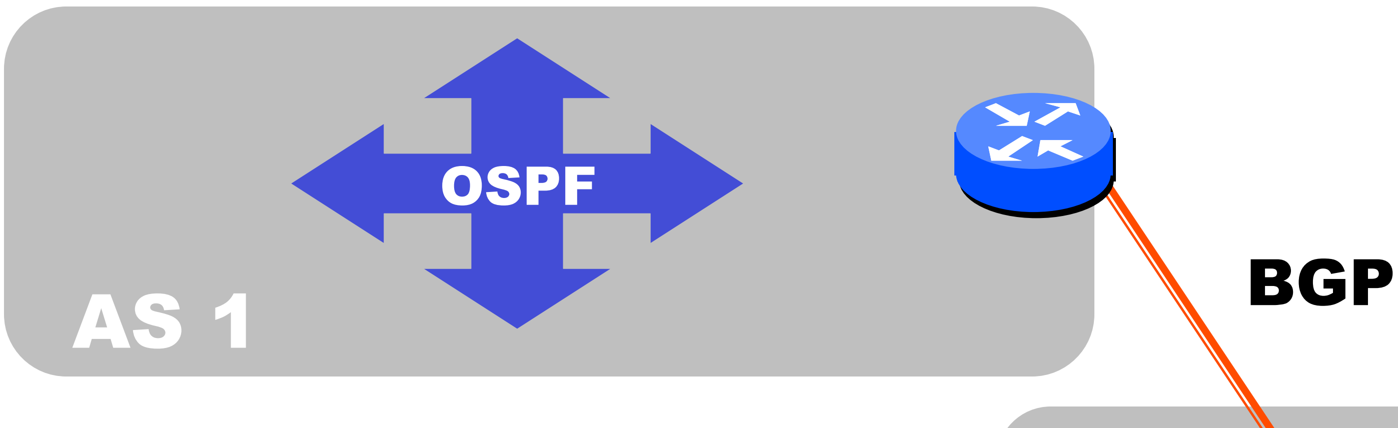
R5 table:

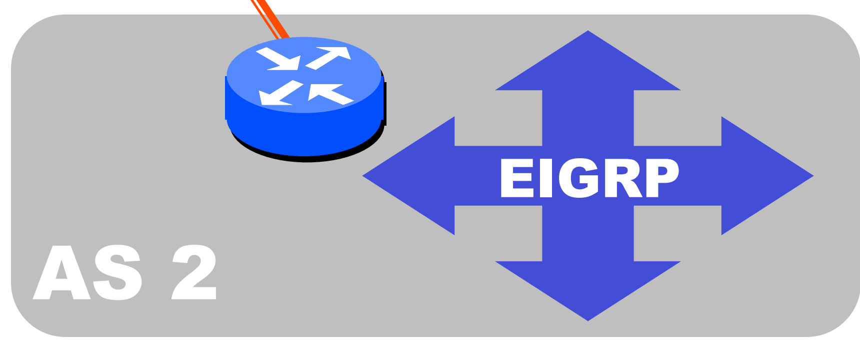
Net	Nxt Hop
A	R4
B	R3
C	R3
D	R4
E	Direct
default	R4

Forwarding: determine next hop

Routing: establish end-to-end paths

Inter-domain and Intra-domain routing





IGP = Interior Gateway Protocol

Metric based: OSPF, IS-IS, RIP, EIGRP (cisco)

EGP = Exterior Gateway Protocol

Policy based: BGP

The Routing Domain of BGP is the entire Internet

Technology of Distributed Routing

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

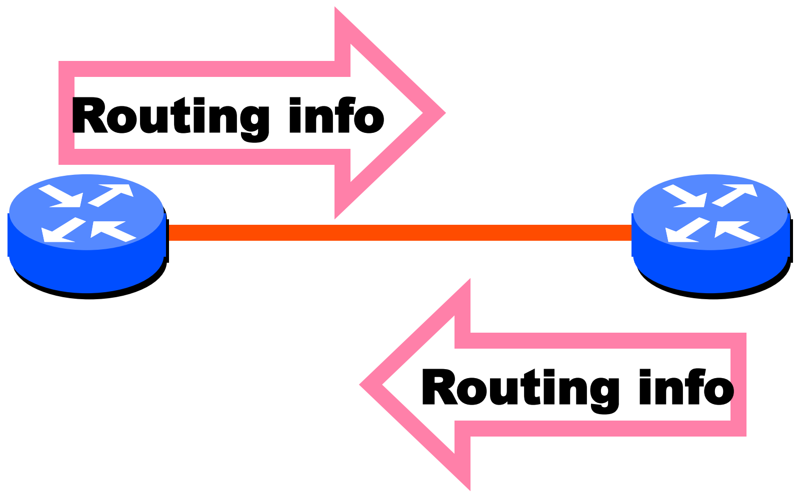
Distance Vector

- 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

The Gang of Four

	Link State	Vectoring
IGP	OSPF IS-IS	RIP
EGP		BGP

Routers Talking to Routers



- Routing computation is distributed among routers within a routing domain
- Computation of best next hop based on routing information is the most CPU/memory intensive task on a router
- Routing messages are usually not routed, but exchanged via layer 2 between physically adjacent routers (internal BGP and multi-hop external BGP are exceptions)

Autonomous Routing Domains (ARDs)

A collection of physical networks glued together using IP, that have a unified administrative routing policy.

- **Campus networks**
- **Corporate networks**
- **ISP Internal networks**
- **...**

Autonomous Systems (ASes)

An autonomous system is an autonomous routing domain that has been assigned an Autonomous System Number (ASN).

> … the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.
>
> RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System

AS Numbers (ASNs)

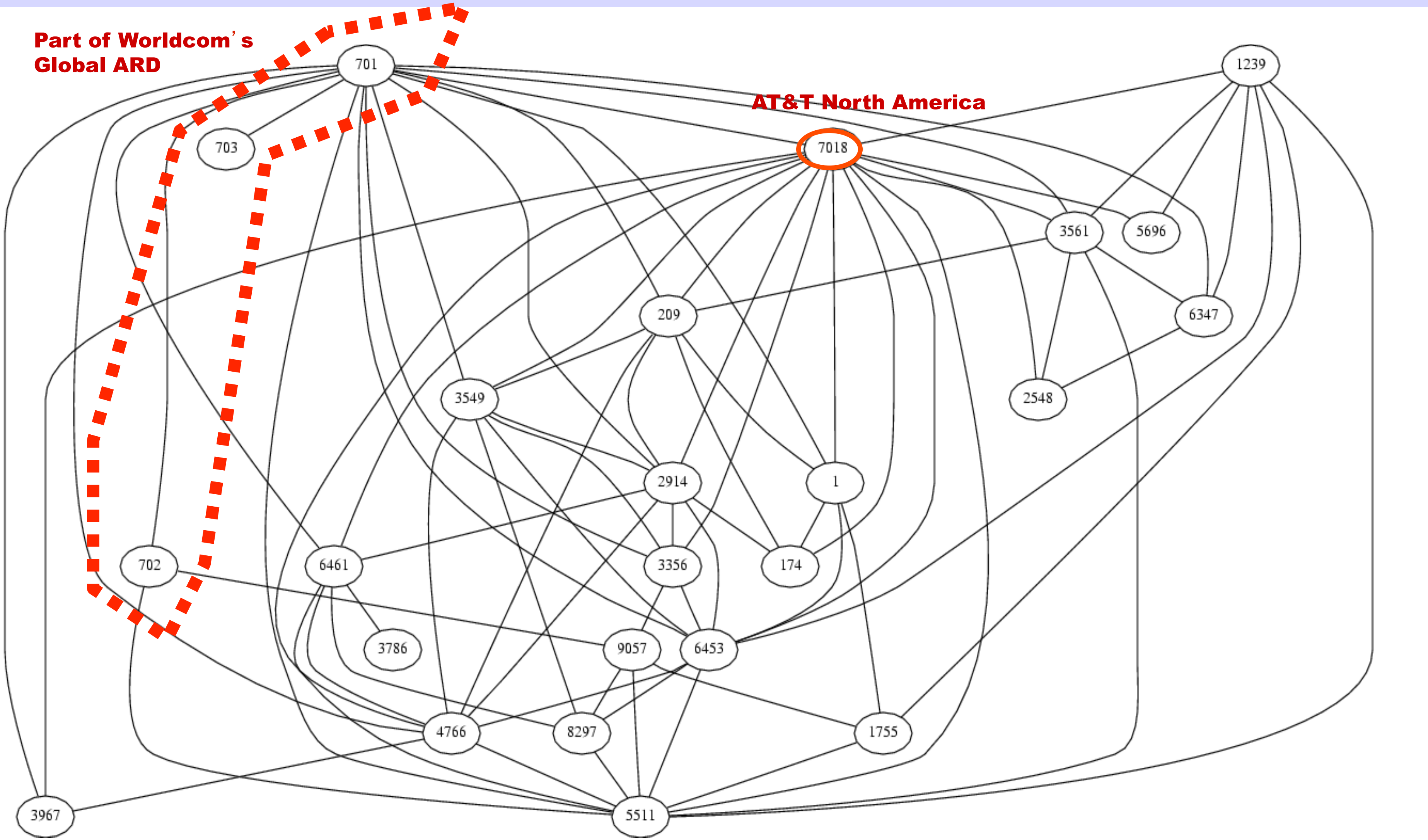
ASNs are 16 bit values.
64512 through 65535 are "private"

Currently over 11,000 in use.

- **Genuity (f.k.a. BBN): 1**
- **MIT: 3**
- **Harvard: 11**
- **UC San Diego: 7377**
- **AT&T: 7018, 6341, 5074, …**
- **UUNET: 701, 702, 284, 12199, …**
- **Sprint: 1239, 1240, 6211, 6242, …**
- **…**

ASNs represent units of routing policy

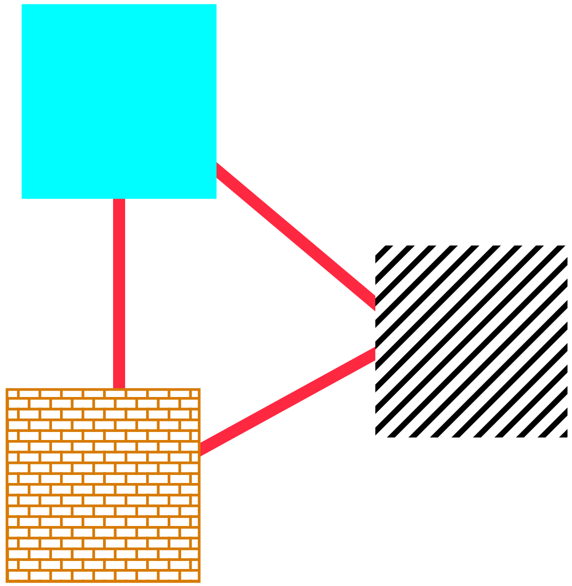
AS Graphs Can Be Fun



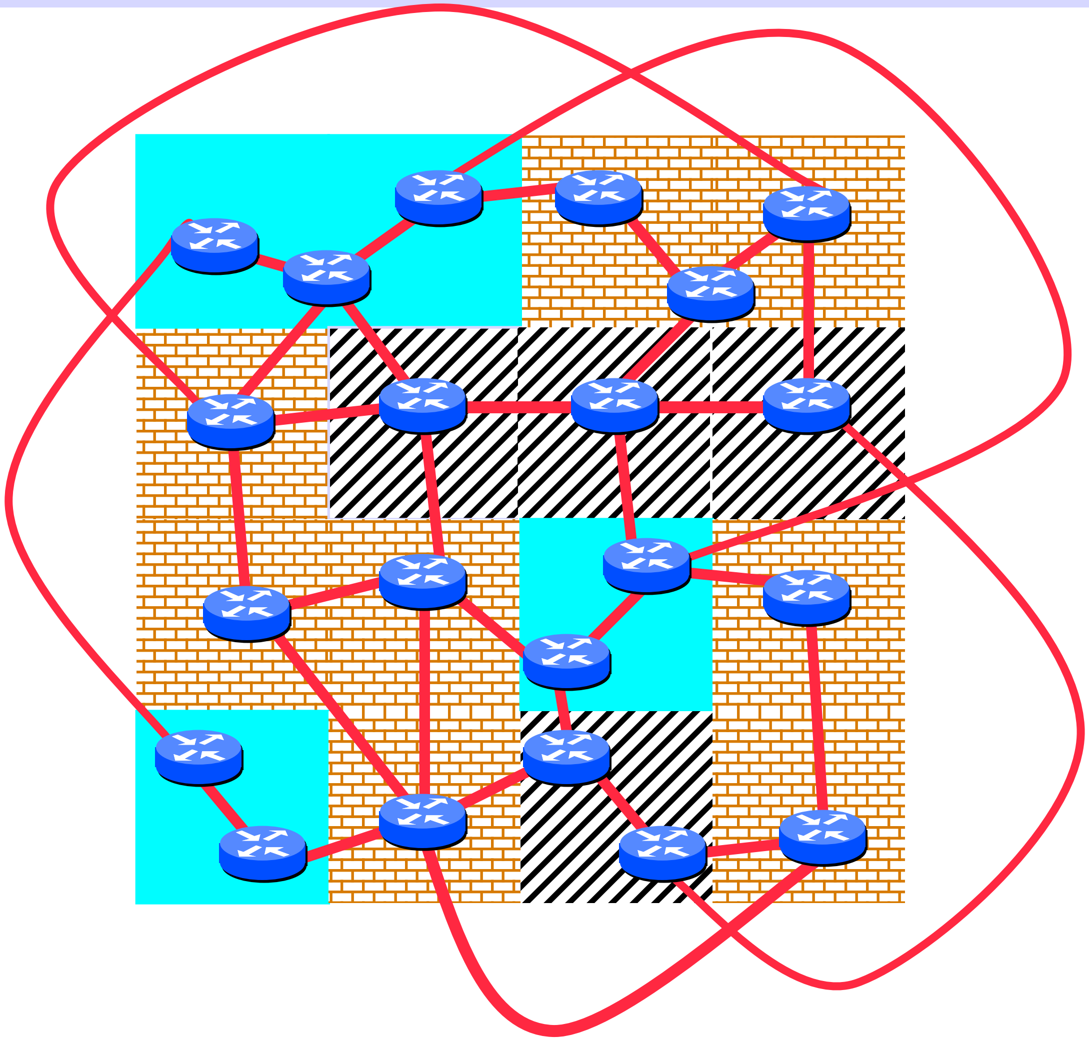
The <u>subgraph</u> showing all ASes that have more than 100 neighbors in full graph of 11,158 nodes. July 6, 2001. Point of view: AT&T route-server

AS Graph != Internet Topology

BGP was designed to throw away information!

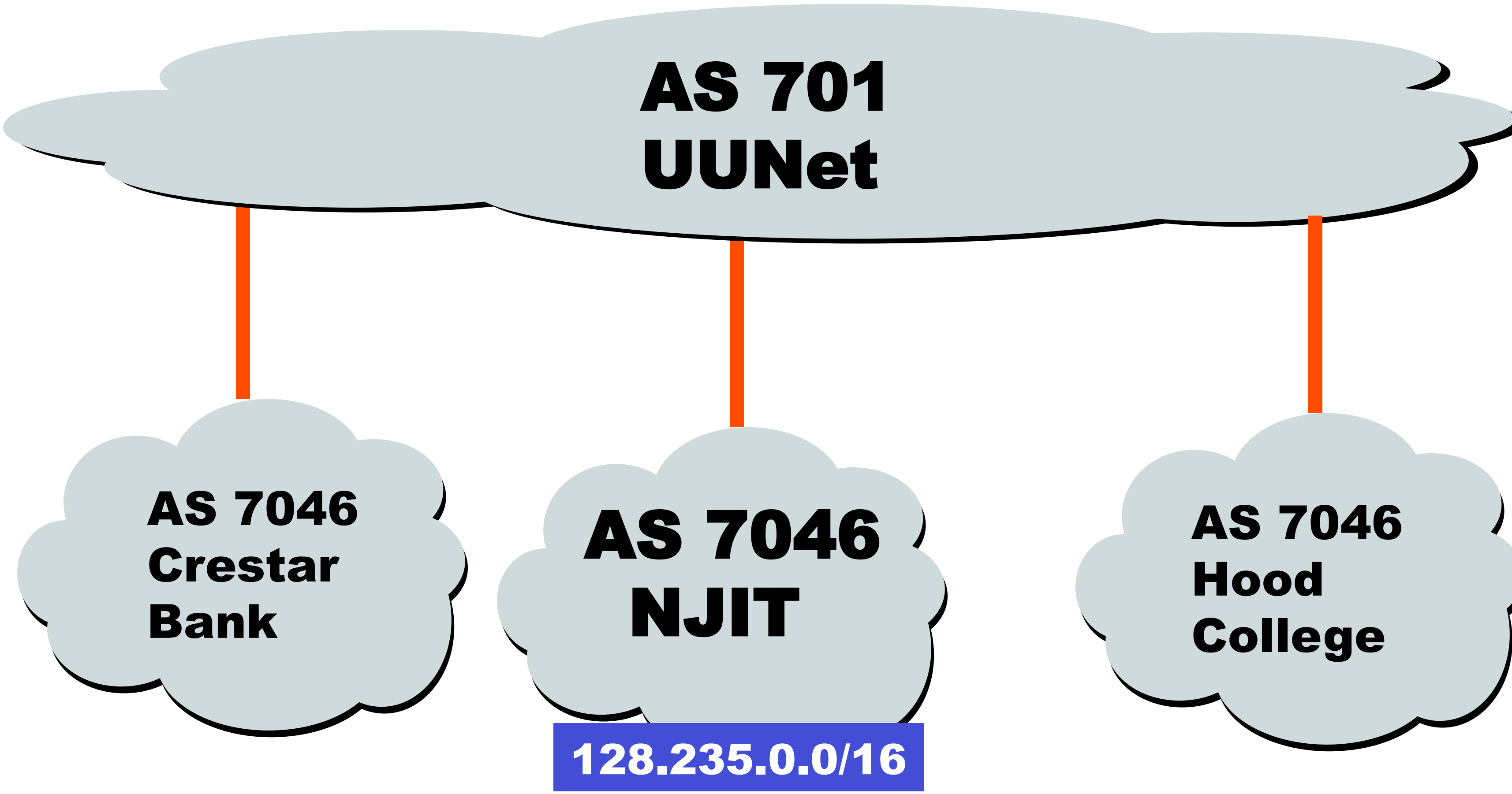


The AS graph may look like this.



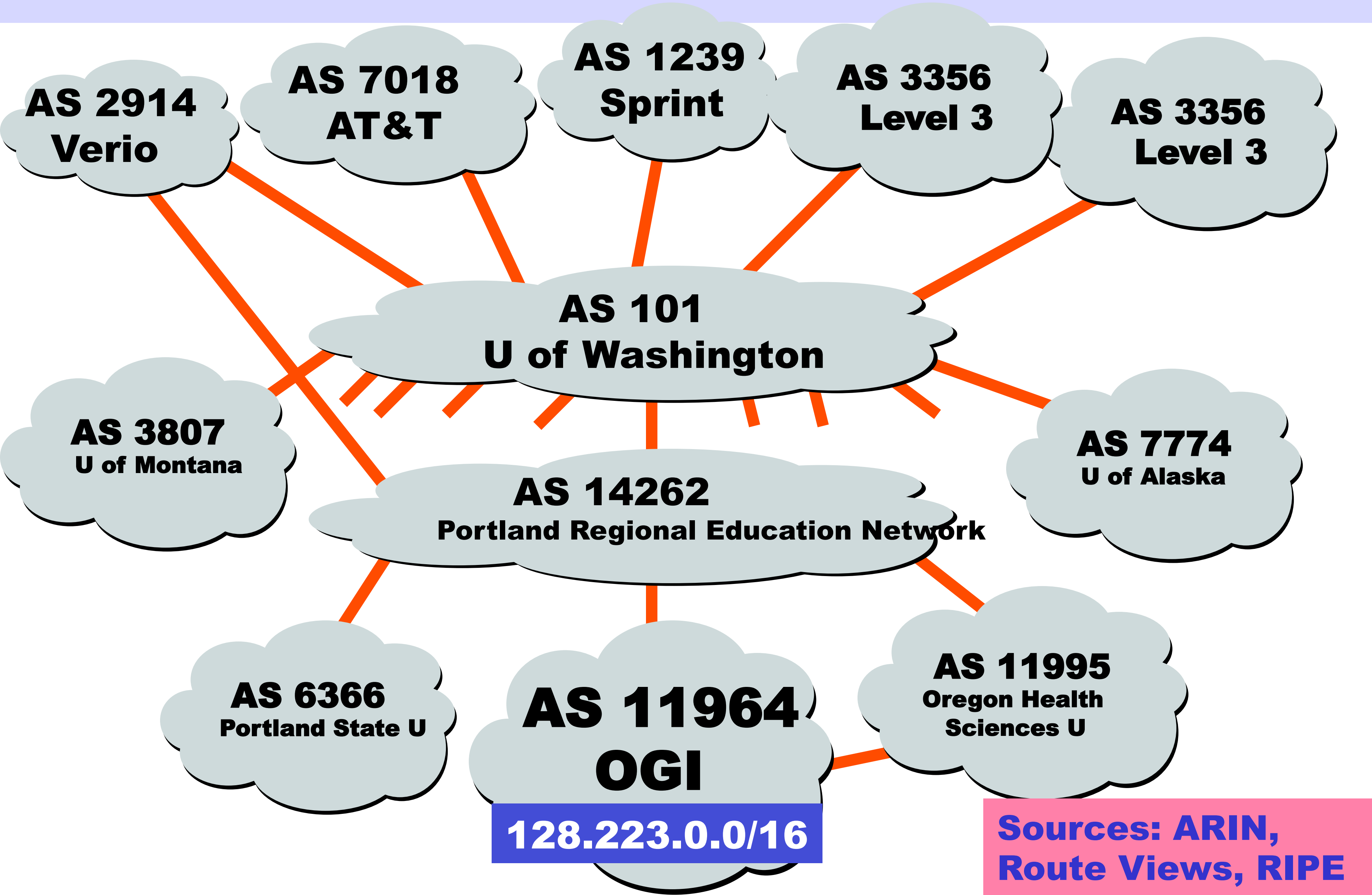
Reality may be closer to this…

ASNs Can Be "Shared" (RFC 2270)



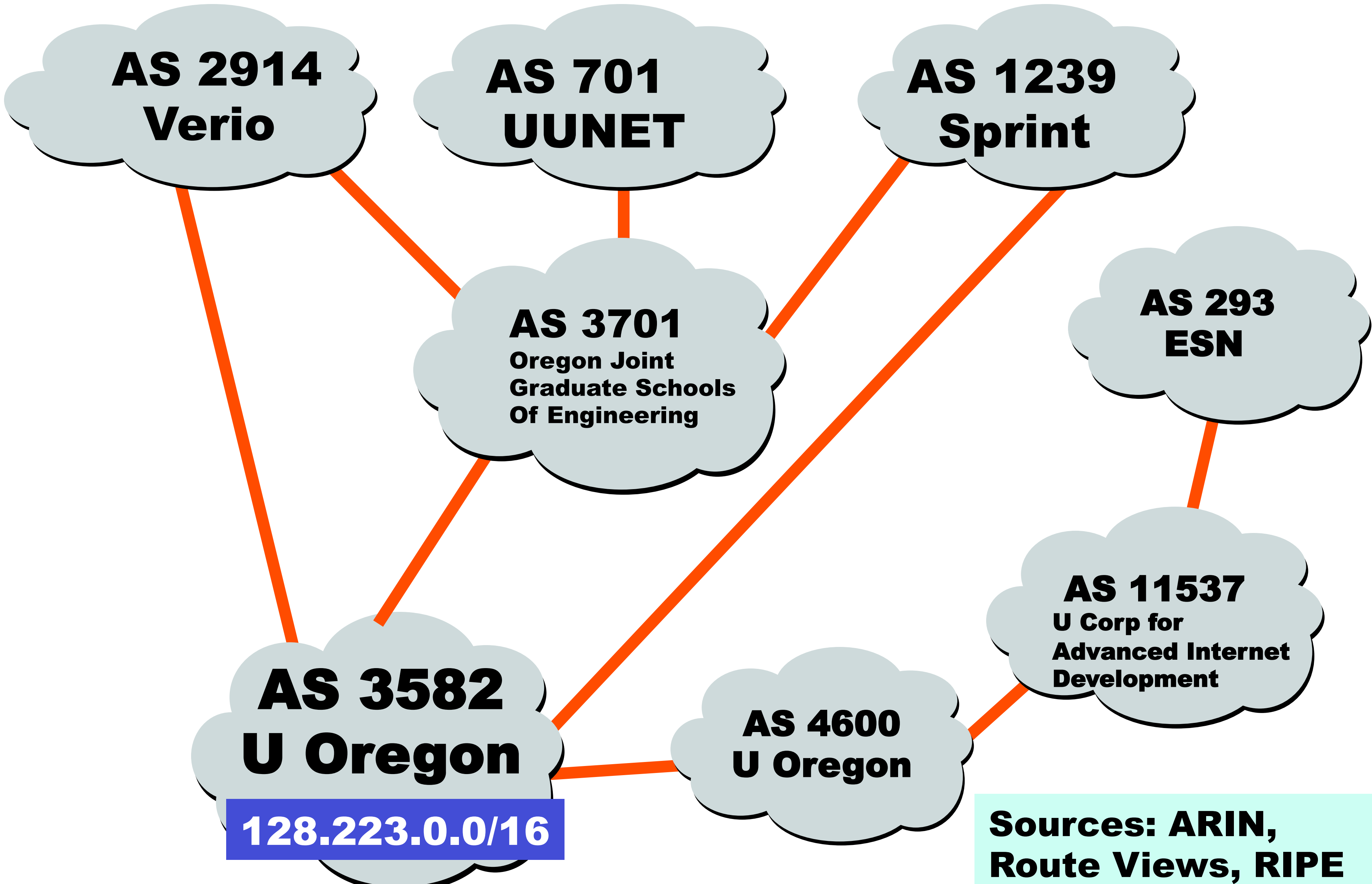
ASN 7046 is assigned to UUNet. It is used by Customers single homed to UUNet, but needing BGP for some reason (load balancing, etc..) [RFC 2270]

A Bit of OGI's AS Neighborhood



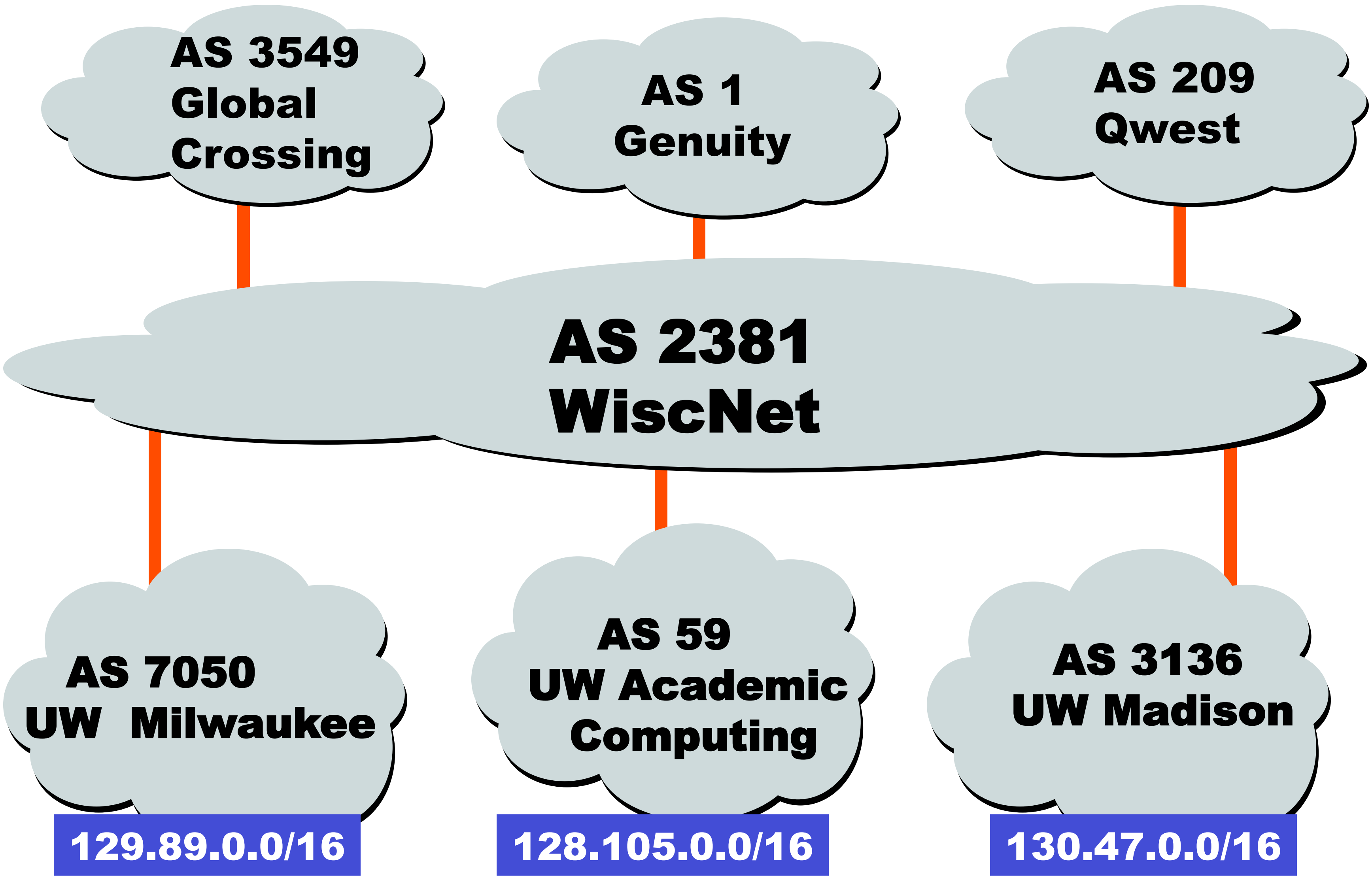
Sources: ARIN, Route Views, RIPE

A Bit of U Oregon's AS Neighborhood



Sources: ARIN, Route Views, RIPE

Partial View of cs.wisc.edu Neighborhood



ARD != AS

- Most ARDs have no ASN (statically routed at Internet edge)
- Some unrelated ARDs share the same ASN (RFC 2270)
- Some ARDs are implemented with multiple ASNs (example: Worldcom)

ASes are an implementation detail of Interdomain routing

PART III

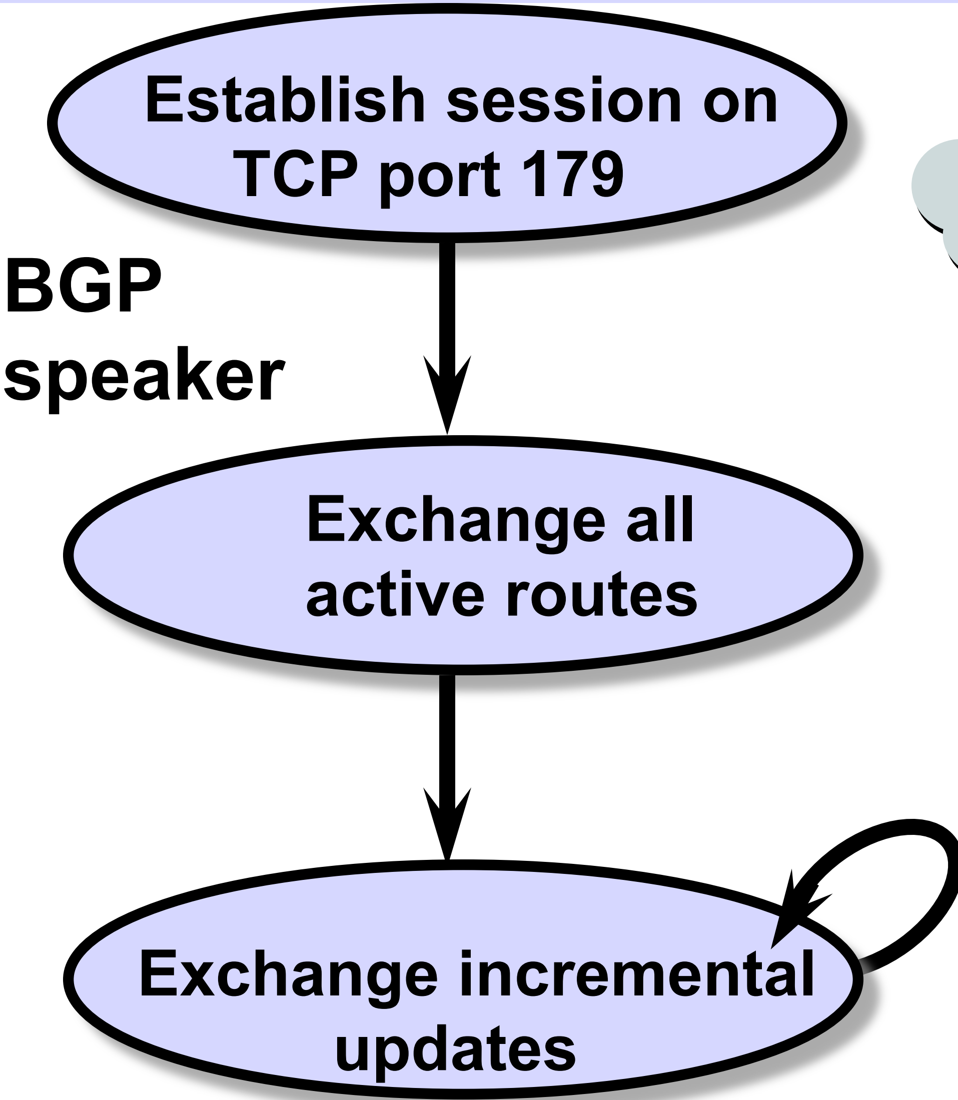
Implementing Inter-Network Relationships with BGP

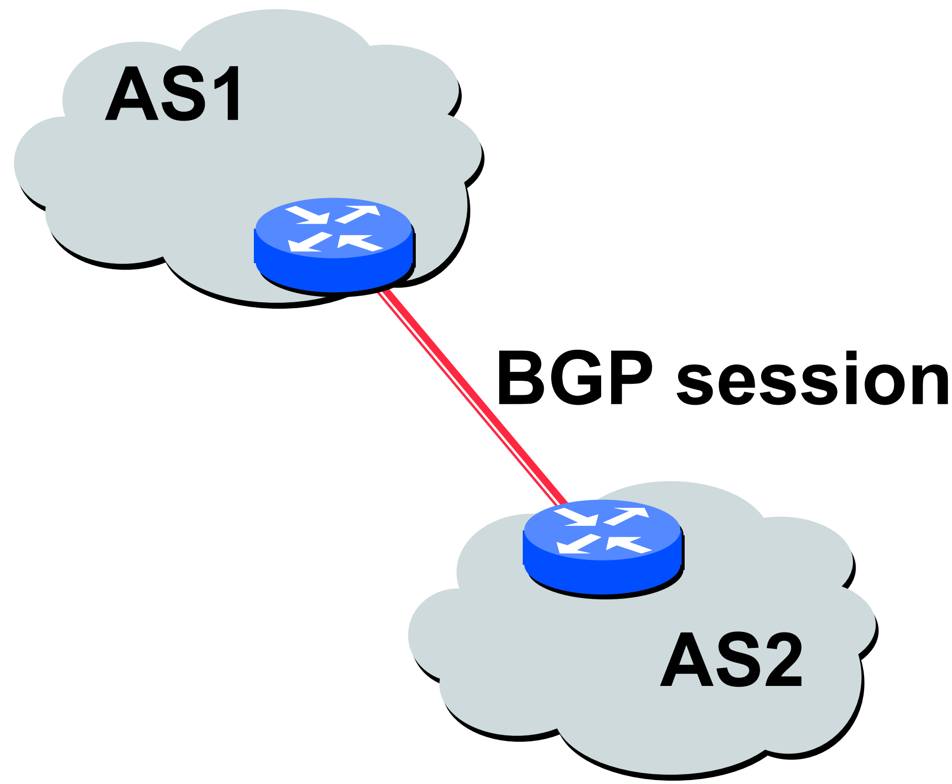
BGP-4

- **BGP** = **B**order **G**ateway **P**rotocol
- Is a **Policy-Based** routing protocol
- Is the **de facto EGP** of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.

- **1989 : BGP-1 [RFC 1105]**
  - **Replacement for EGP (1984, RFC 904)**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771]**
  - **Support for Classless Interdomain Routing (CIDR)**

BGP Operations (Simplified)



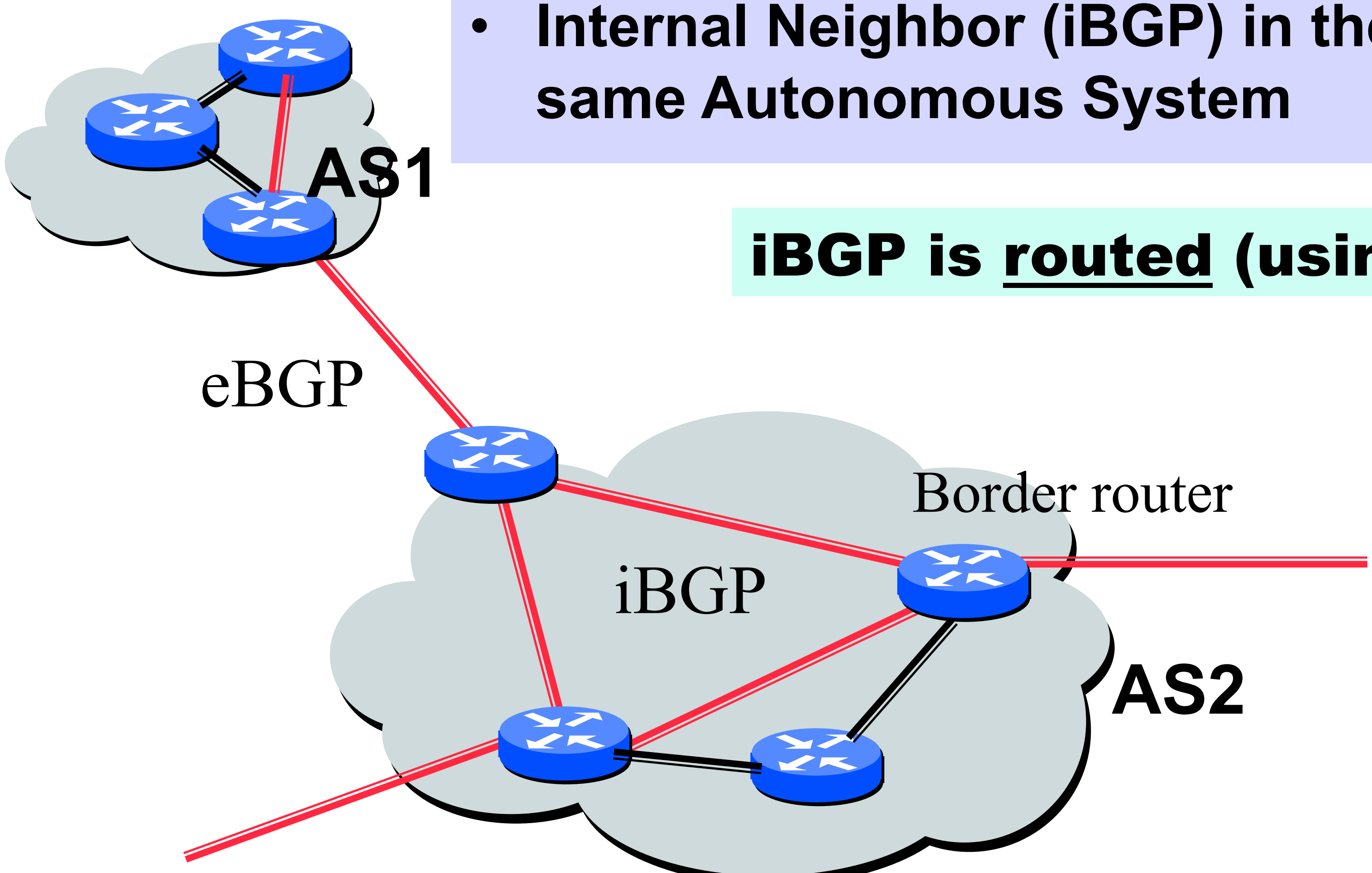


While connection is ALIVE exchange route UPDATE messages

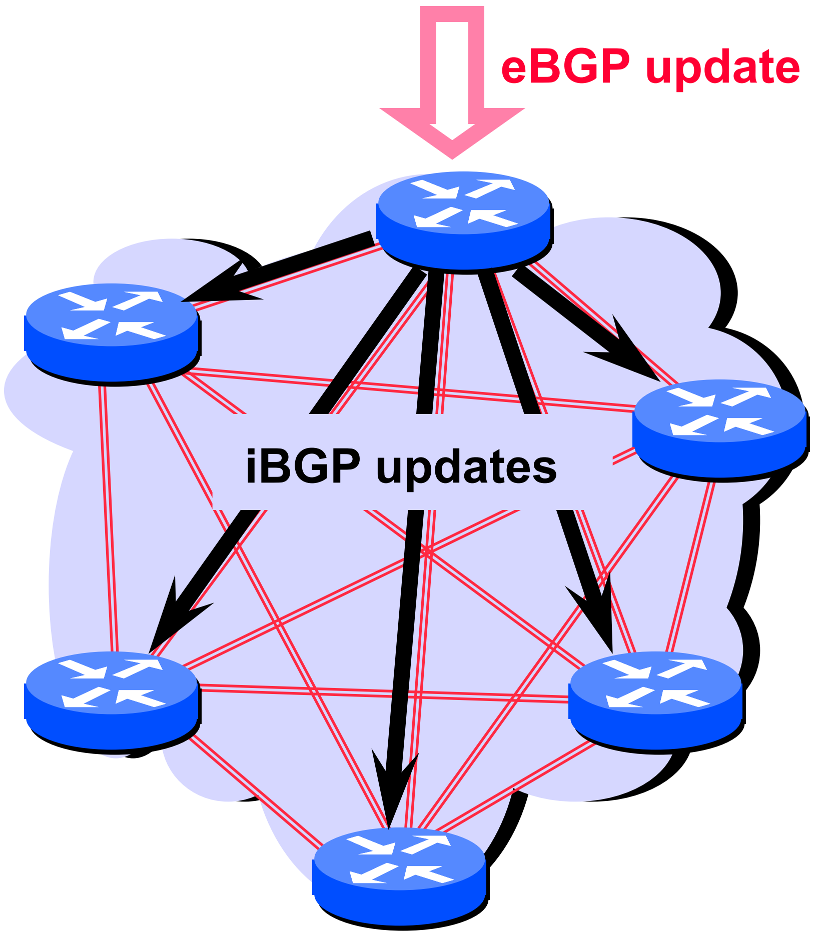
Two Types of BGP Neighbor Relationships

- **External Neighbor (eBGP) in a different Autonomous Systems**
- **Internal Neighbor (iBGP) in the same Autonomous System**

iBGP is <u>routed</u> (using IGP!)



iBGP Mesh Does Not Scale



- N border routers means N(N-1)/2 peering sessions
- Each router must have N-1 iBGP sessions configured
- The addition a single iBGP speaker requires configuration changes to all other iBGP speakers
- Size of iBGP routing table can be order N larger than number of best routes (remember alternate routes!)
- Each router has to listen to update noise from each neighbor

Currently four solutions:
(0) Buy bigger routers!
(1) Break AS into smaller ASes
(2) BGP Route reflectors
(3) BGP confederations

Four Types of BGP Messages

- **Open : Establish a peering session.**
- **Keep Alive : Handshake at regular intervals.**
- **Notification : Shuts down a peering session.**
- **Update : <u>Announcing</u> new routes or <u>withdrawing</u> previously announced routes.**

announcement

=

prefix + <u>attributes values</u>

BGP Attributes

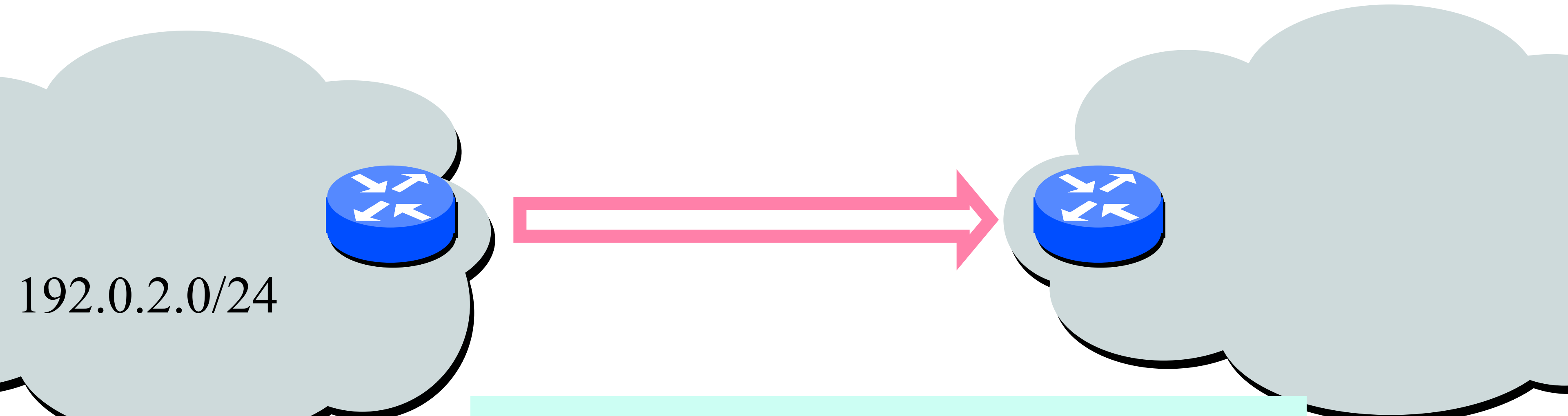
Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	[Rosen]
...		
255	reserved for development	

From IANA: http://www.iana.org/assignments/bgp-parameters

Most important attributes

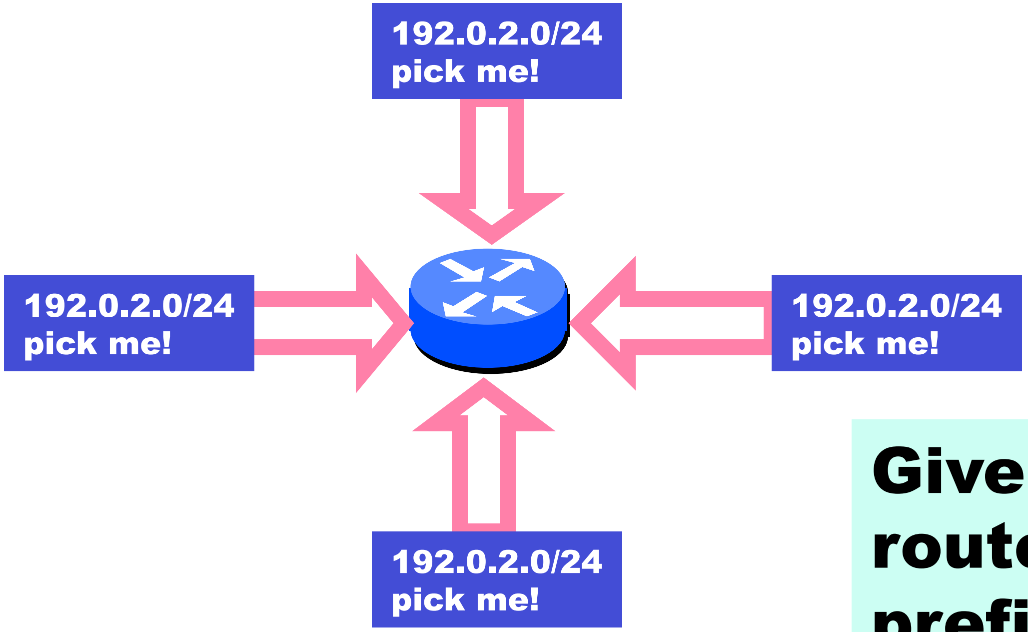
Not all attributes need to be present in every announcement

Announcing a route



For 192.0.2.0/24
Next Hop
AS Path
Other attributes

Attributes are Used to Select Best Routes



Given multiple routes to the same prefix, a BGP speaker must pick at most <u>one</u> best route

(Note: it could reject them all!)

Route Selection Summary

Highest Local Preference	Enforce relationships
Shortest ASPATH Lowest MED i-BGP < e-BGP Lowest IGP cost to BGP egress	traffic engineering
Lowest router ID	Throw up hands and break ties

BGP Route Processing

Open ended programming.
Constrained only by vendor configuration language

Receive BGP Updates

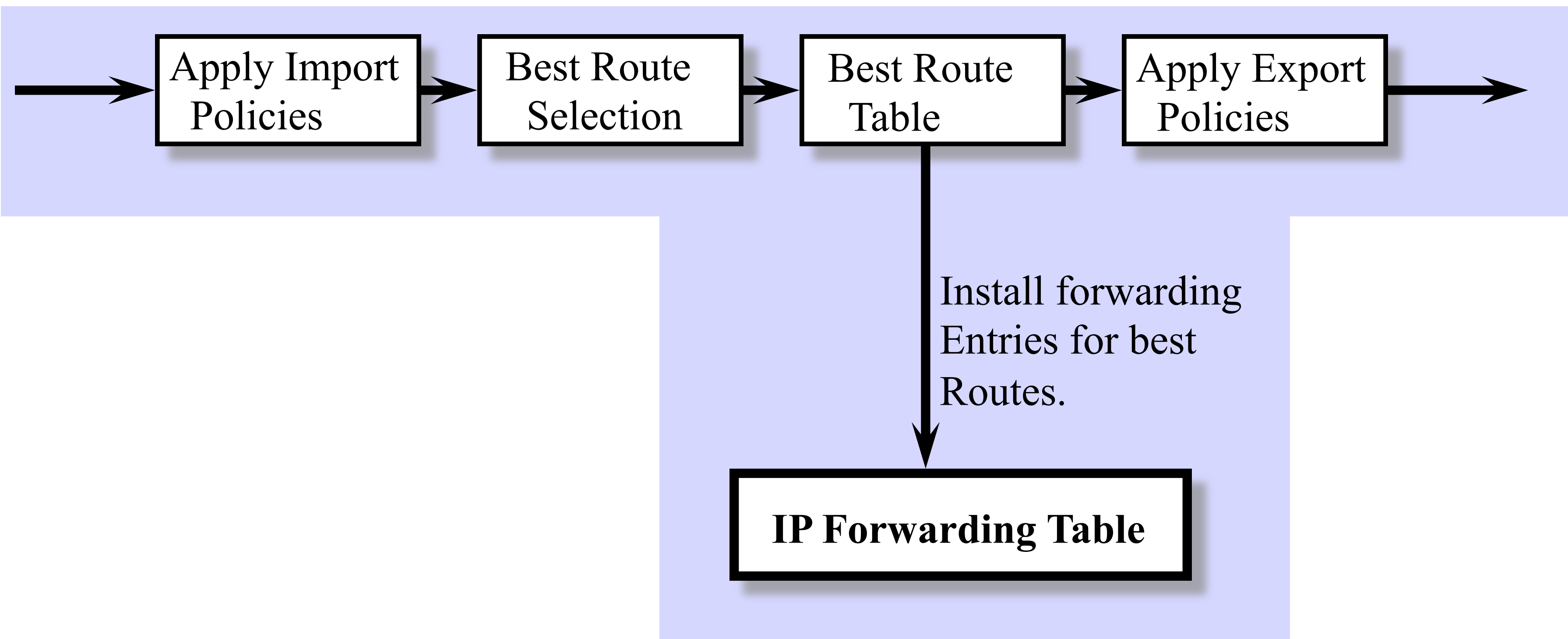
Apply Policy = filter routes & tweak attributes

Based on Attribute Values

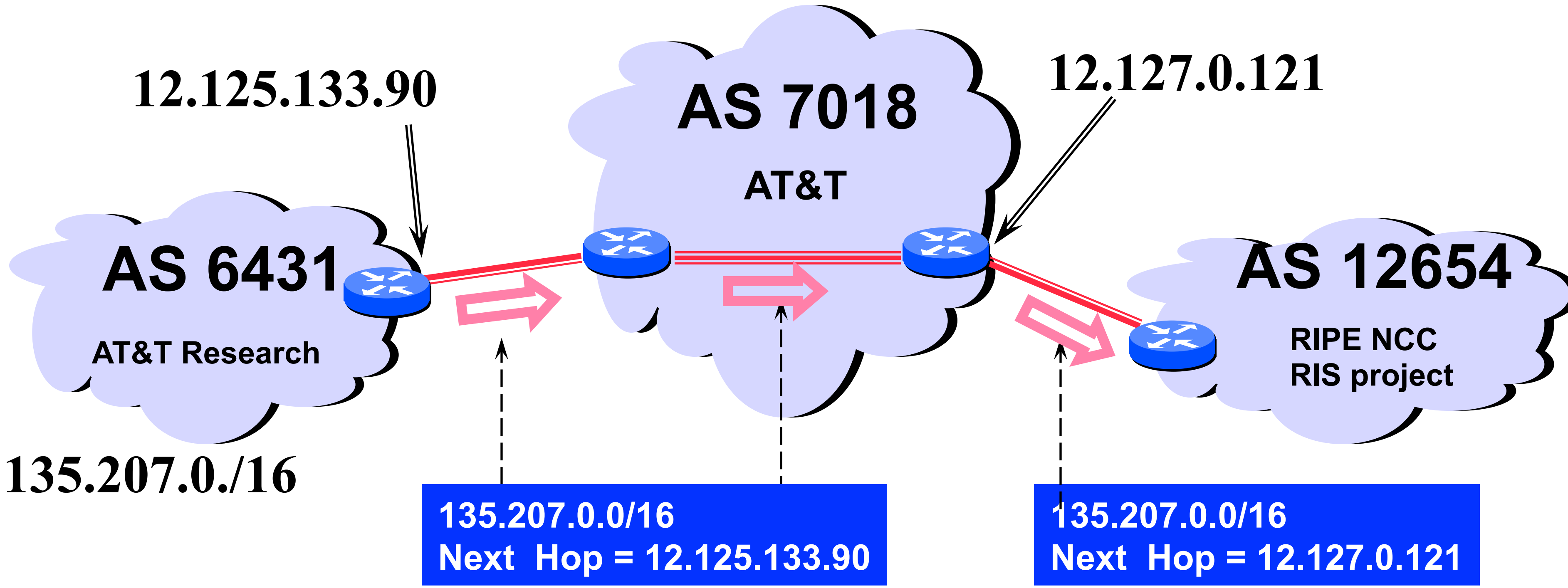
Best Routes

Apply Policy = filter routes & tweak attributes

Transmit BGP Updates

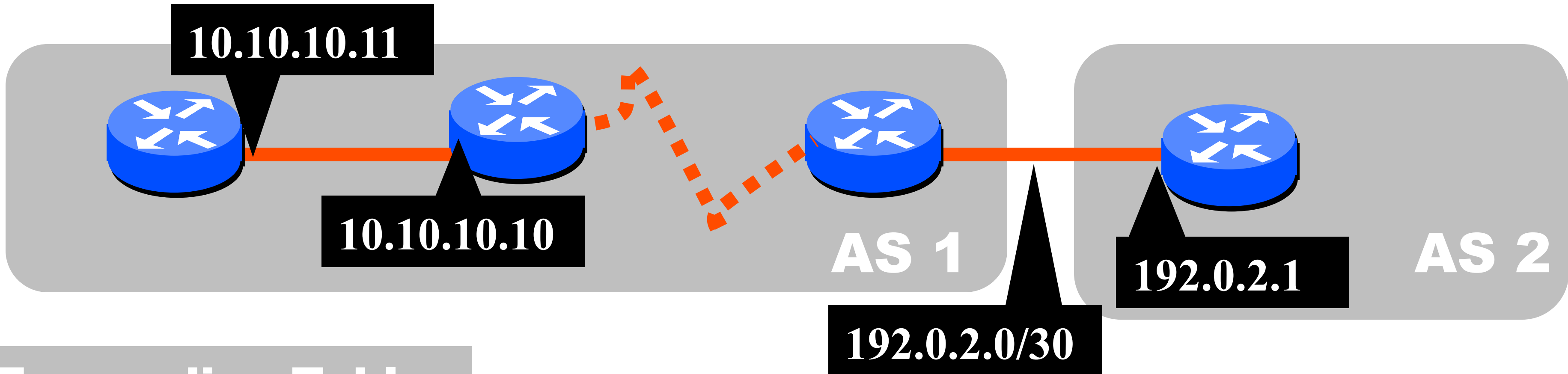


BGP Next Hop Attribute



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.

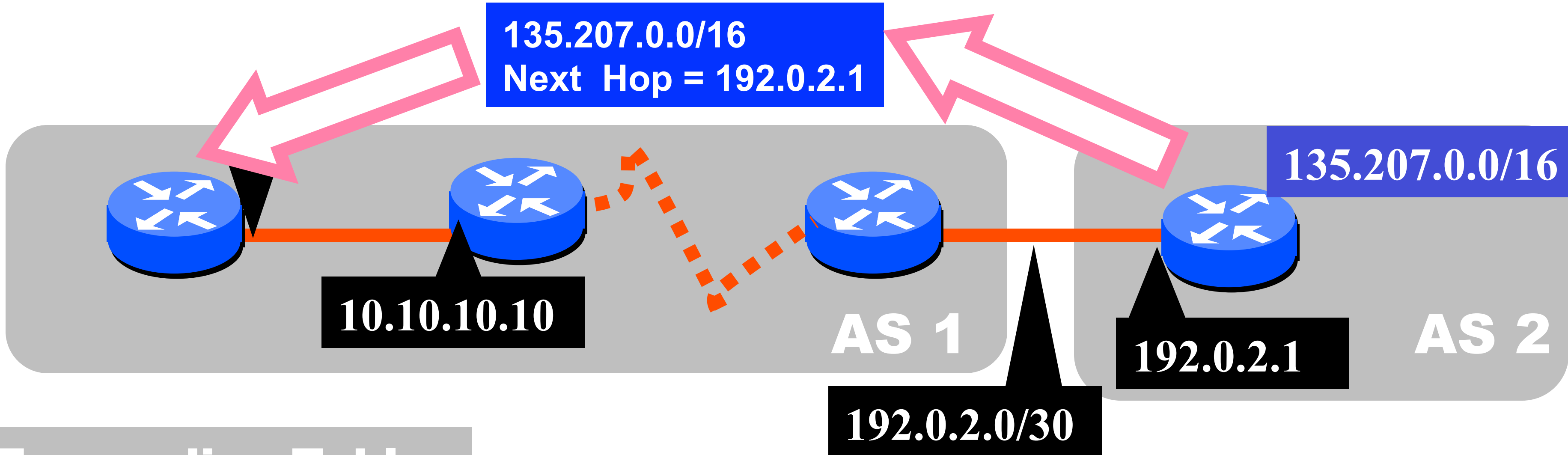
Join EGP with IGP For Connectivity



Forwarding Table

destination	next hop
192.0.2.0/30	10.10.10.10

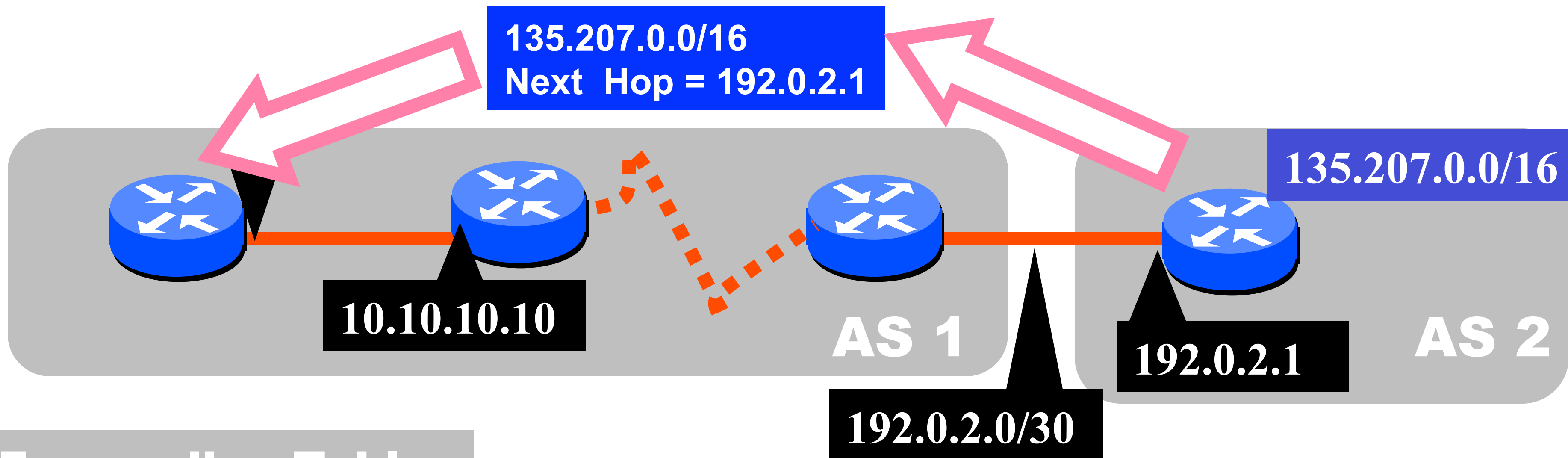
Join EGP with IGP For Connectivity



Forwarding Table

destination	next hop
192.0.2.0/30	10.10.10.10

Join EGP with IGP For Connectivity



Forwarding Table

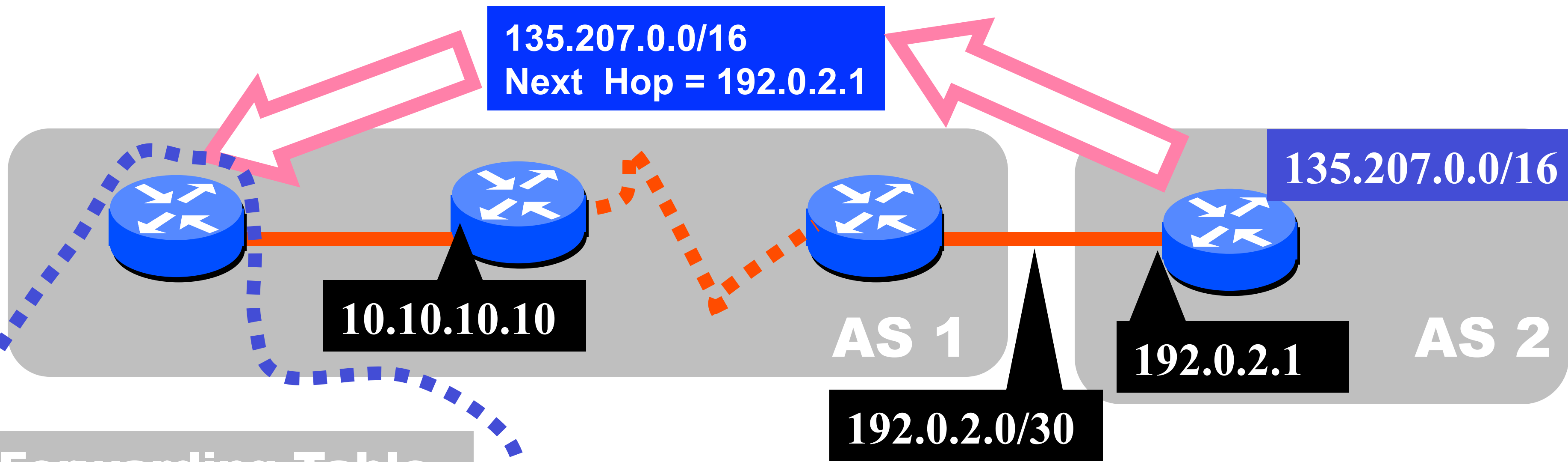
destination	next hop
192.0.2.0/30	10.10.10.10

+

EGP

destination	next hop
135.207.0.0/16	192.0.2.1

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

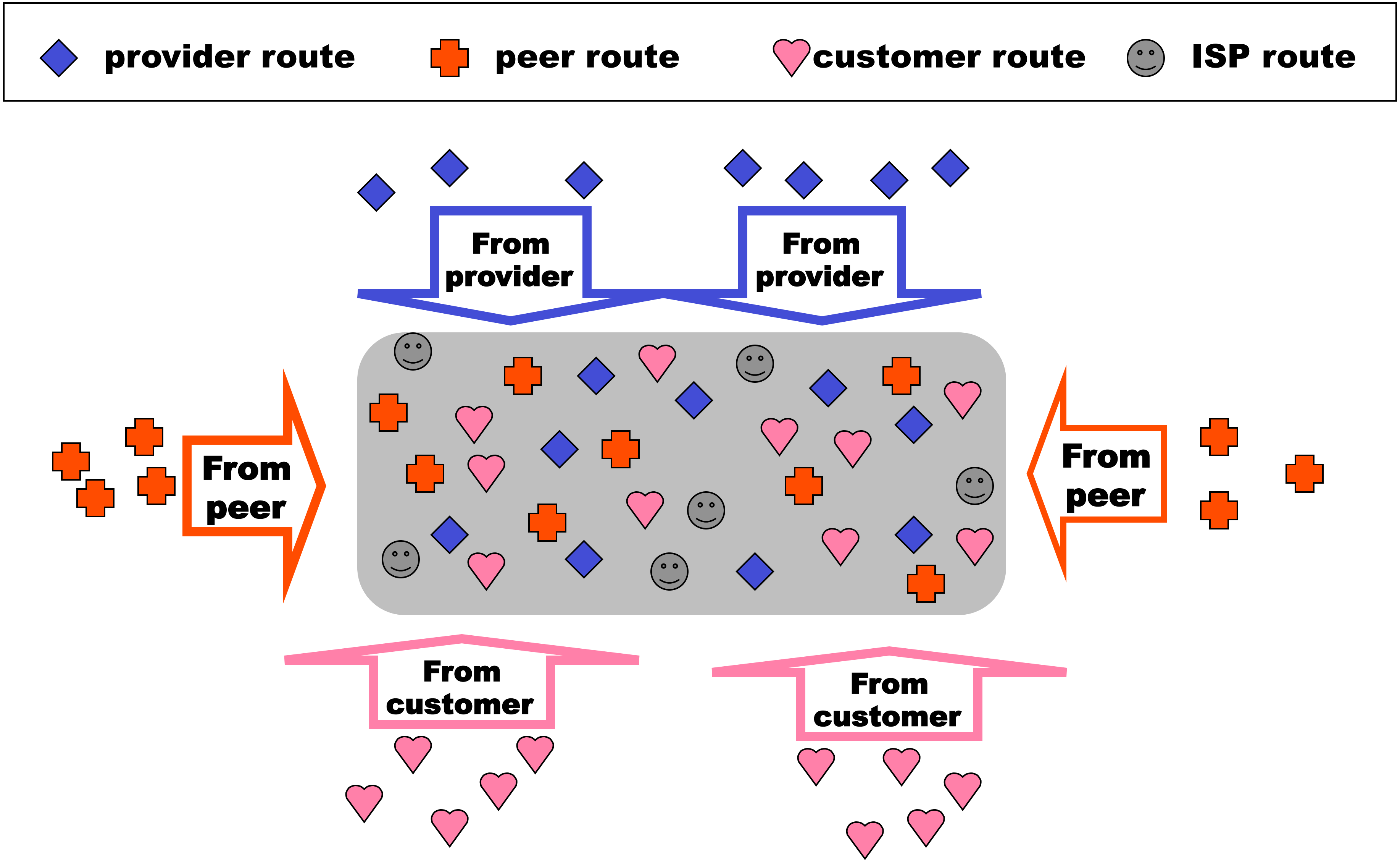
destination	next hop
135.207.0.0/16	10.10.10.10
192.0.2.0/30	10.10.10.10

Implementing Customer/Provider and Peer/Peer relationships

Two parts:

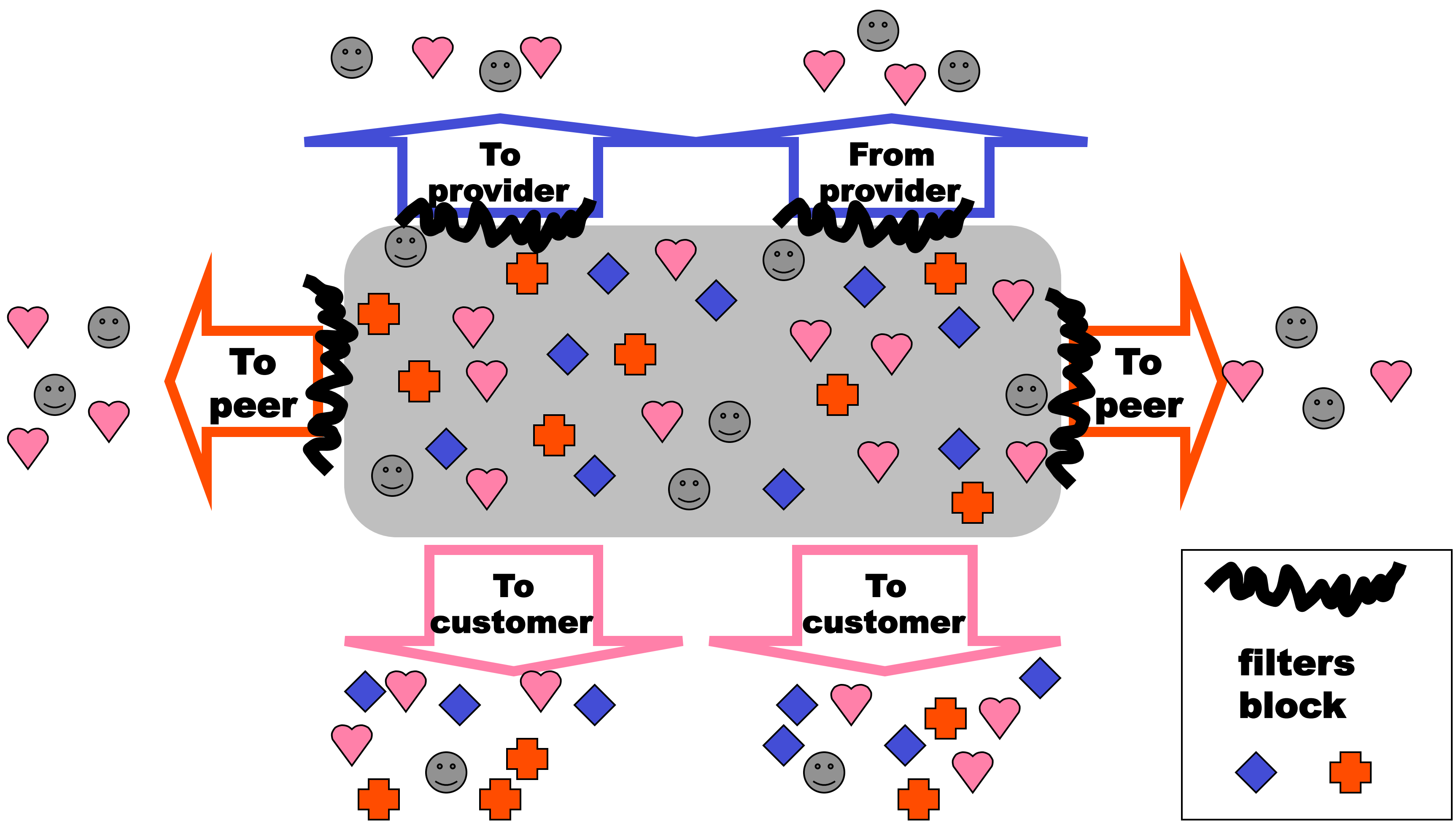
- Enforce transit relationships
  - Outbound route filtering
- Enforce order of route preference
  - provider < peer < customer

Import Routes

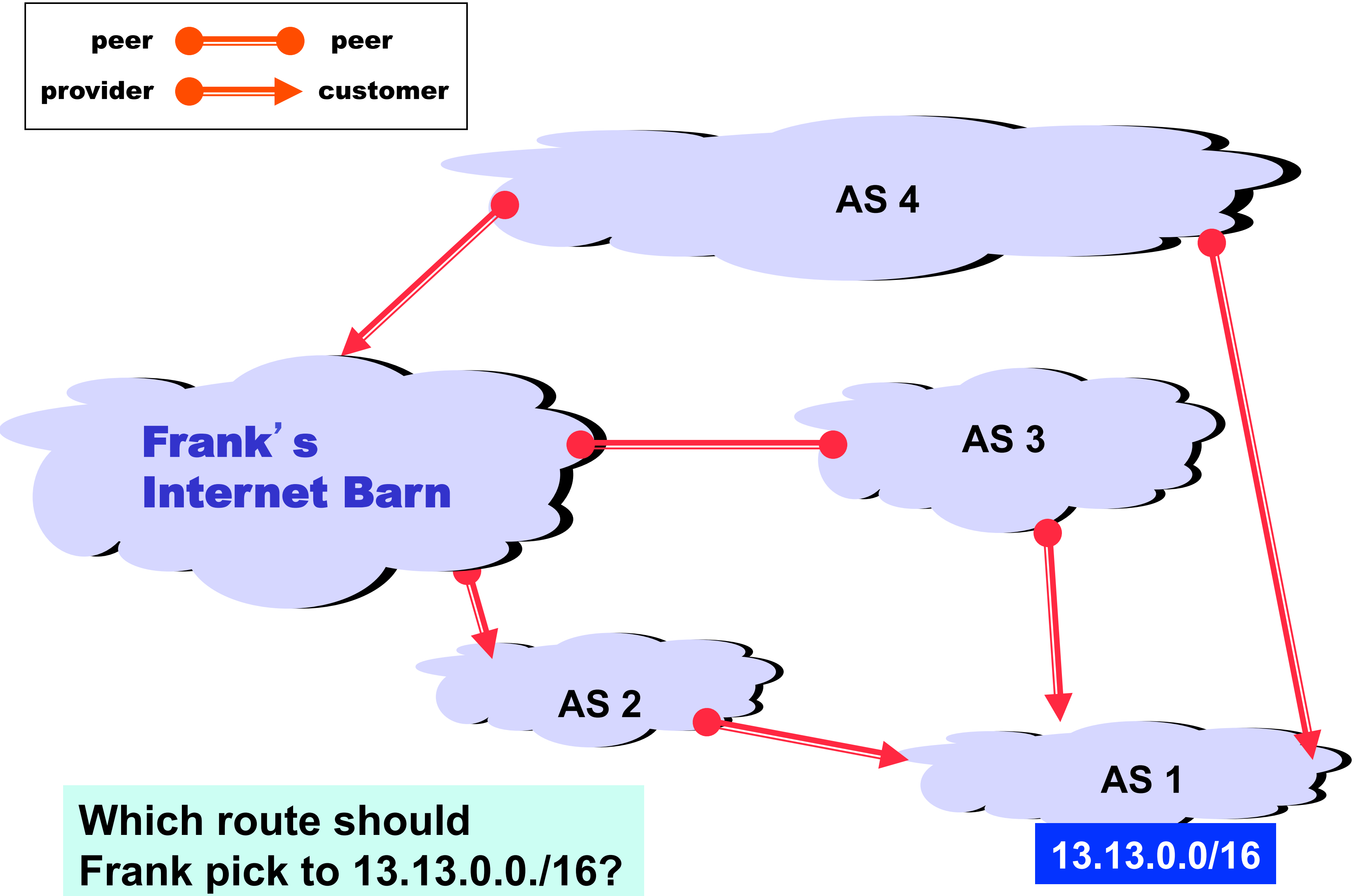


Export Routes

provider route | peer route | customer route | ISP route



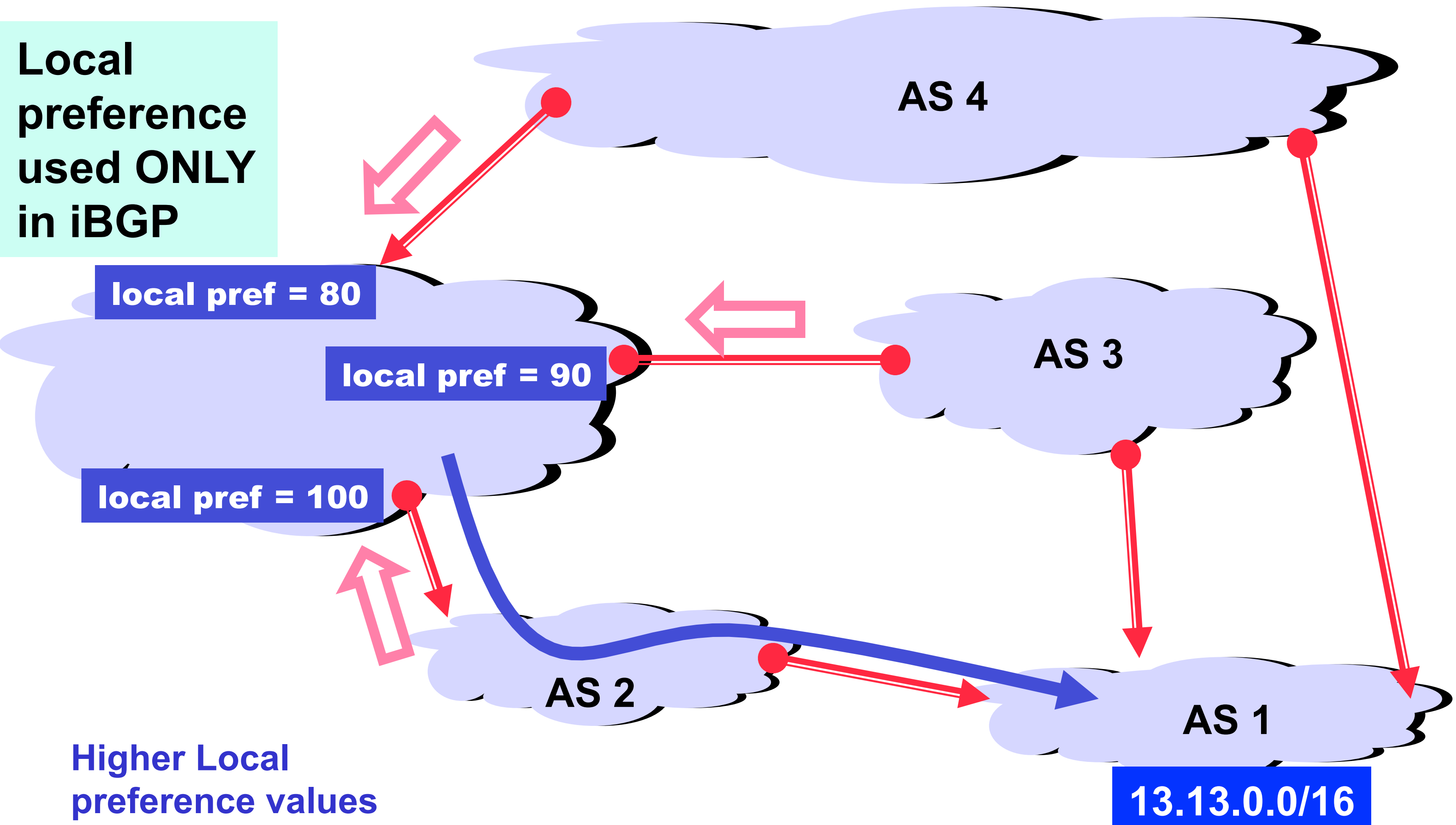
So Many Choices



Which route should
Frank pick to 13.13.0.0./16?

LOCAL PREFERENCE

Local preference used ONLY in iBGP



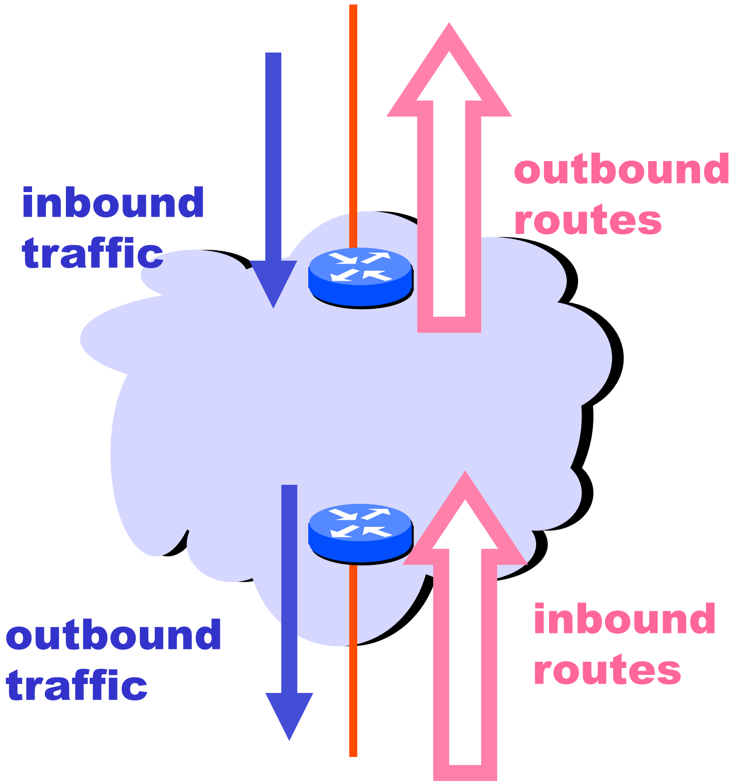
Higher Local preference values are more preferred

PART IV

Traffic Engineering with BGP

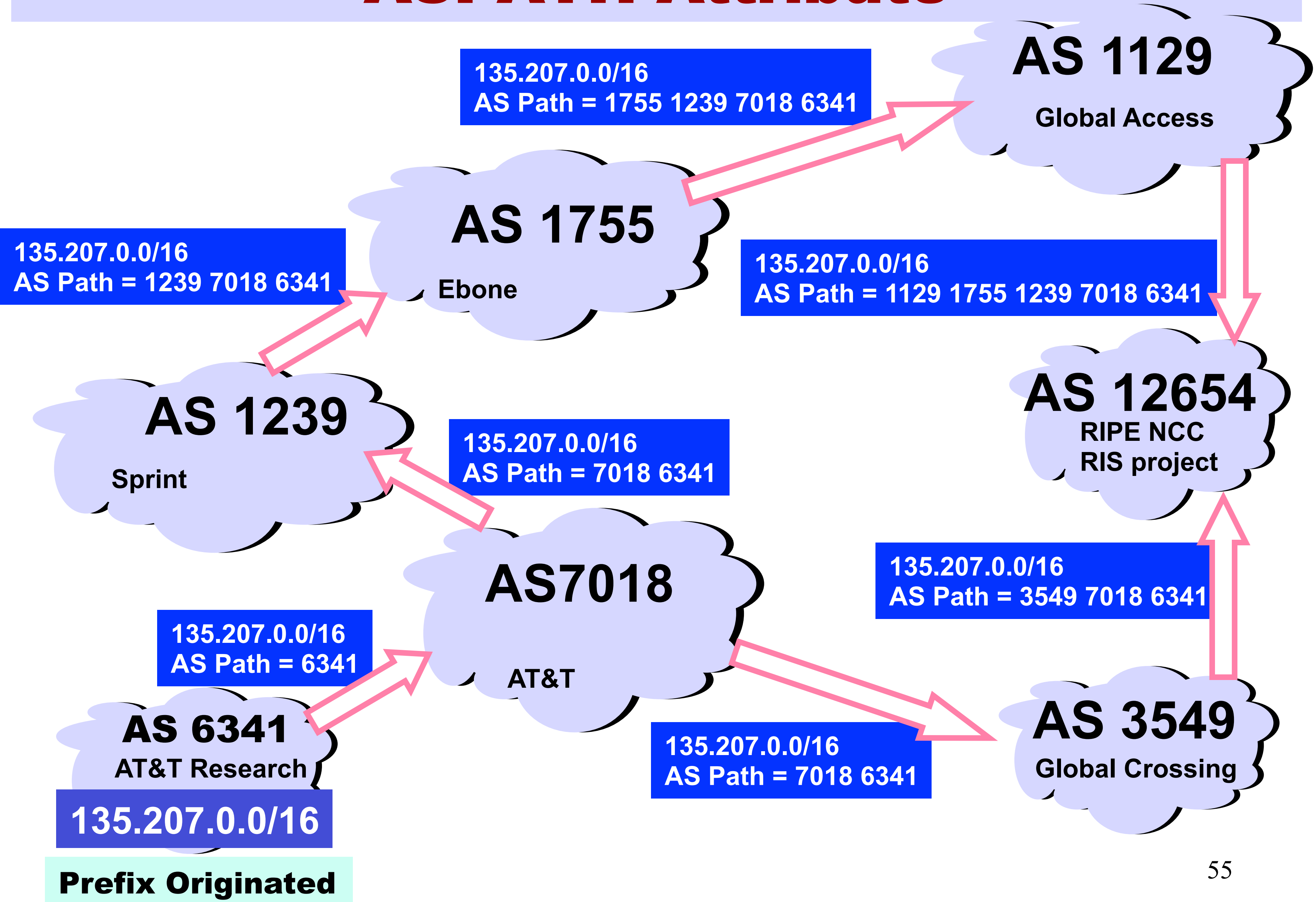
Tweak Tweak Tweak

- For <u>inbound</u> traffic
  - Filter outbound routes
  - Tweak attributes on <u>outbound</u> routes in the hope of influencing your neighbor's best route selection
- For <u>outbound</u> traffic
  - Filter <u>inbound</u> routes
  - Tweak attributes on <u>inbound</u> routes to influence best route selection

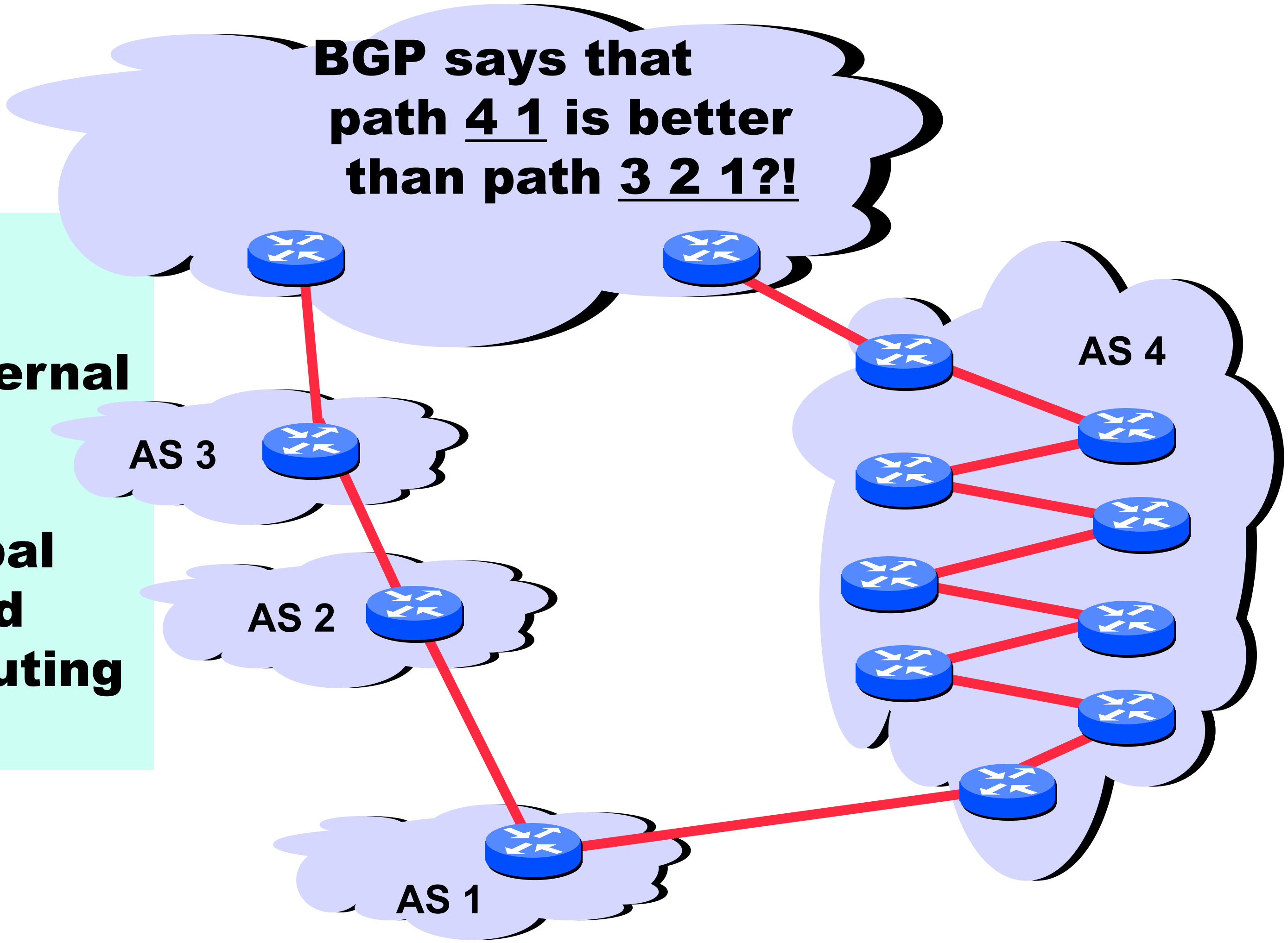


In general, an AS has more control over outbound traffic

ASPATH Attribute



Shorter Doesn't Always Mean Shorter

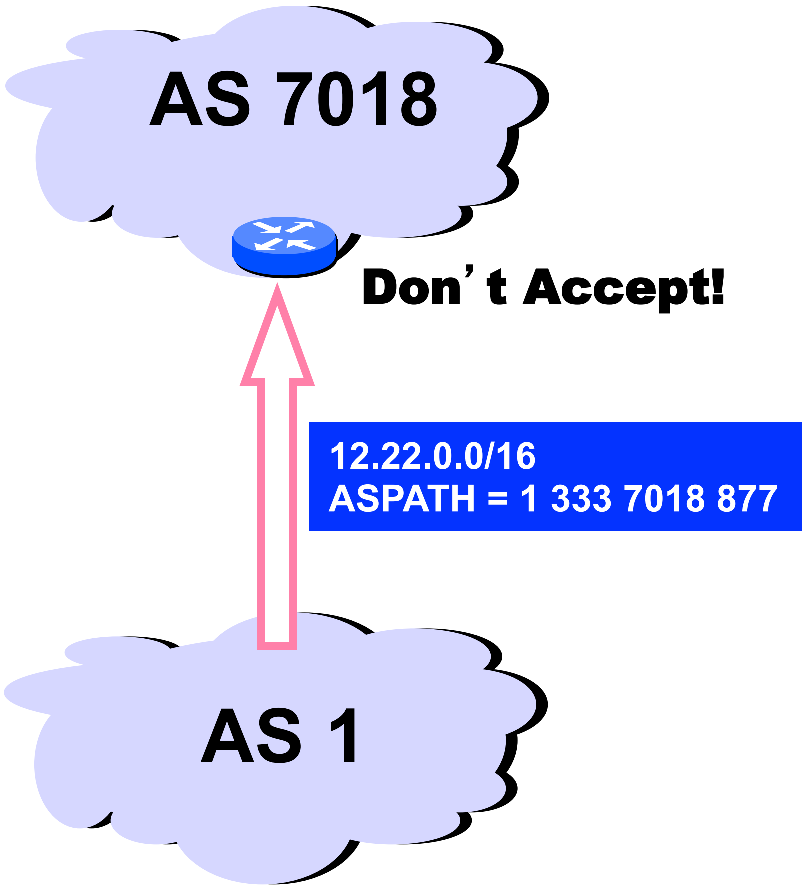


But …

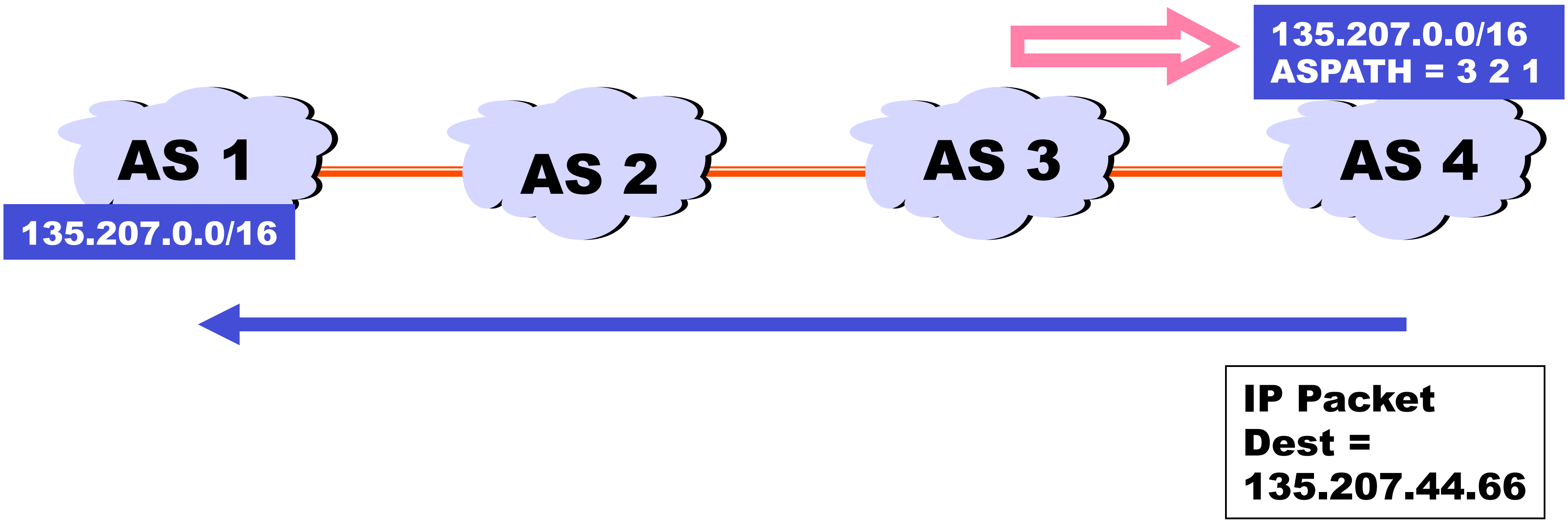
Exporting internal state would dramatically increase global instability and amount of routing state

Interdomain Loop Prevention

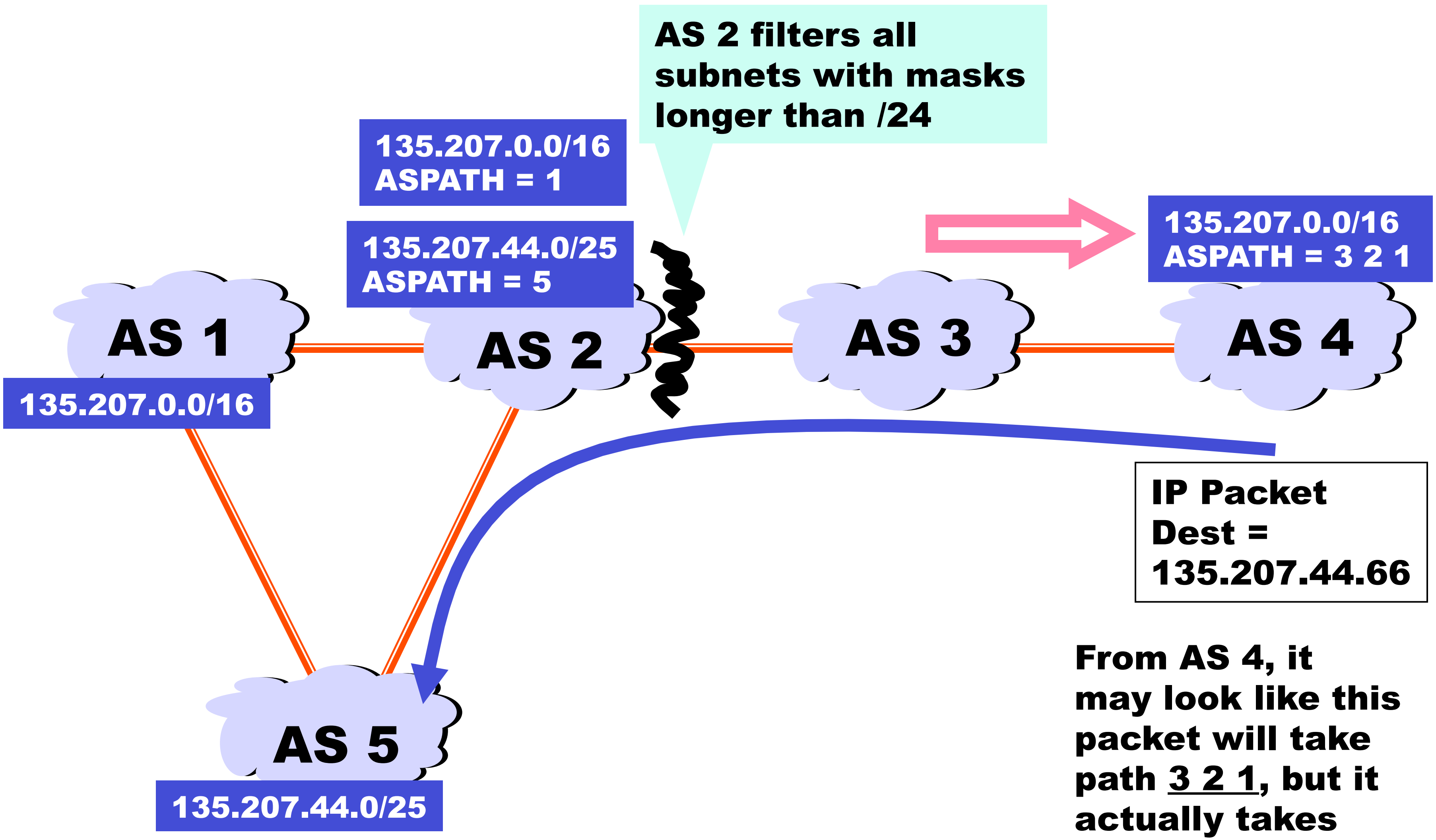
BGP at AS YYY will never accept a route with ASPATH containing YYY.



Traffic Often Follows ASPATH

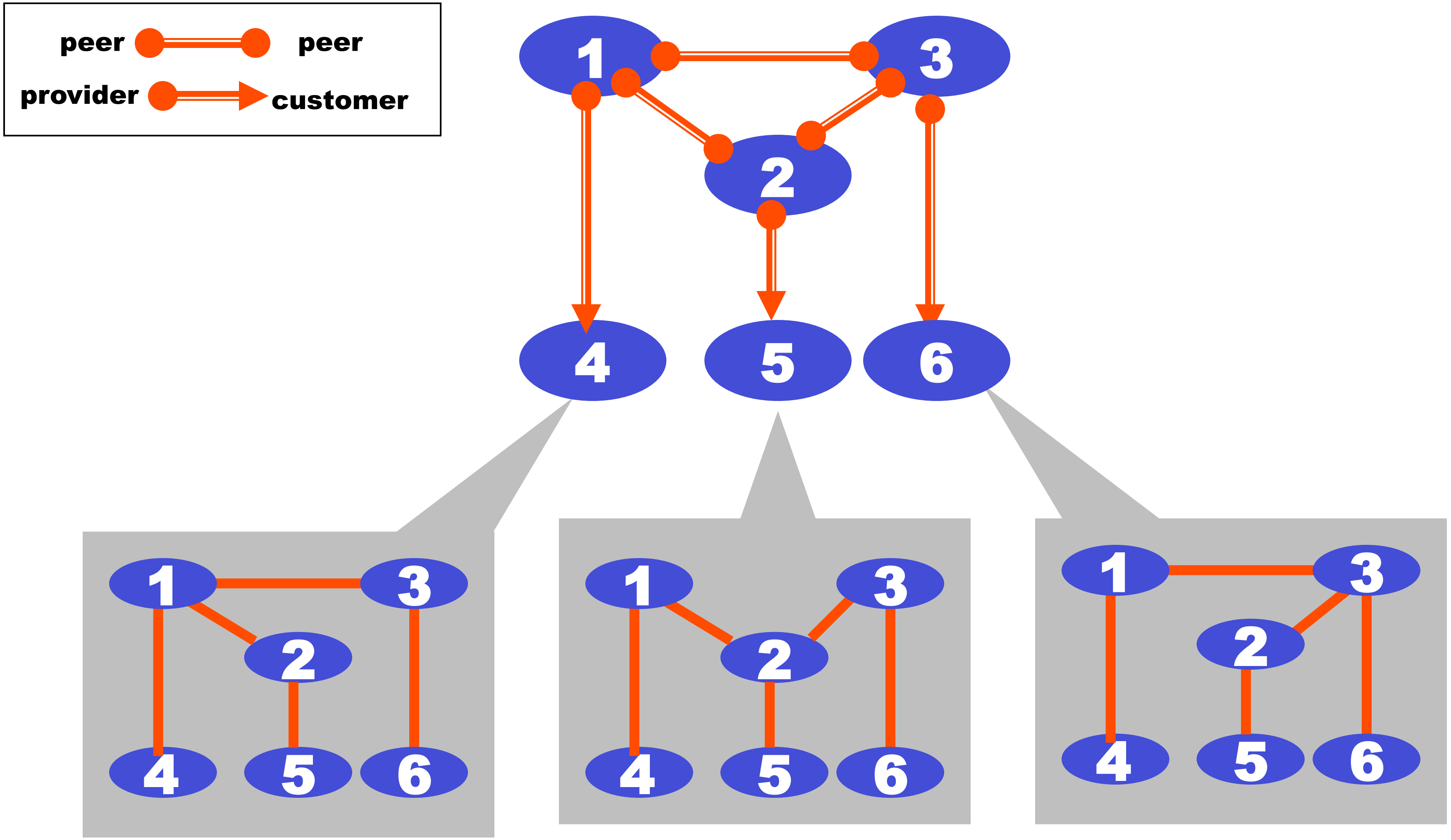


… But It Might Not

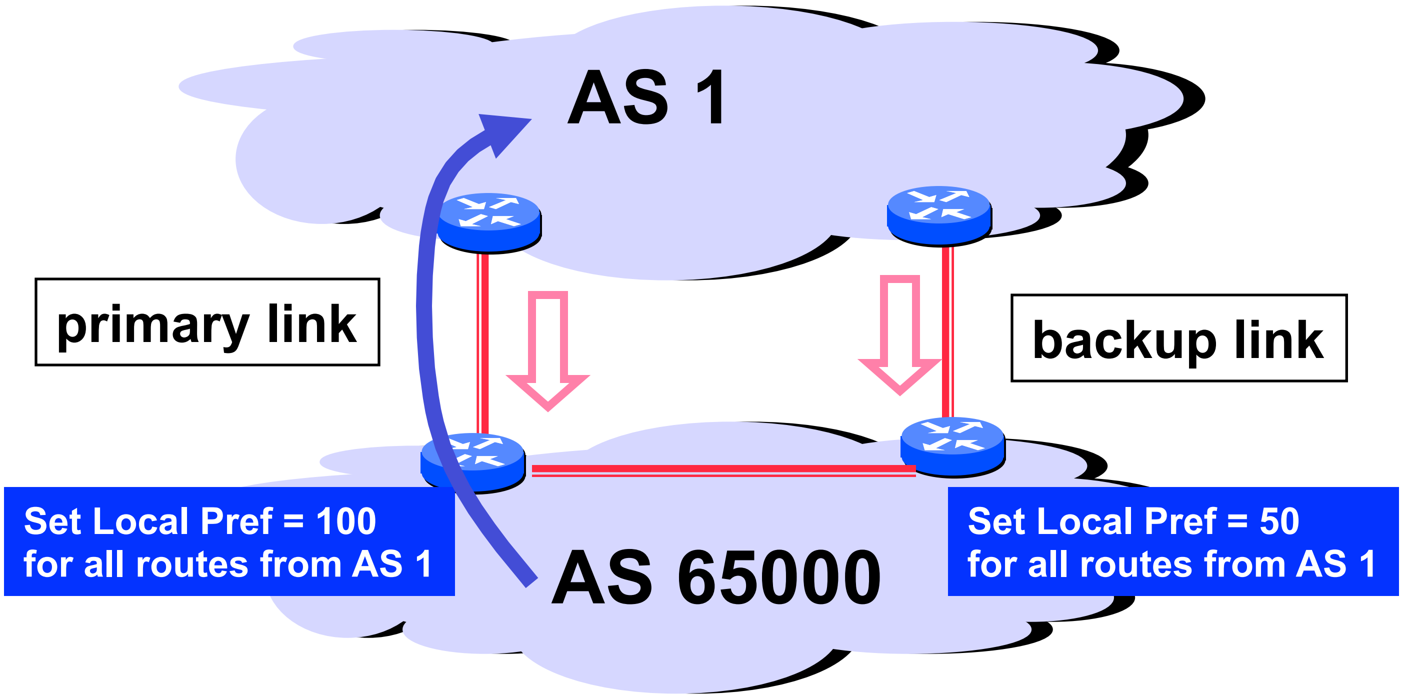


From AS 4, it may look like this packet will take path 3 2 1, but it actually takes path 3 2 5

AS Graphs Depend on Point of View



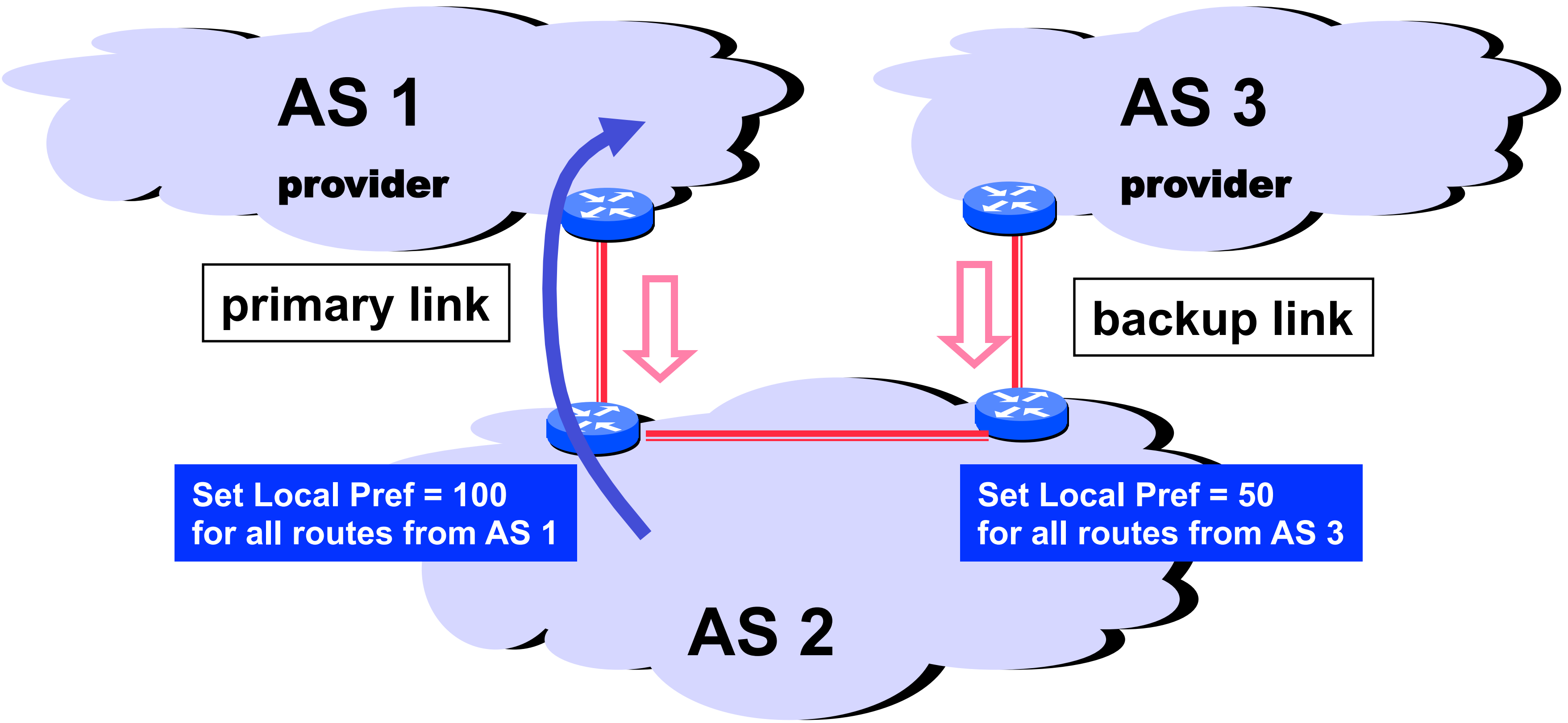
Implementing Backup Links with Local Preference (Outbound Traffic)



Forces outbound traffic to take primary link, unless link is down.

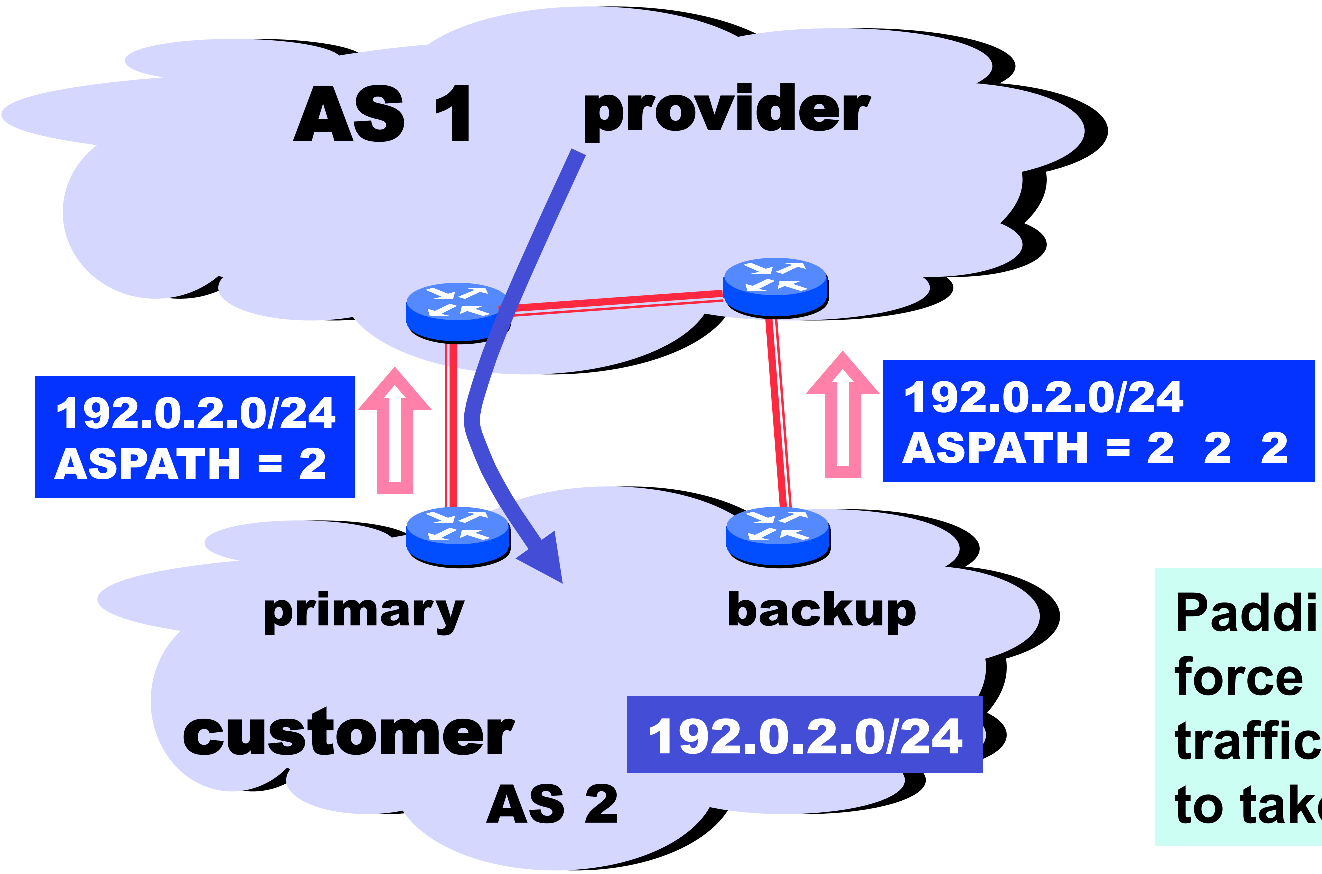
We'll talk about inbound traffic soon …

Multihomed Backups (Outbound Traffic)



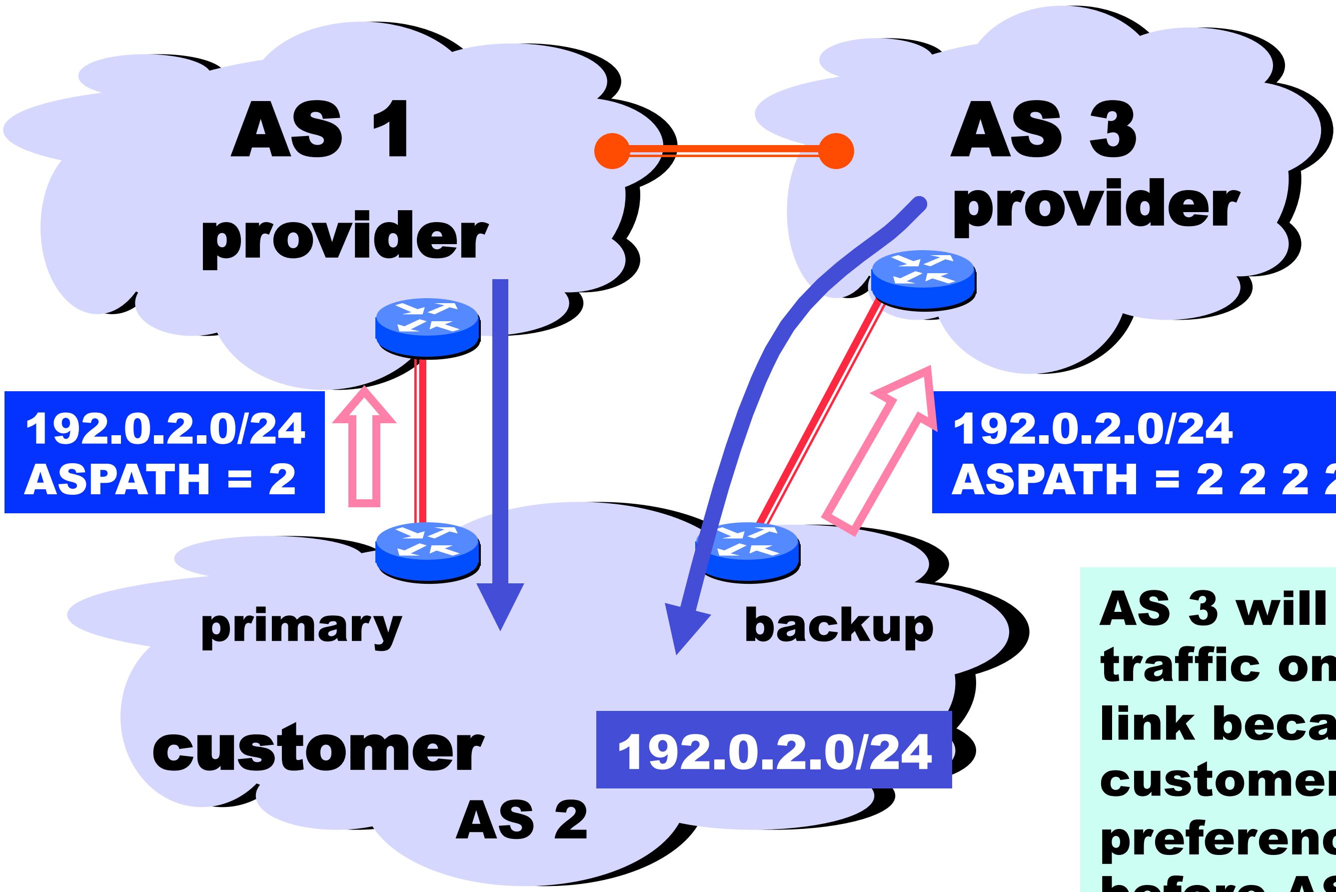
Forces outbound traffic to take primary link, unless link is down.

Shedding Inbound Traffic with ASPATH Padding. Yes, this is a Glorious Hack …



Padding will (usually) force inbound traffic from AS 1 to take primary link

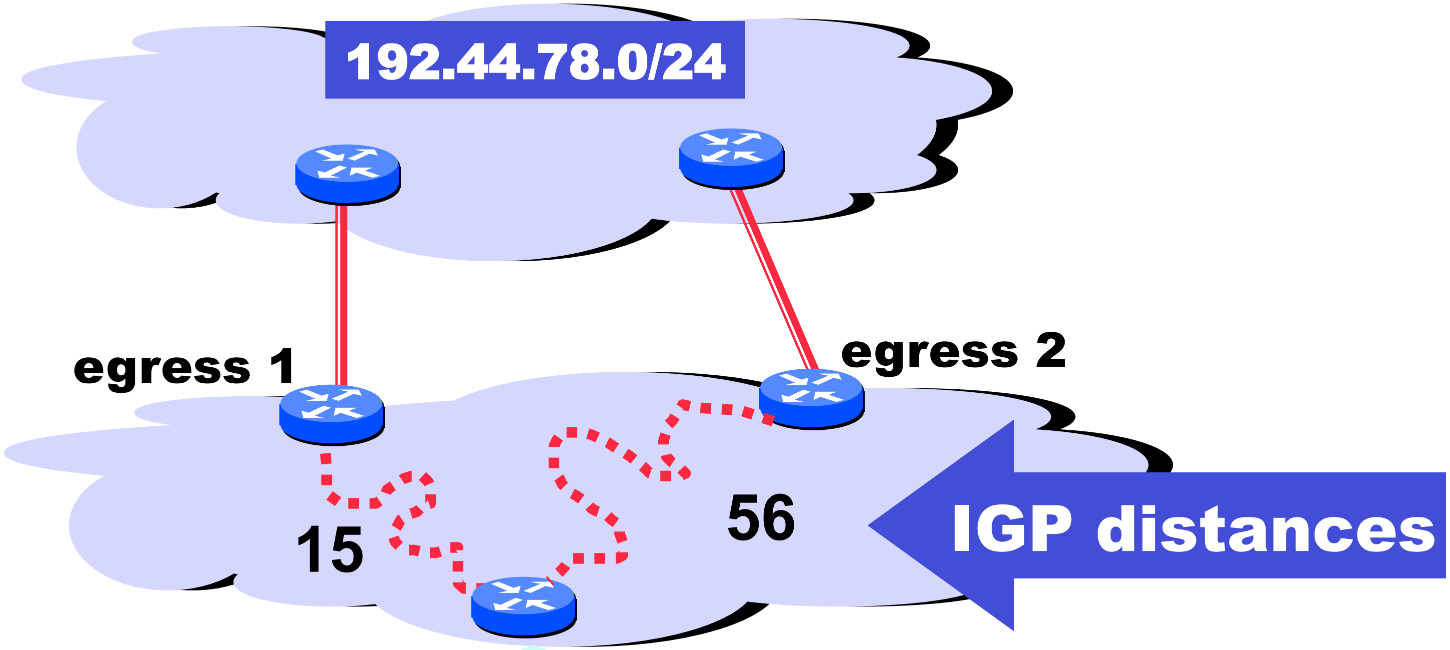
… But Padding Does Not Always Work



AS 3 will send traffic on "backup" link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing

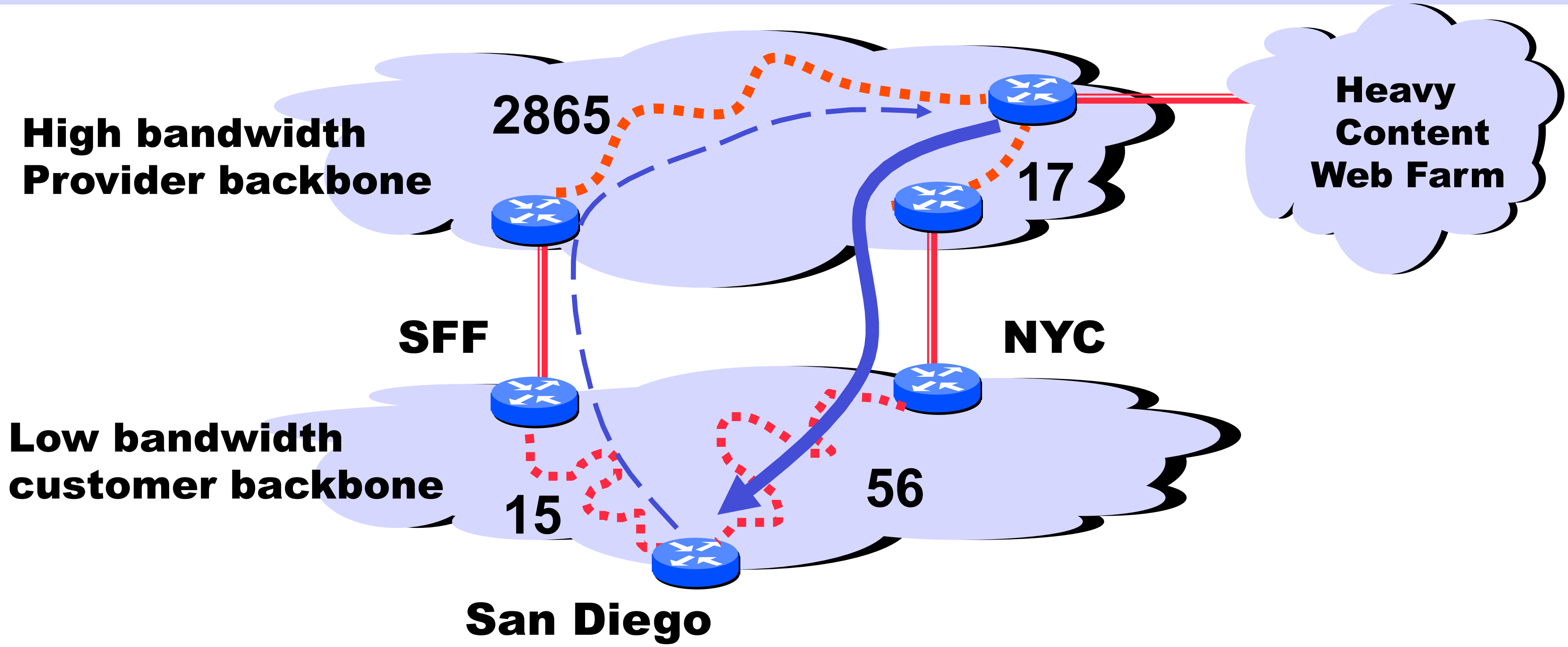
Hot Potato Routing: Go for the Closest Egress Point



This Router has two BGP routes to 192.44.78.0/24.

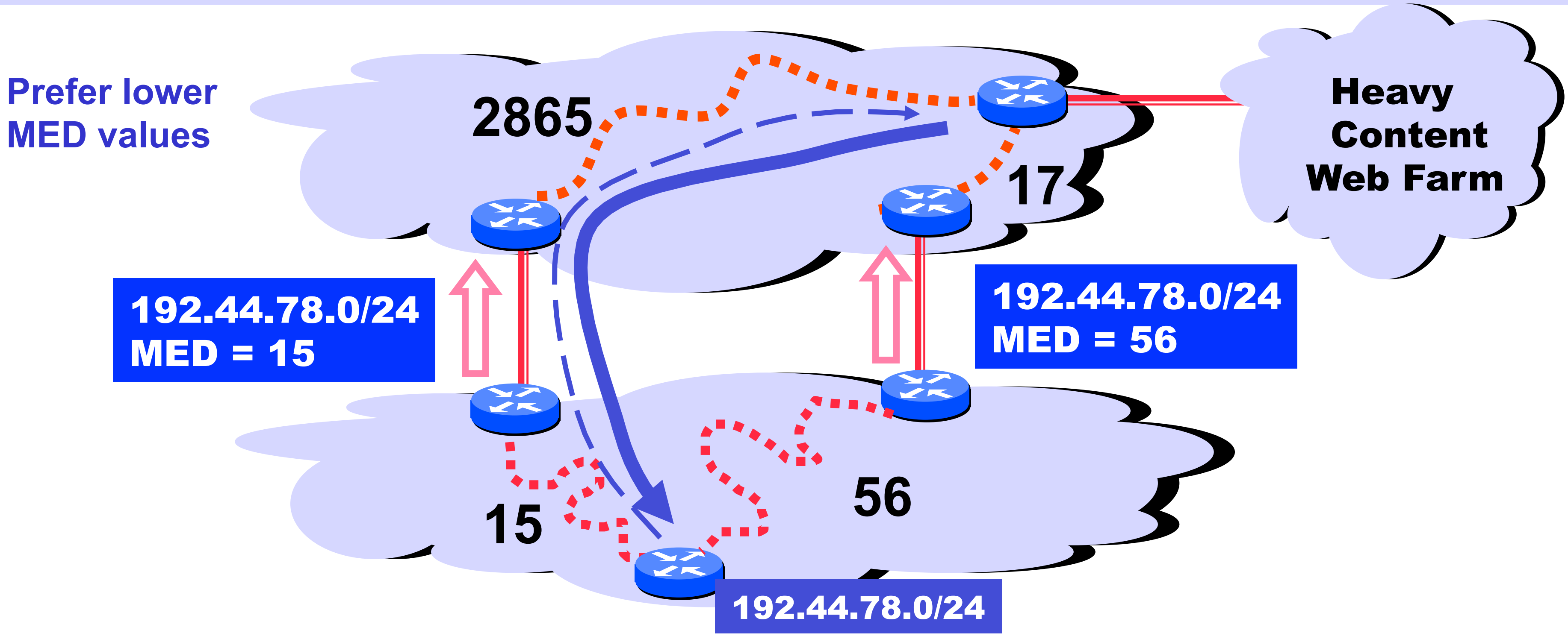
Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!

Getting Burned by the Hot Potato



Many customers want their provider to carry the bits!

Cold Potato Routing with MEDs (Multi-Exit Discriminator Attribute)



This means that MEDs must be considered BEFORE IGP distance!

Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance