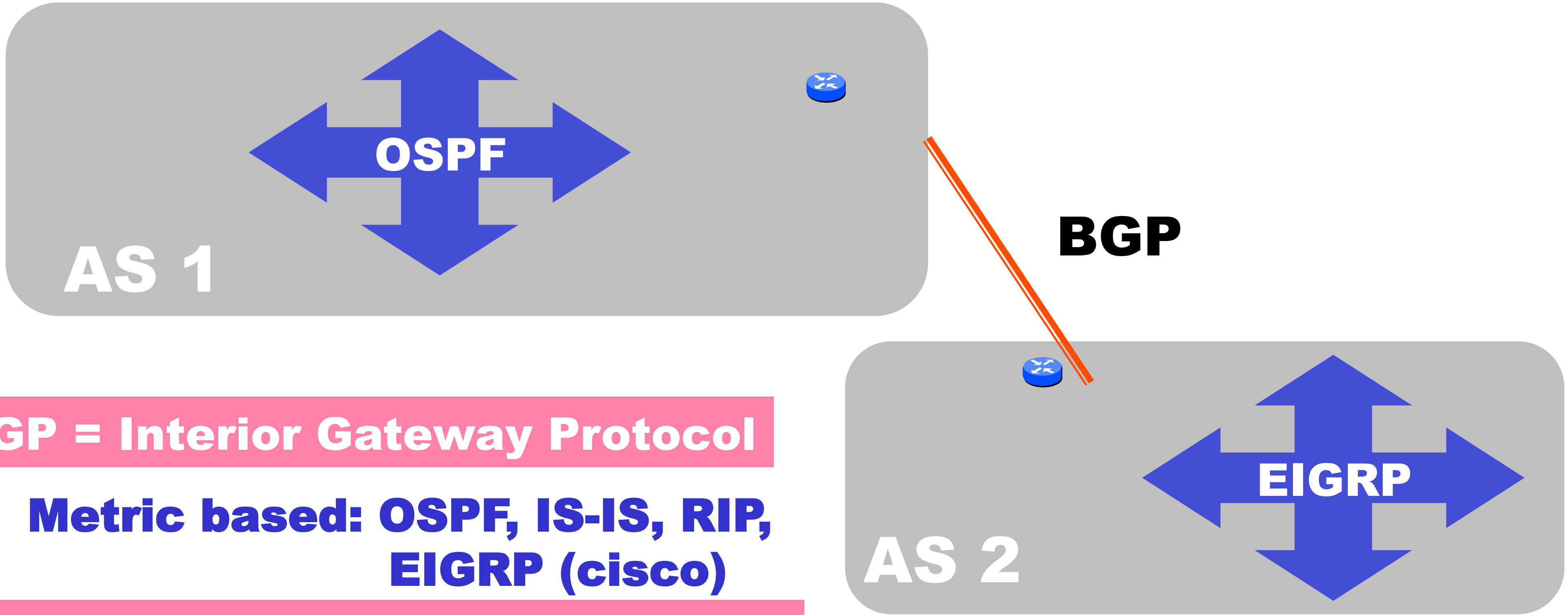
Inter-Domain Routing & BGP

Outline

The glue that holds the Internet together : interdomain routing with The Border Gateway Protocol (BGP)

Architecture of Dynamic 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

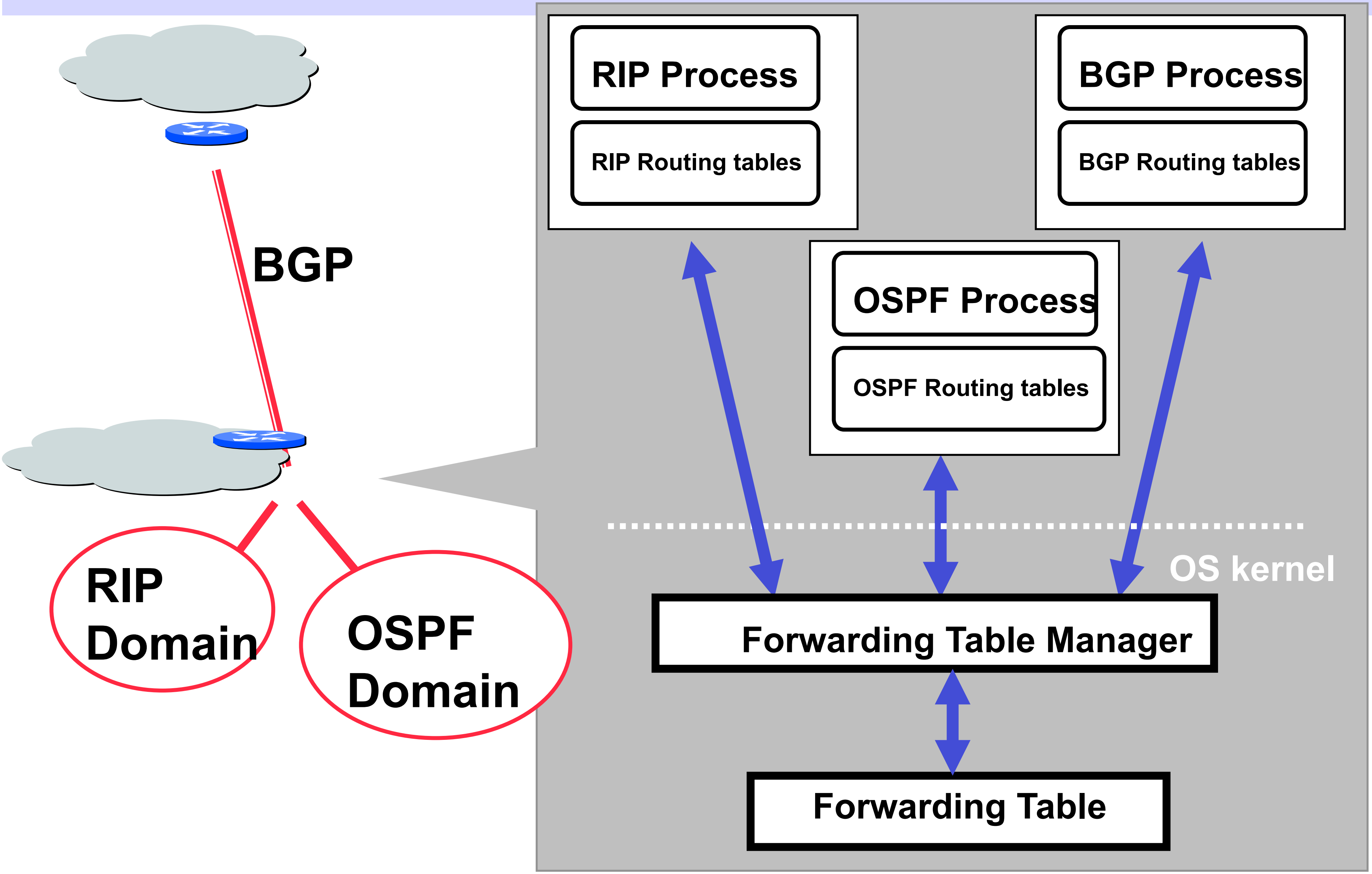
Vectoring

- 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

The Gang of Four

	Link State	Vectoring
IGP	OSPF IS-IS	RIP
EGP		BGP

Many Routing Processes Can Run on a Single Router



Internet Hierarchy

- What is an Autonomous System (AS)?
  - A set of routers under a single technical administration, using an *intra-domain routing protocol* (IGP) and common metrics to route packets within the AS and using an *inter-domain routing protocol* (EGP) to route packets to other ASes
  - Sometimes ASes use multiple intra-domain routing protocols and metrics, but appear as a single AS to other ASes
- Each AS is assigned a unique ID

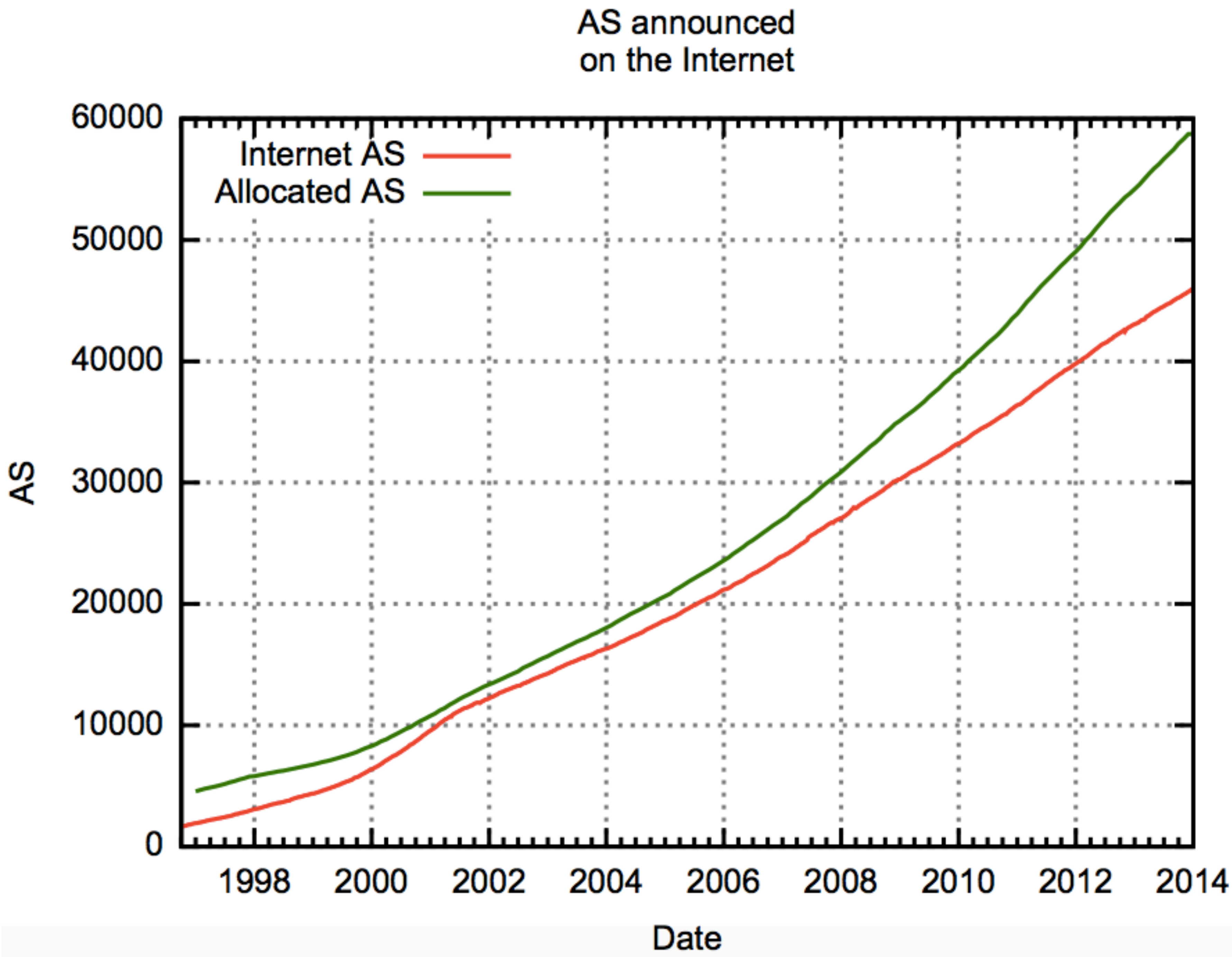
AS Numbers (ASNs)

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

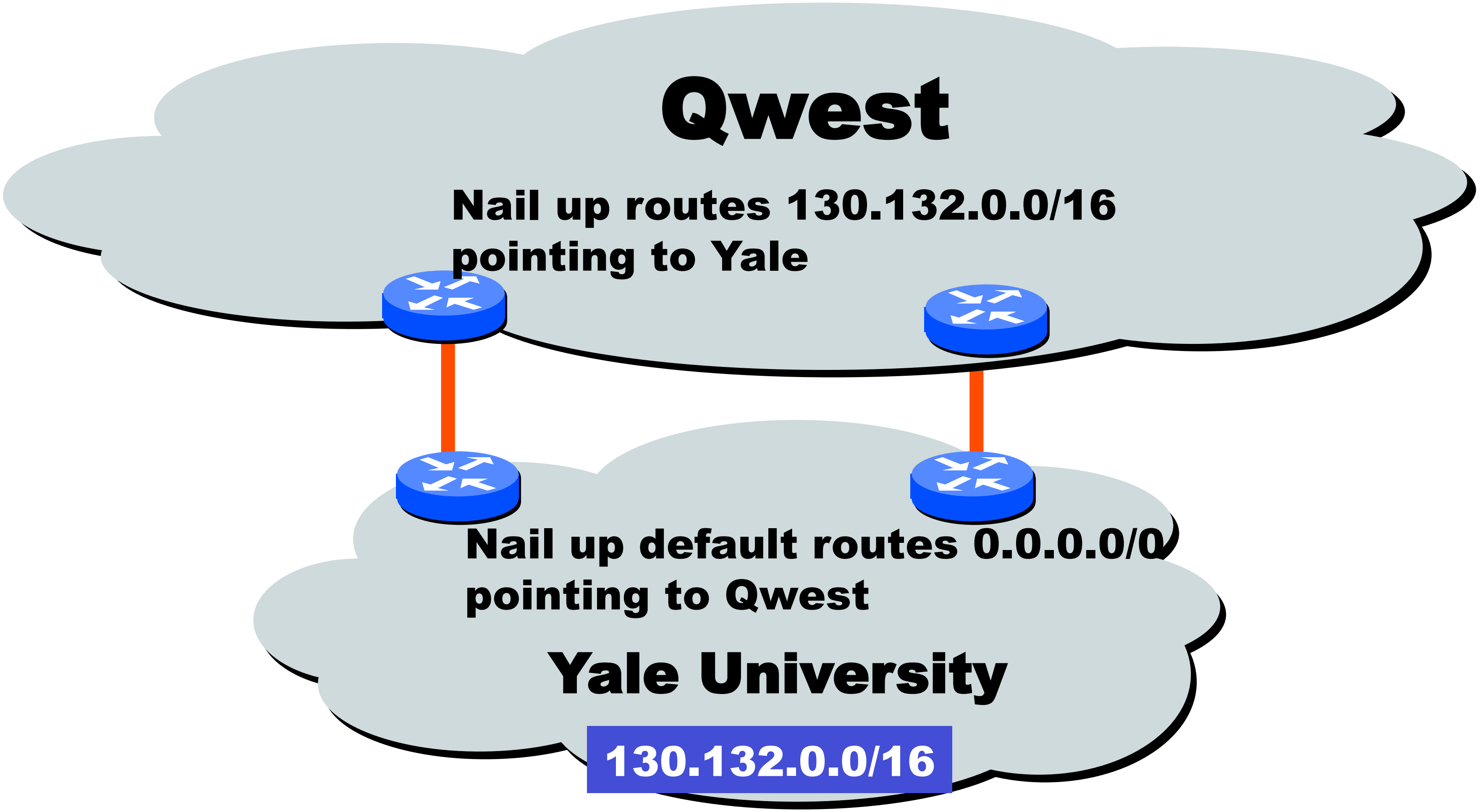
Currently over 12,000 in use.

- **Yale: 29**
- **MIT: 3**
- **Harvard: 11**
- **Genuity: 1**
- **AT&T: 7018, 6341, 5074, …**
- **UUNET: 701, 702, 284, 12199, …**
- **Sprint: 1239, 1240, 6211, 6242, …**
- **…**

ASNs represent units of routing policy

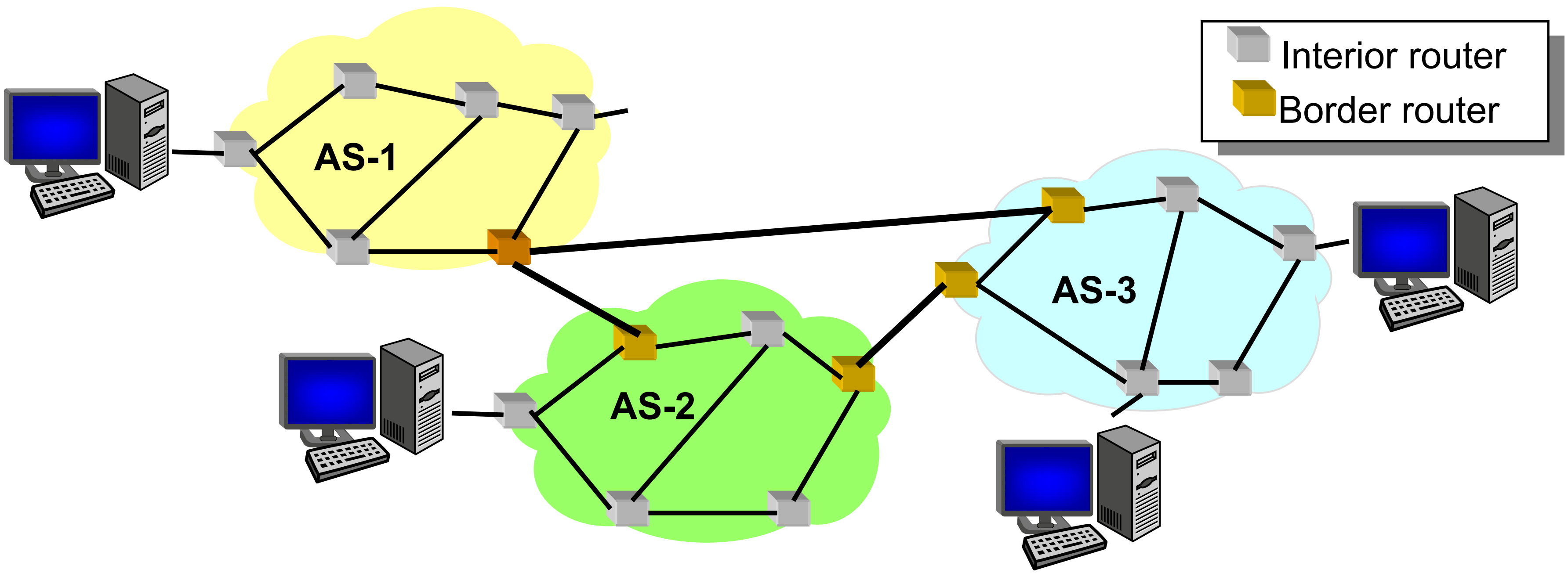

AS announced
on the Internet
Internet AS
Allocated AS
AS
Date
0
10000
20000
30000
40000
50000
60000
1998
2000
2002
2004
2006
2008
2010
2012
2014

Autonomous Routing Domains Don't Always Need BGP or an ASN



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 ...

Picture of the Internet



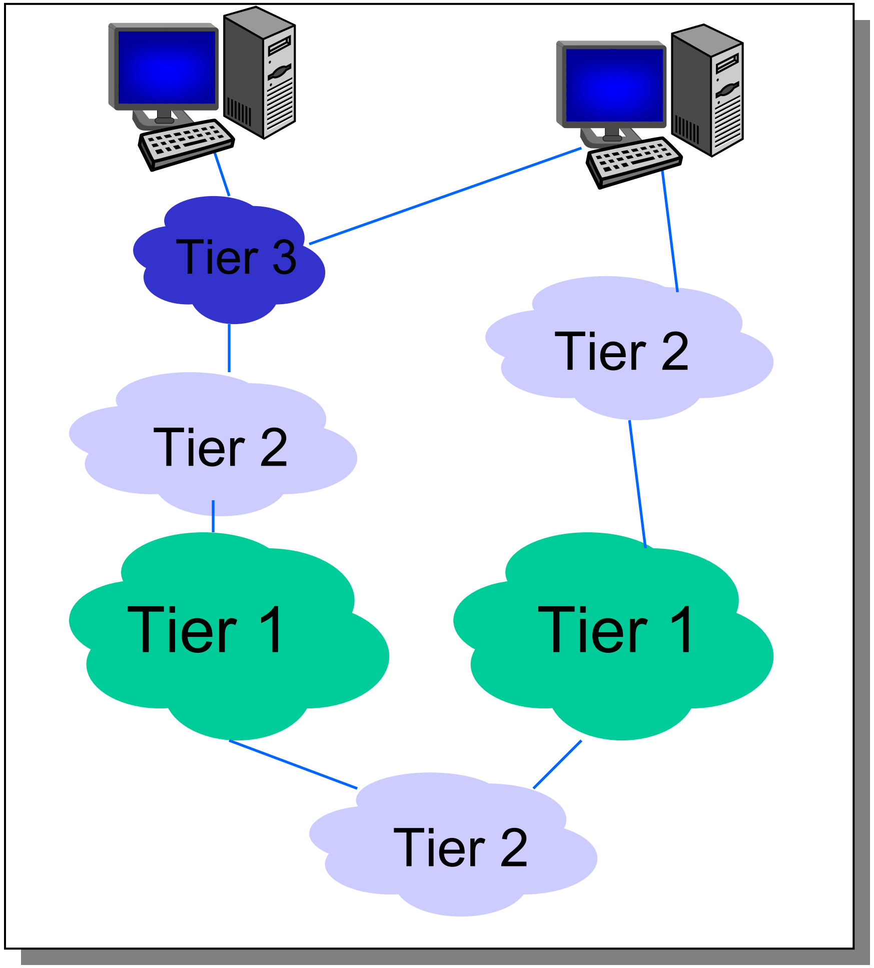
Figures by MIT OpenCourseWare.

- Intra-domain routing inside an AS
- Iner-domain routing between ASes

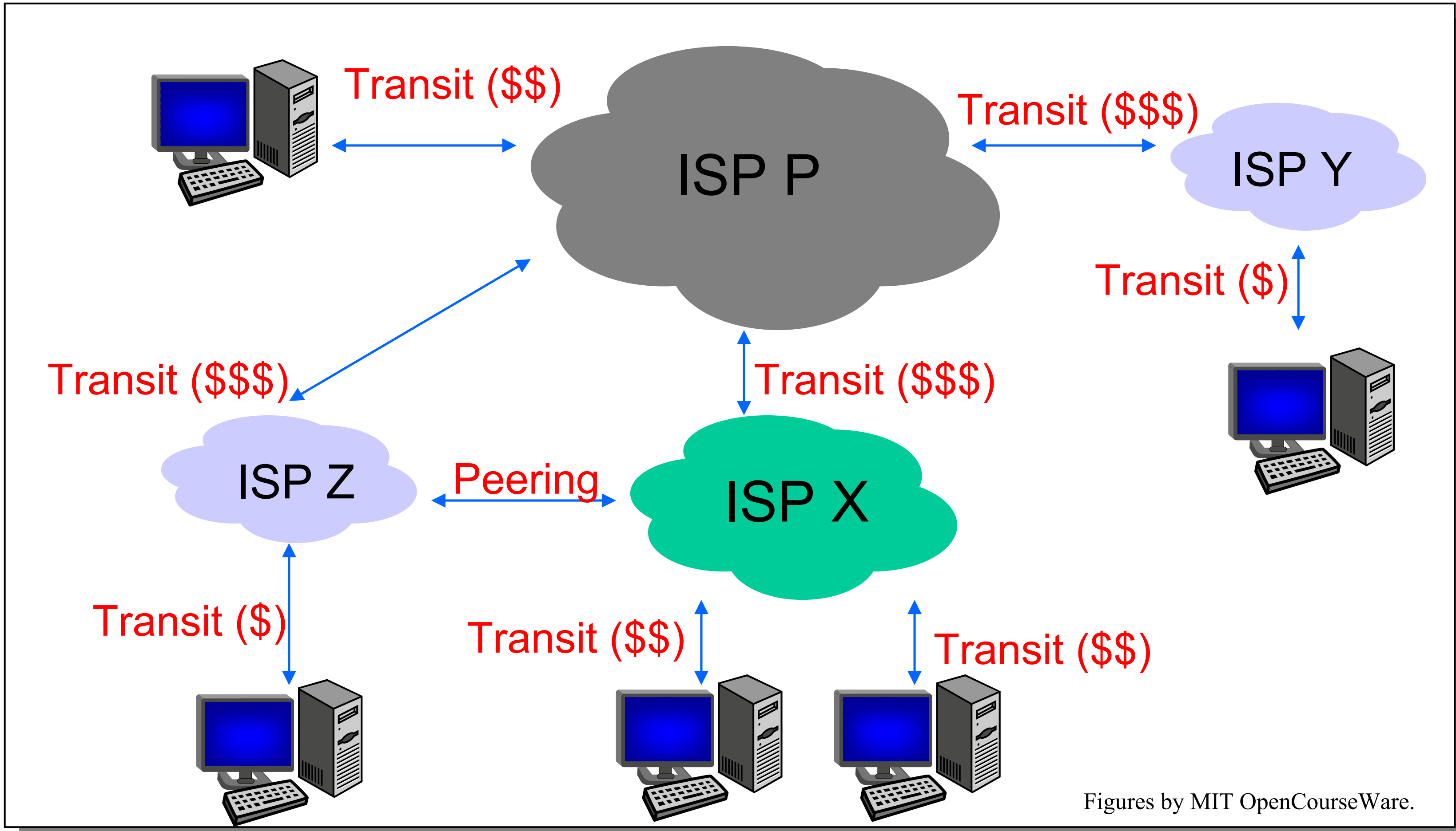
A Logical View of the Internet

- Internet connectivity is provided by commercial entities called ISPs, who compete for profit yet have to cooperate to provide connectivity
  - Each ISP has its own AS (sometimes multiple ASes)
- Not all ISPs are created equal
  - Tier 1 ISP
    - "Default-free" global reachability info
  - Tier 2 ISP
    - Regional or country-wide
  - Tier 3 ISP
    - Local

Figures by MIT OpenCourseWare.



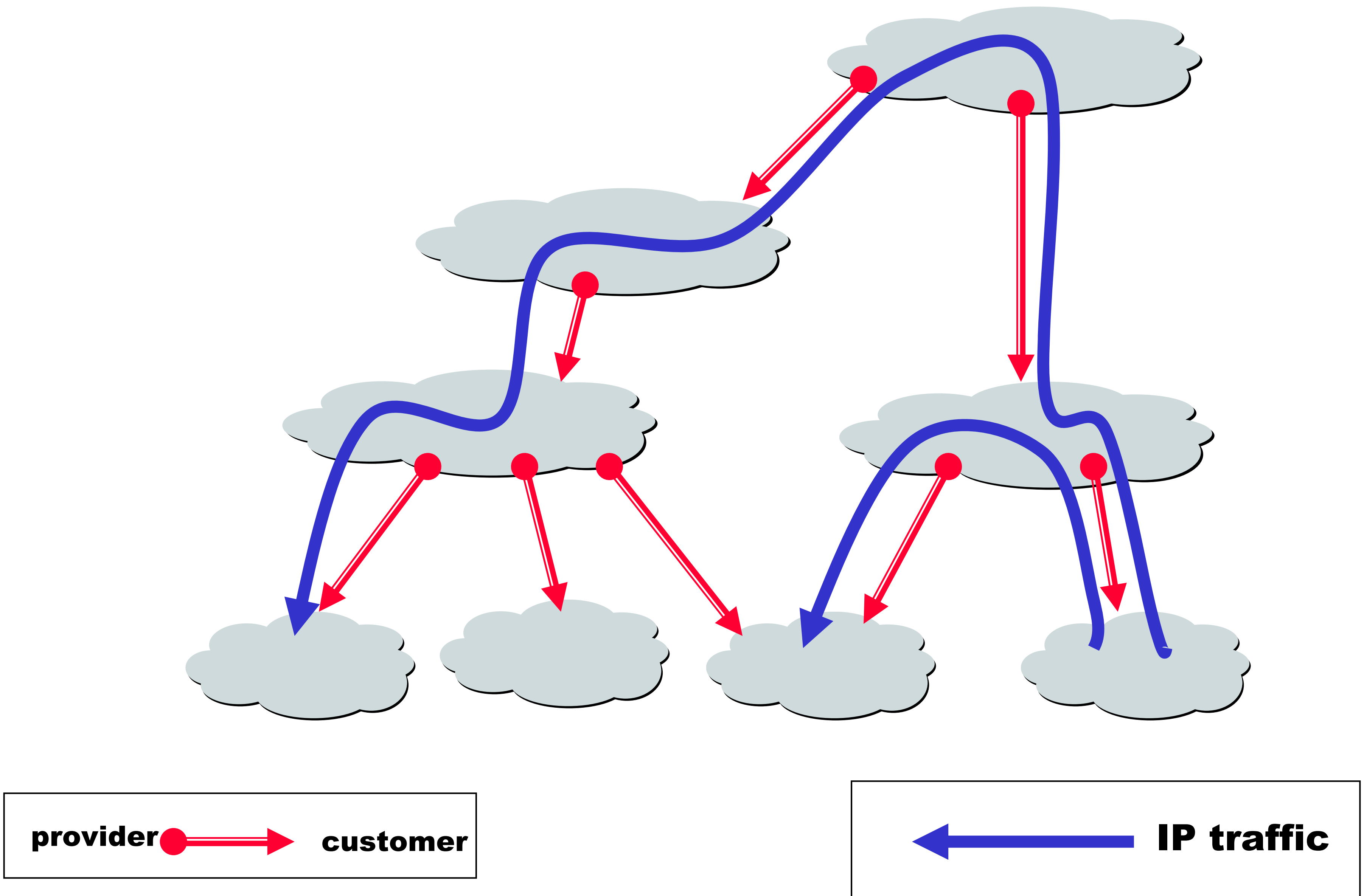
Inter-AS Relationship: Transit vs. Peering



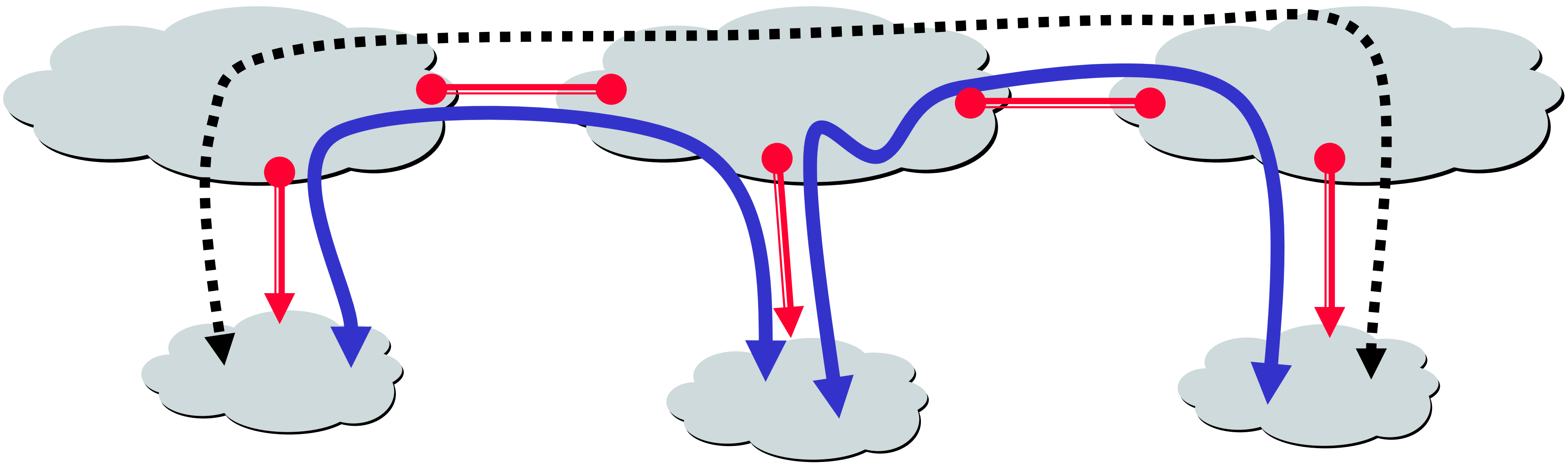
Policy Impact on Routing

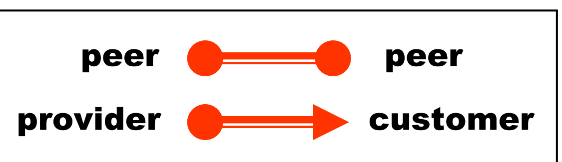
- AS relationships
  - Customer-provider
  - Peers
- Want "Valley-free" routes
  - Number links as (+1, 0, -1) for provider, peer and customer links
  - In any path, you should only see sequence of +1, followed by at most one 0, followed by sequence of -1

Customer-Provider Hierarchy



The Peering Relationship



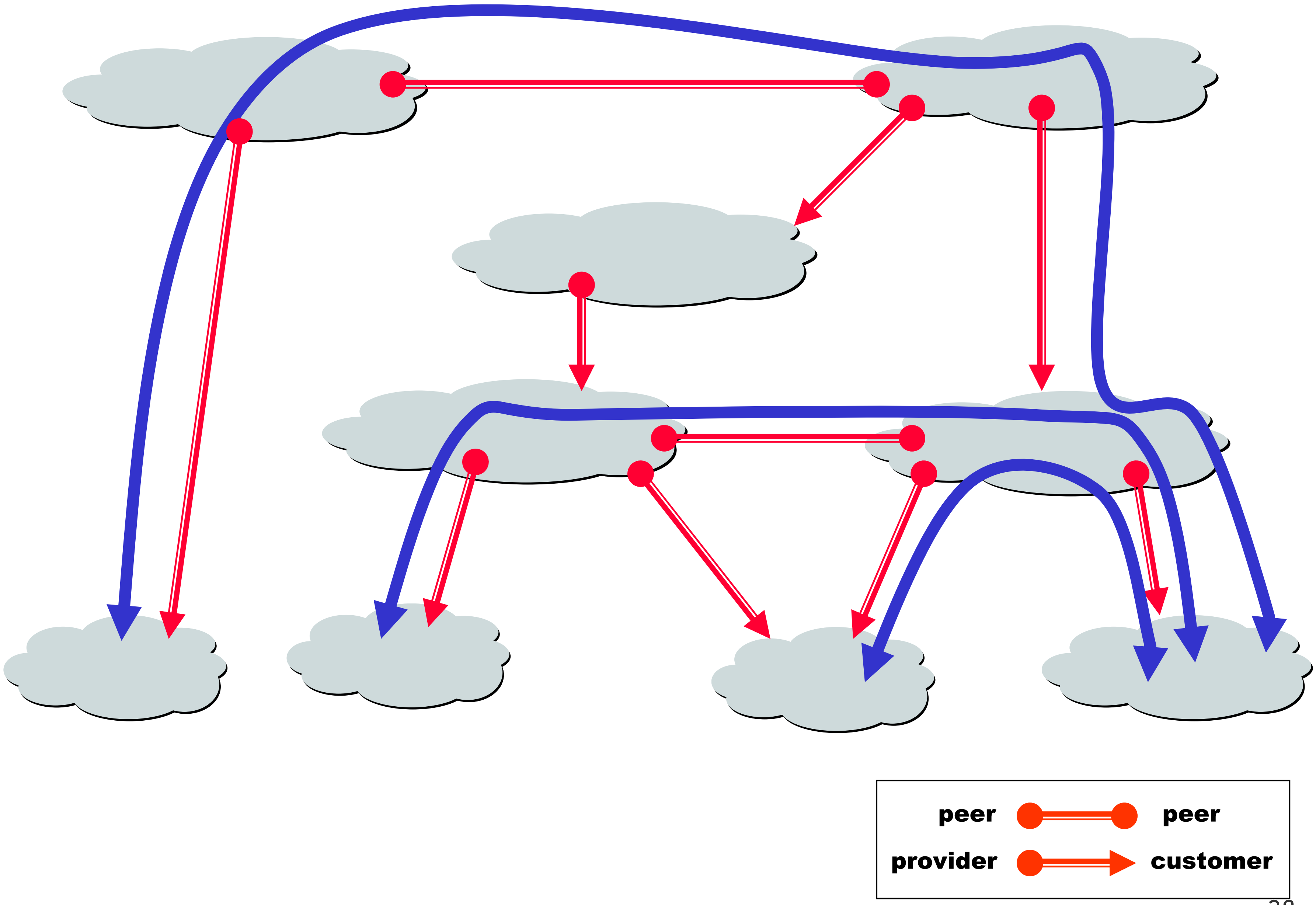


Peers provide transit between their respective customers

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

Peers (often) do not exchange $$$

Peering Provides Shortcuts



Policy-Based Routing

- Policies are used to force customer-provider-peer relationships, backup links, load balancing, …
- Can't use shortest path routing
  - No universal metric – policy-based decisions
  - Main characteristic of shortest path does not hold ($i \rightarrow x \rightarrow j$ is shortest route, then $x \rightarrow j$ is shortest route)
- Problems with distance-vector:
  - Bellman-Ford algorithm may not converge, and may loop
- Problems with link state:
  - Metric used by different routers are not the same → loops
  - LS database too large – entire Internet
  - May expose policies to other AS's

BGP: Distance Vector with Path

- Each routing update carries the entire path
  - e.g.,: destination 18.26/16 is reachable using {AS1, AS3, AS11}
- When AS receives a routing update
  - Reject routes with loops
    - To detect loops check whether my AS is already in path
- AS remembers loop-free routes
- For each destination, the AS chooses the best route according its policies.
- AS advertises a neighbor routes to a subset of all the destinations, depending on its policy
  - E.g., I might hide from you that I know how to get to destination X, because I don't want to deliver your messages to X
- AS advertises to neighbors only those routes that it uses
  - Ensures that if i→x→j is the used route, then x→j is the used route
  - *What happens if an AS advertises routes that it doesn't use?*
- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops

Implementing Customer/Provider and Peer/Peer relationships using BGP

- BGP provides capability for enforcing various policies
- Policies are <u>**not**</u> part of BGP: they are provided to BGP as configuration information
- BGP enforces policies by
  1. choosing paths from multiple alternatives (importing routers)
  2. controlling advertisement to other AS's (exporting routes)

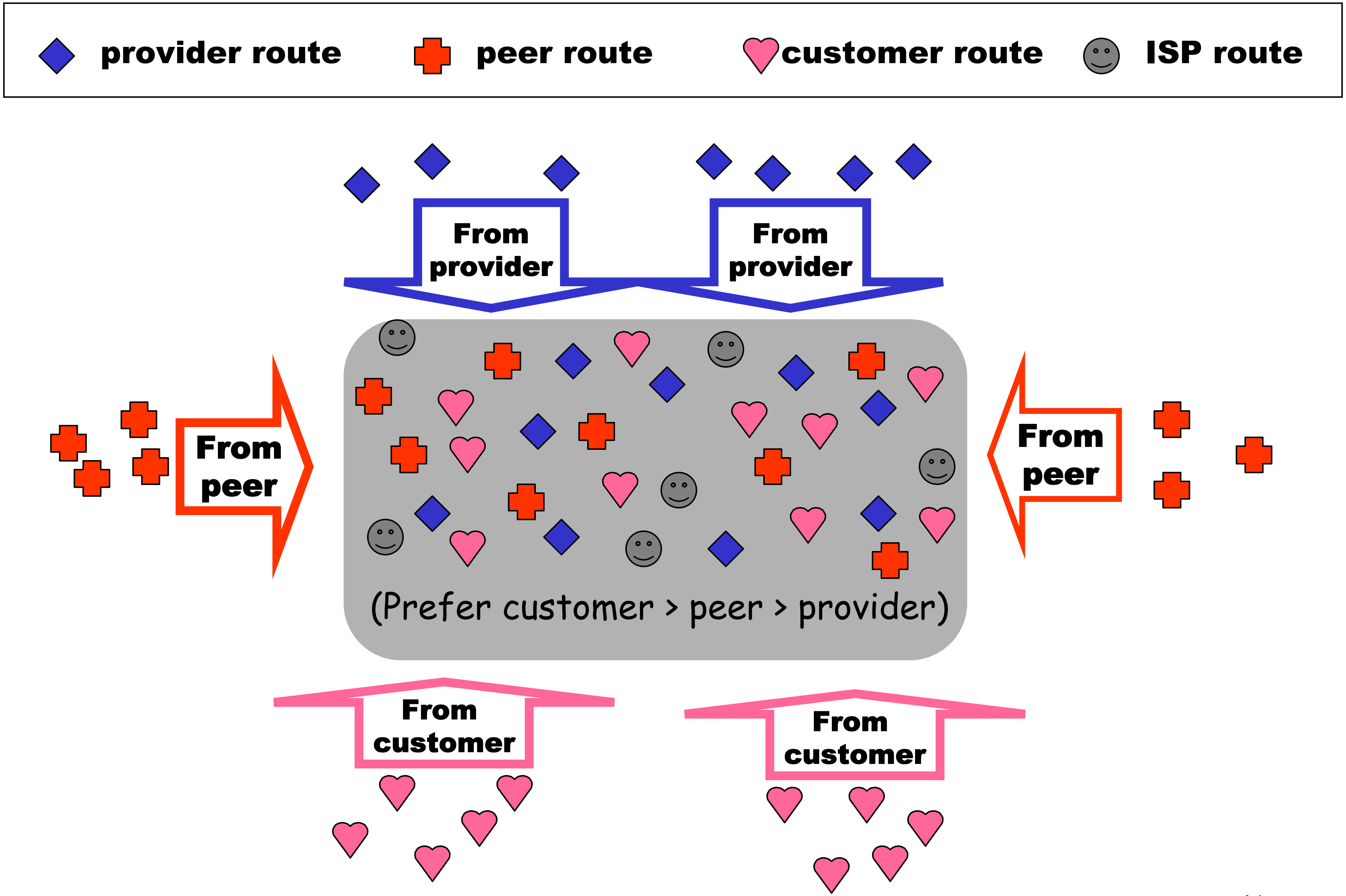
Importing Routes

- Based on route attributes
  - First, Prefer customer > peer > provider
  - Then, Shortest AS PATH length
  - Then, look at other route attributes

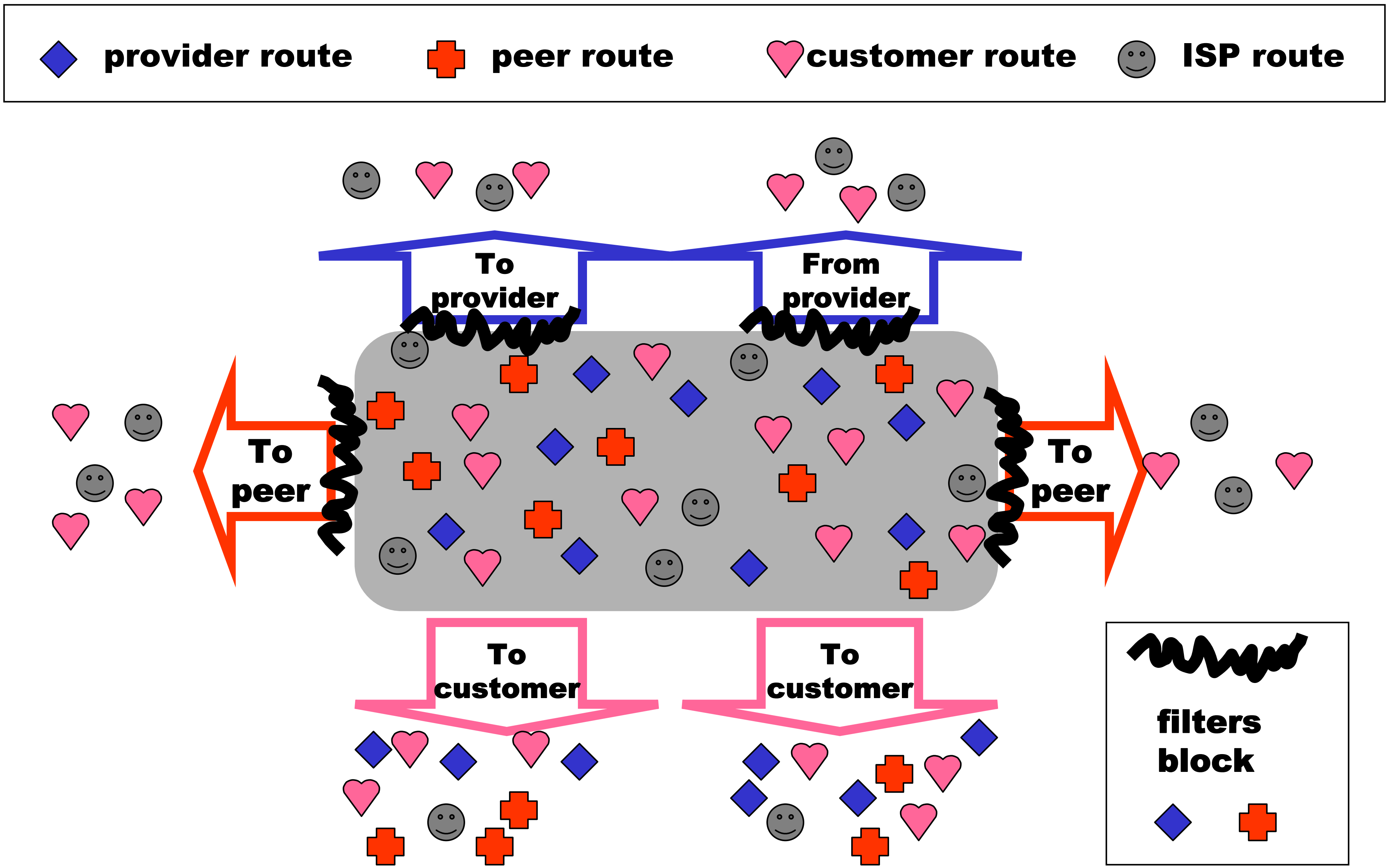
Exporting Routes

- When an AS exports a route, others can use the AS to forward packets along that route
- Rules:
  - Export customers routes to everyone
    - why?
  - Export routes to your own addresses to everyone
    - Why?
  - Don't export routes advertised to you by your provider (may advertise them to customers)
    - Why?
  - Don't export routes advertised to you by your peer (may advertise them to customers)
    - Why?

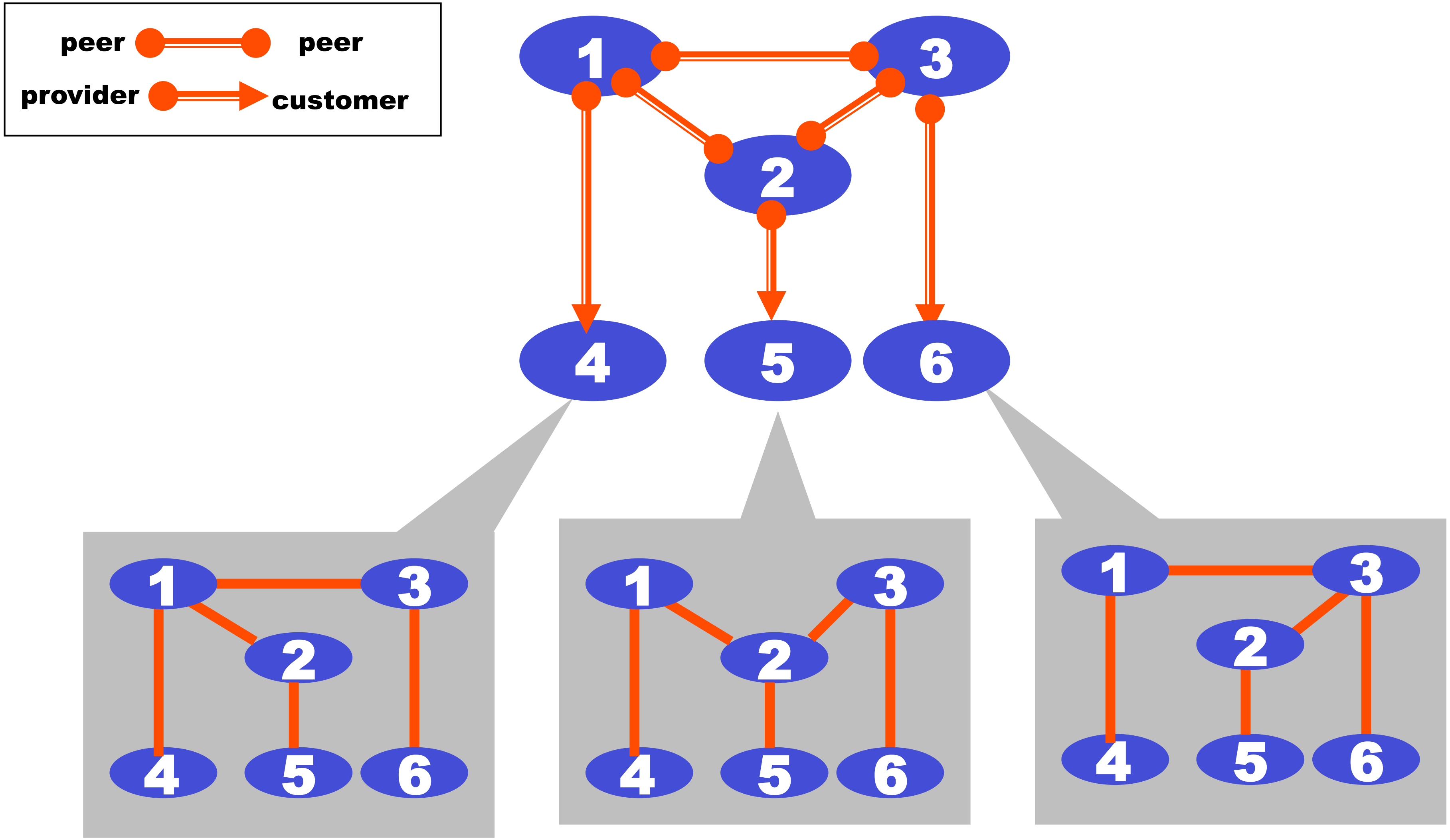
Import Routes



Export Routes



AS Graphs Depend on Point of View



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)

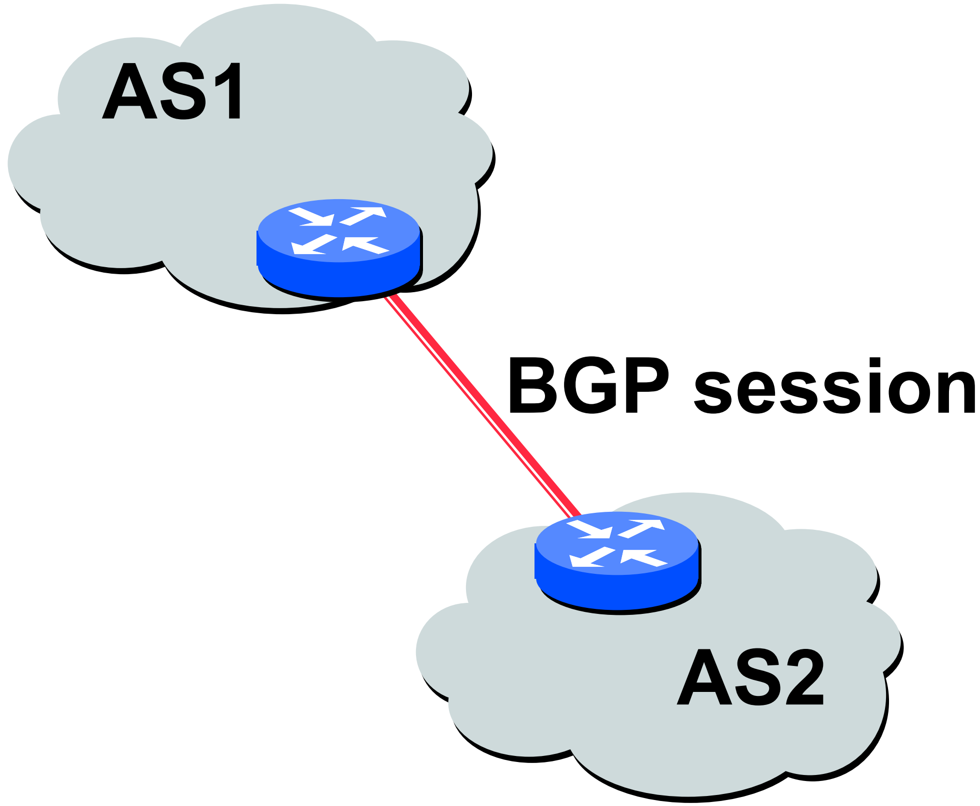
Establish session on TCP port 179

↓

Exchange all active routes

↓

Exchange incremental updates



While connection is ALIVE exchange route UPDATE messages

Four Types of BGP Messages

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

announcement

=

prefix + attributes values

BGP Attributes

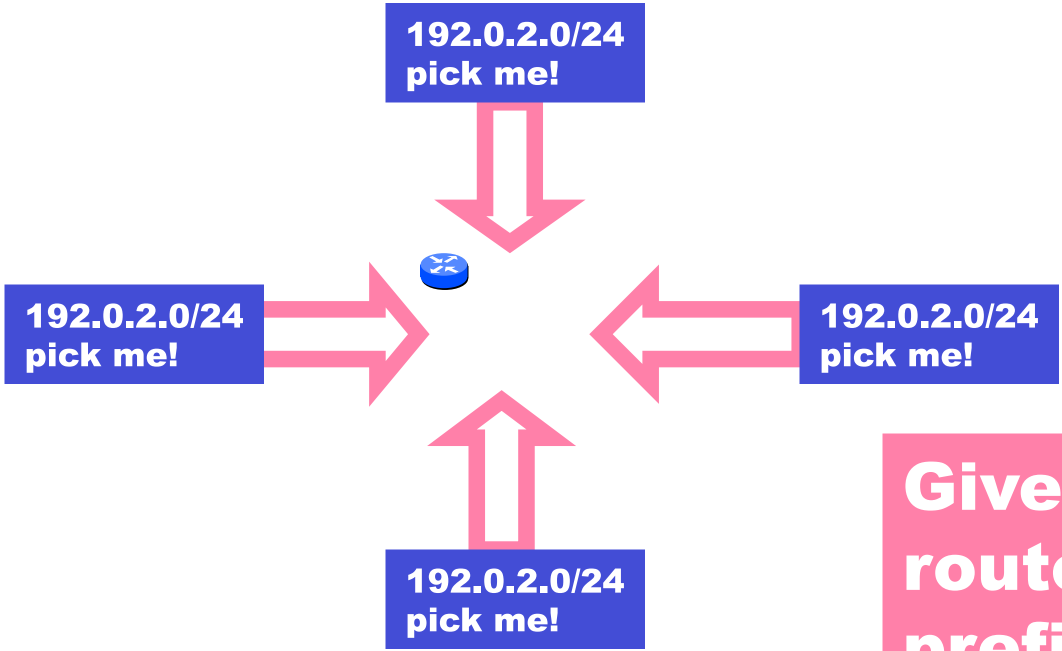
Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / 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

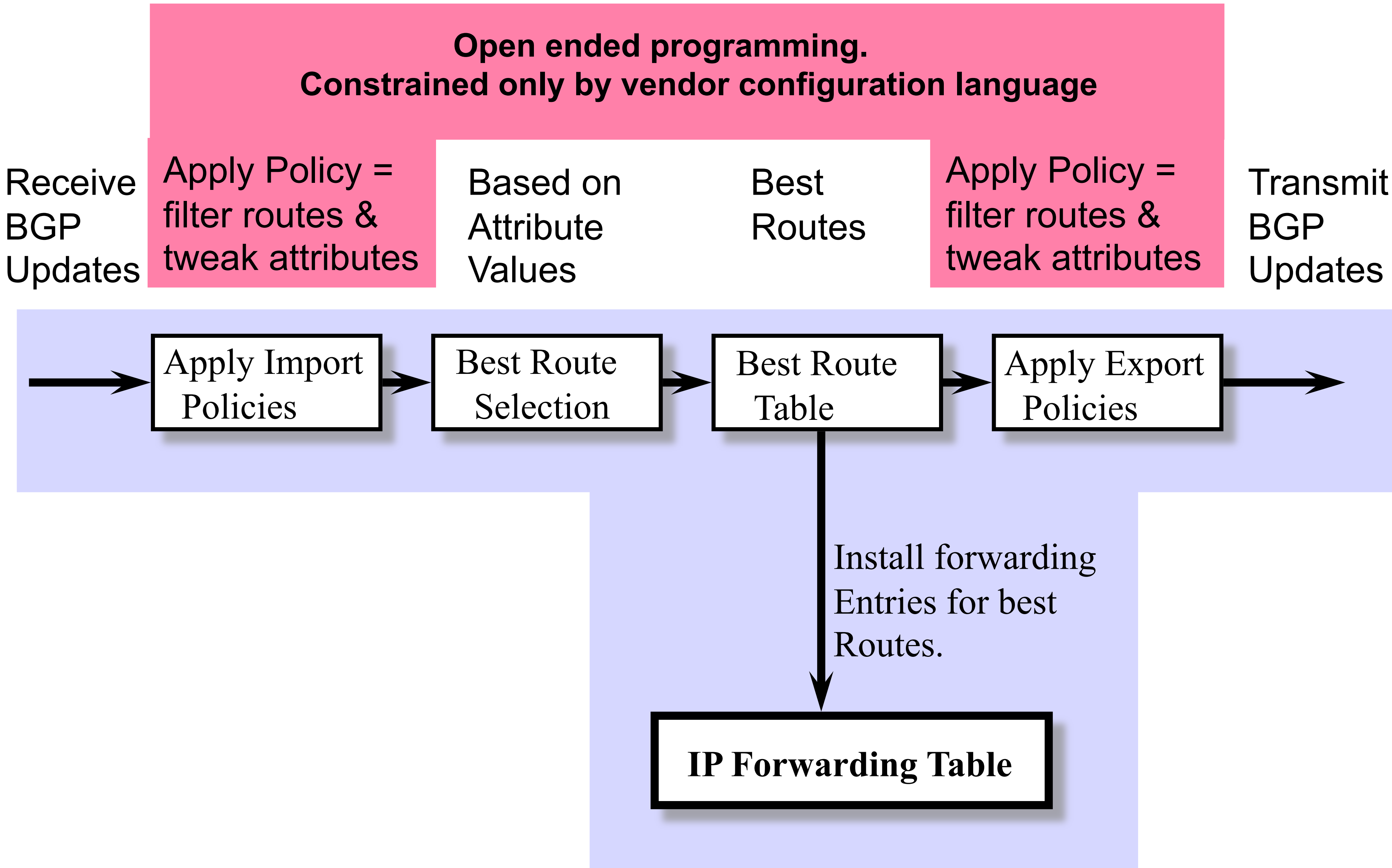
Attributes are Used to Select Best Routes



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

(Note: it could reject them all!)

BGP Route Processing

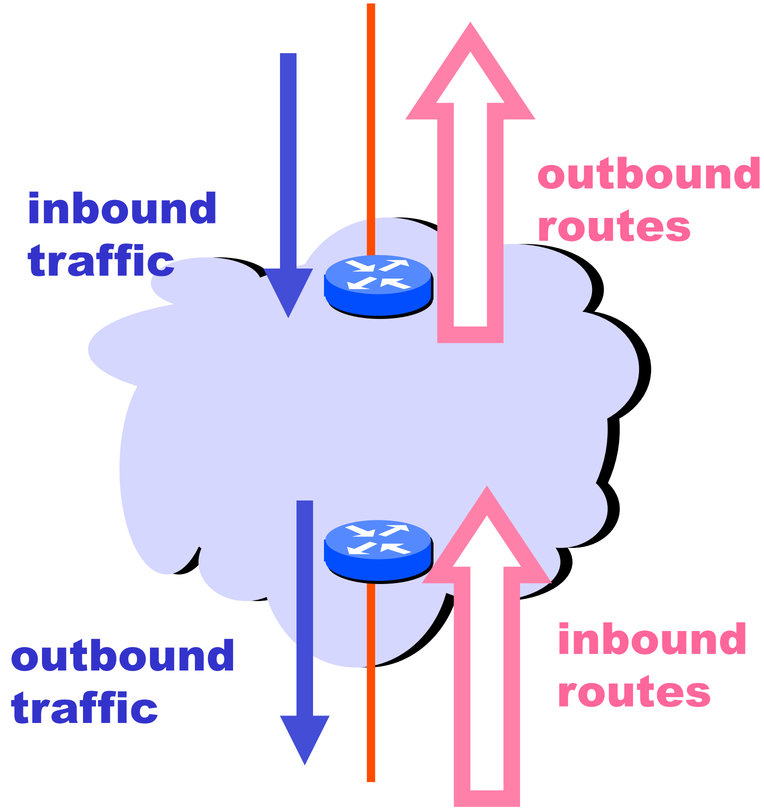


Route Selection Summary

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

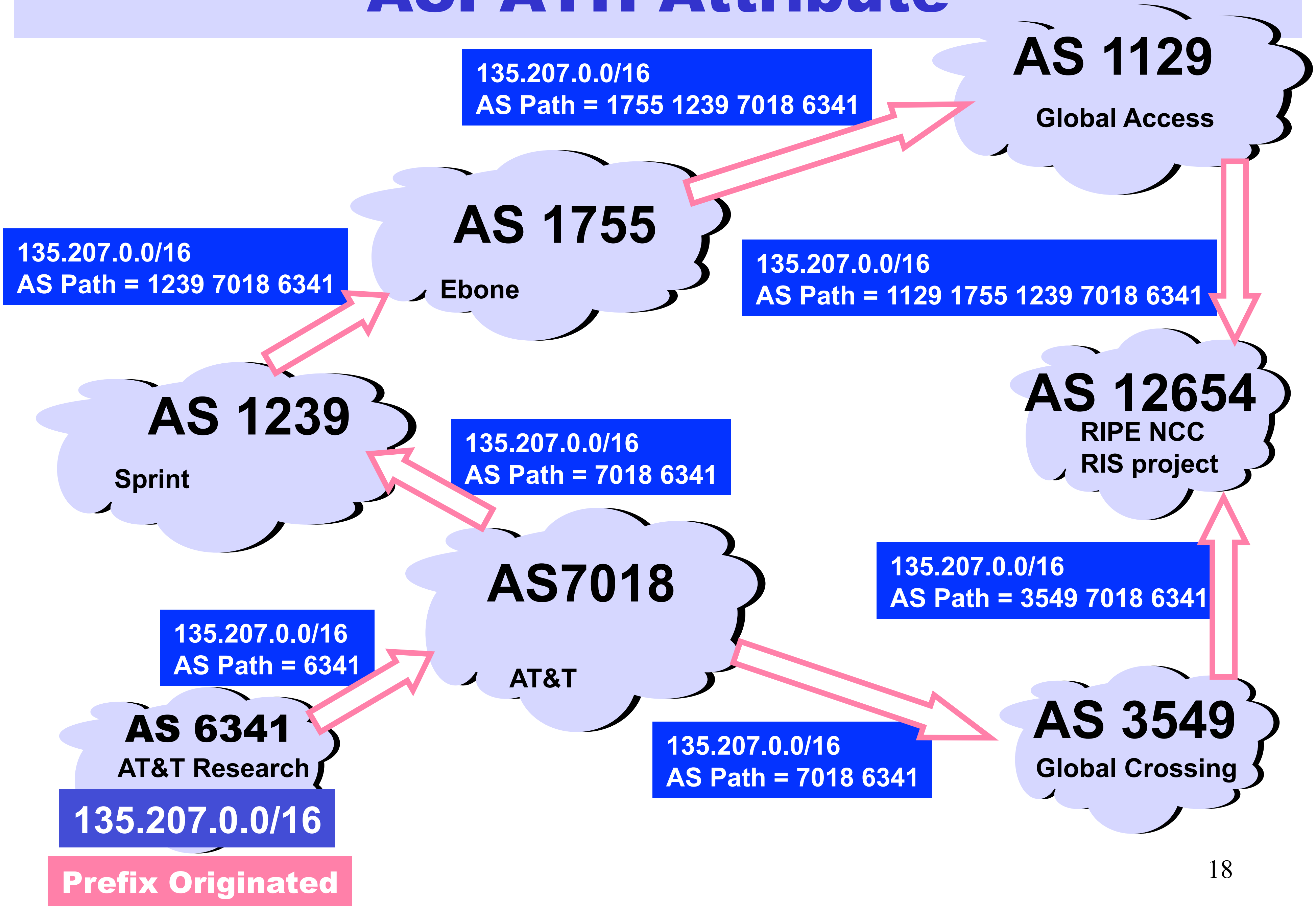
Tweak Tweak Tweak

- 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



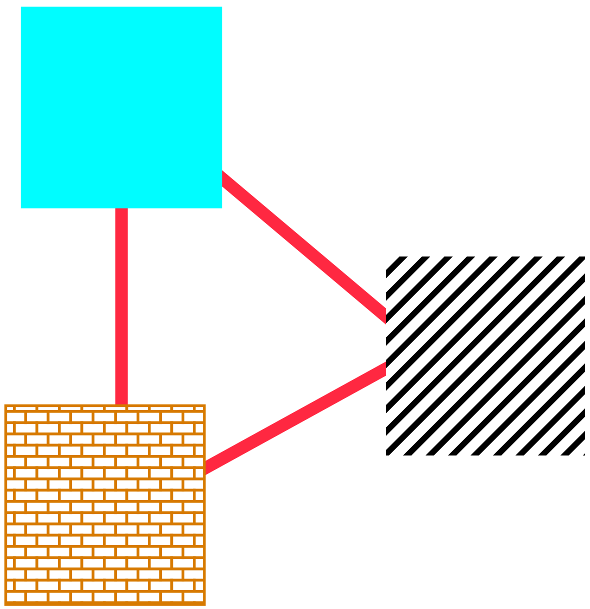
In general, an AS has more control over outbound traffic

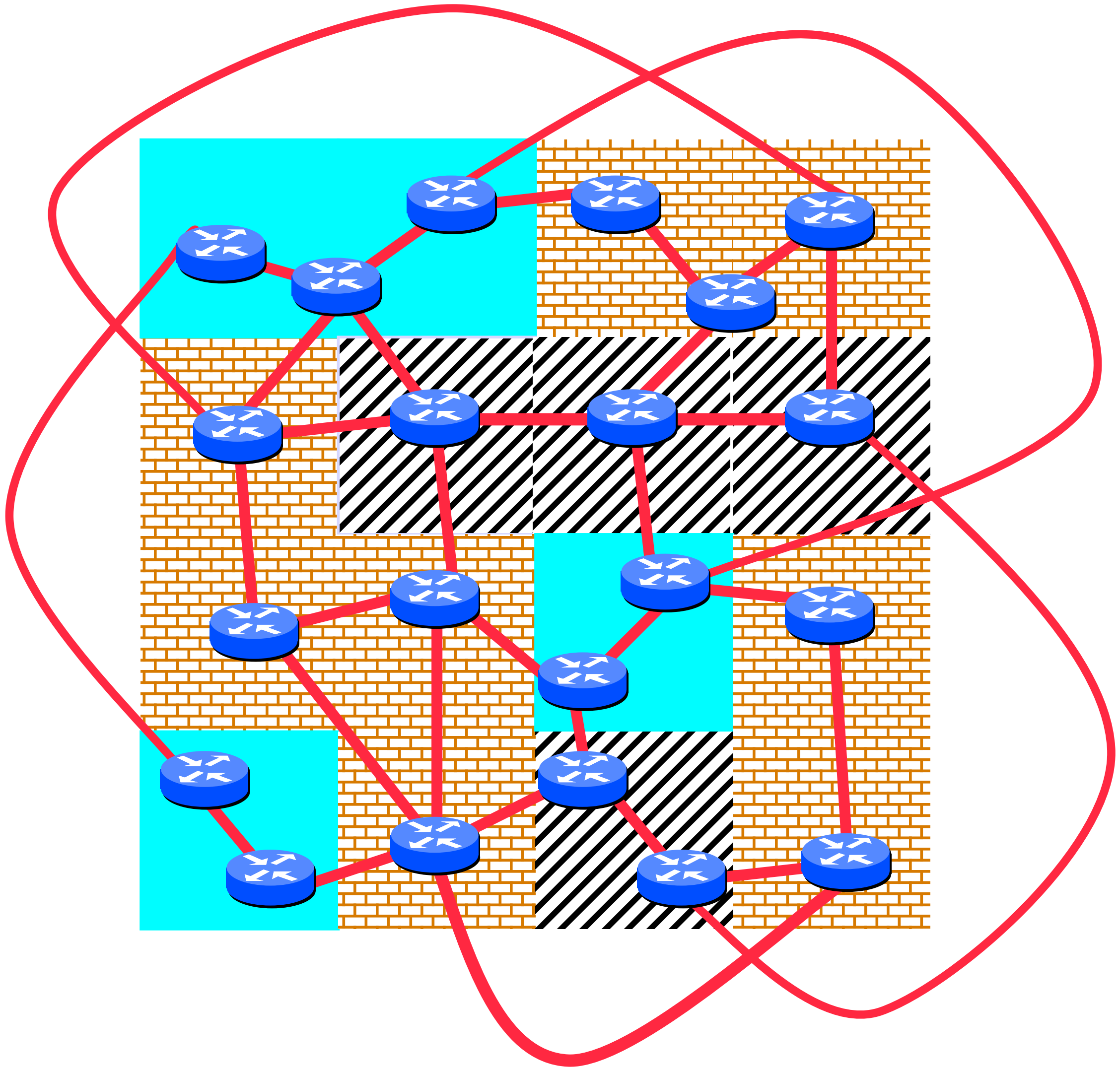
ASPATH Attribute



AS Graphs Do Not Show Topology!

BGP was designed to throw away information!





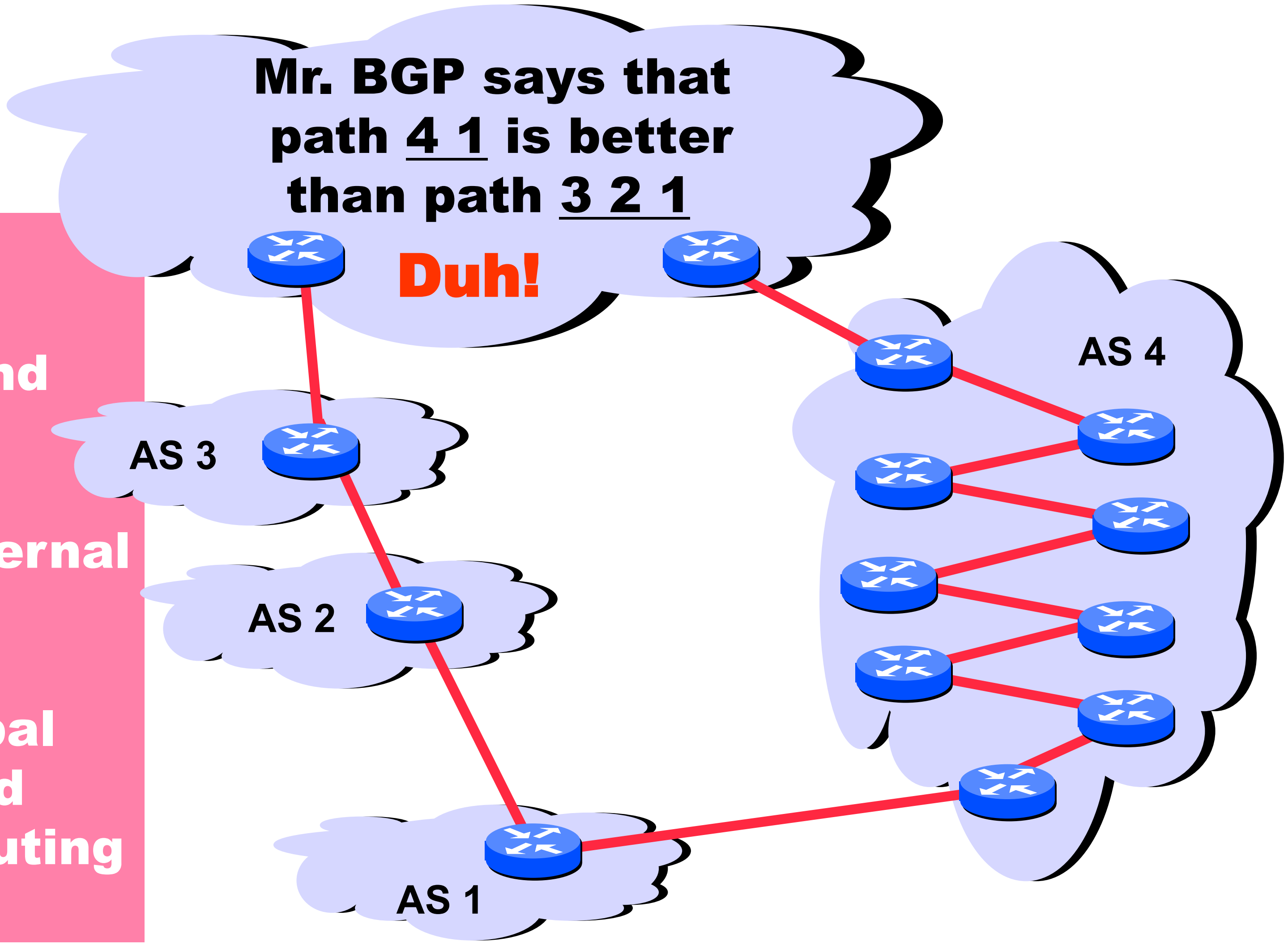
The AS graph may look like this.

Reality may be closer to this...

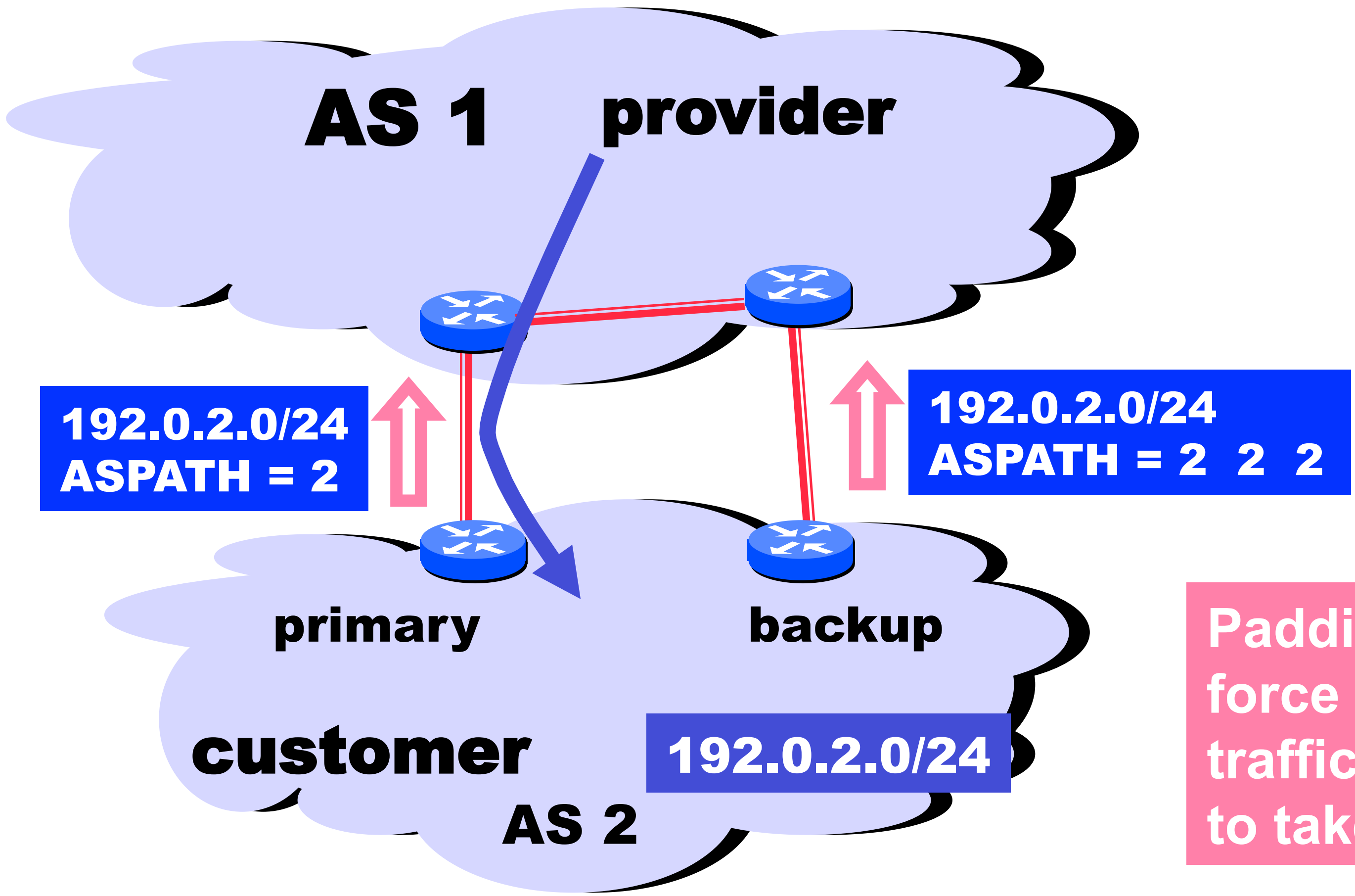
Shorter Doesn't Always Mean Shorter

In fairness: could you do this "right" and still scale?

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

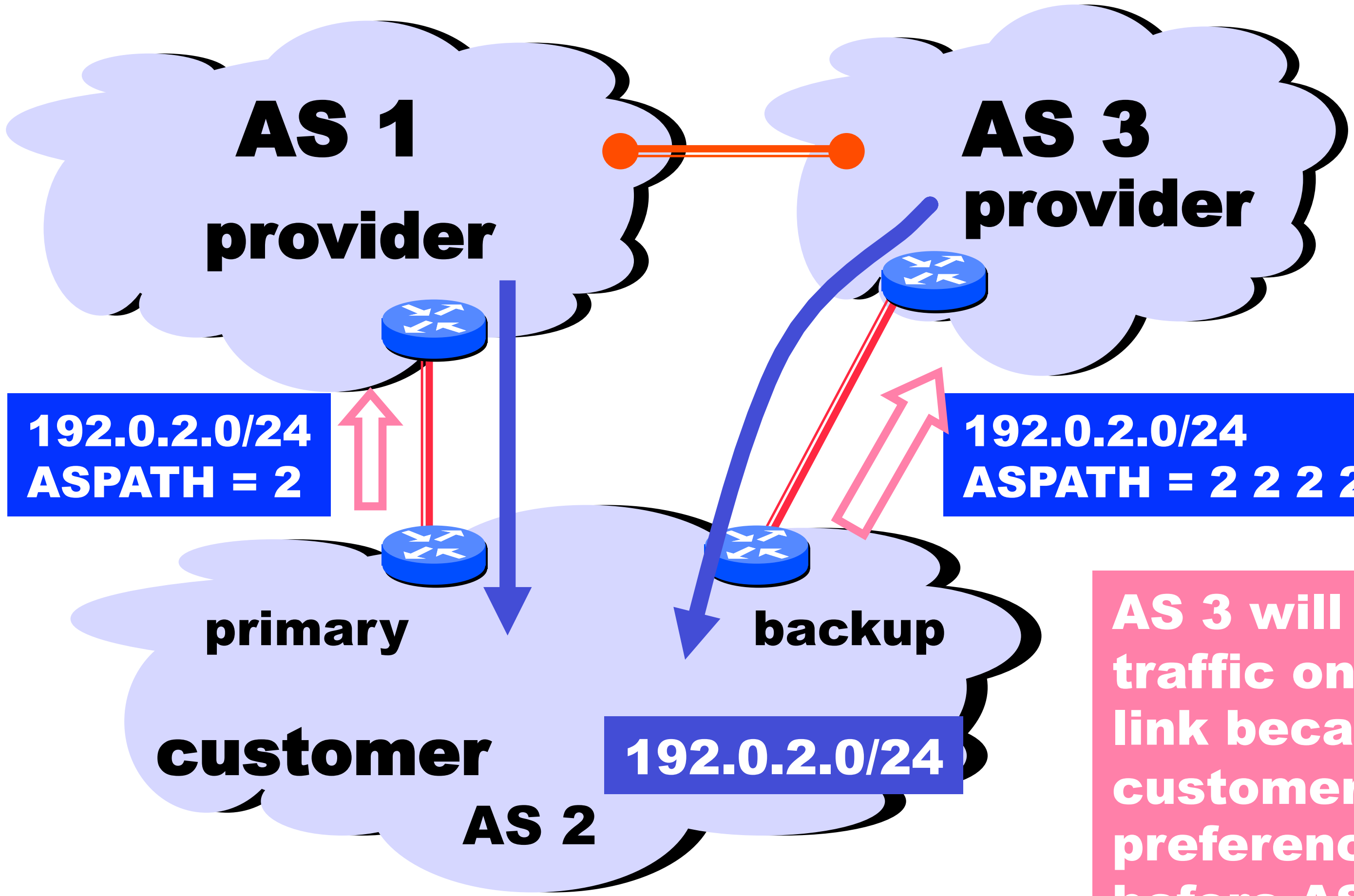


Shedding Inbound Traffic with ASPATH Padding Hack



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

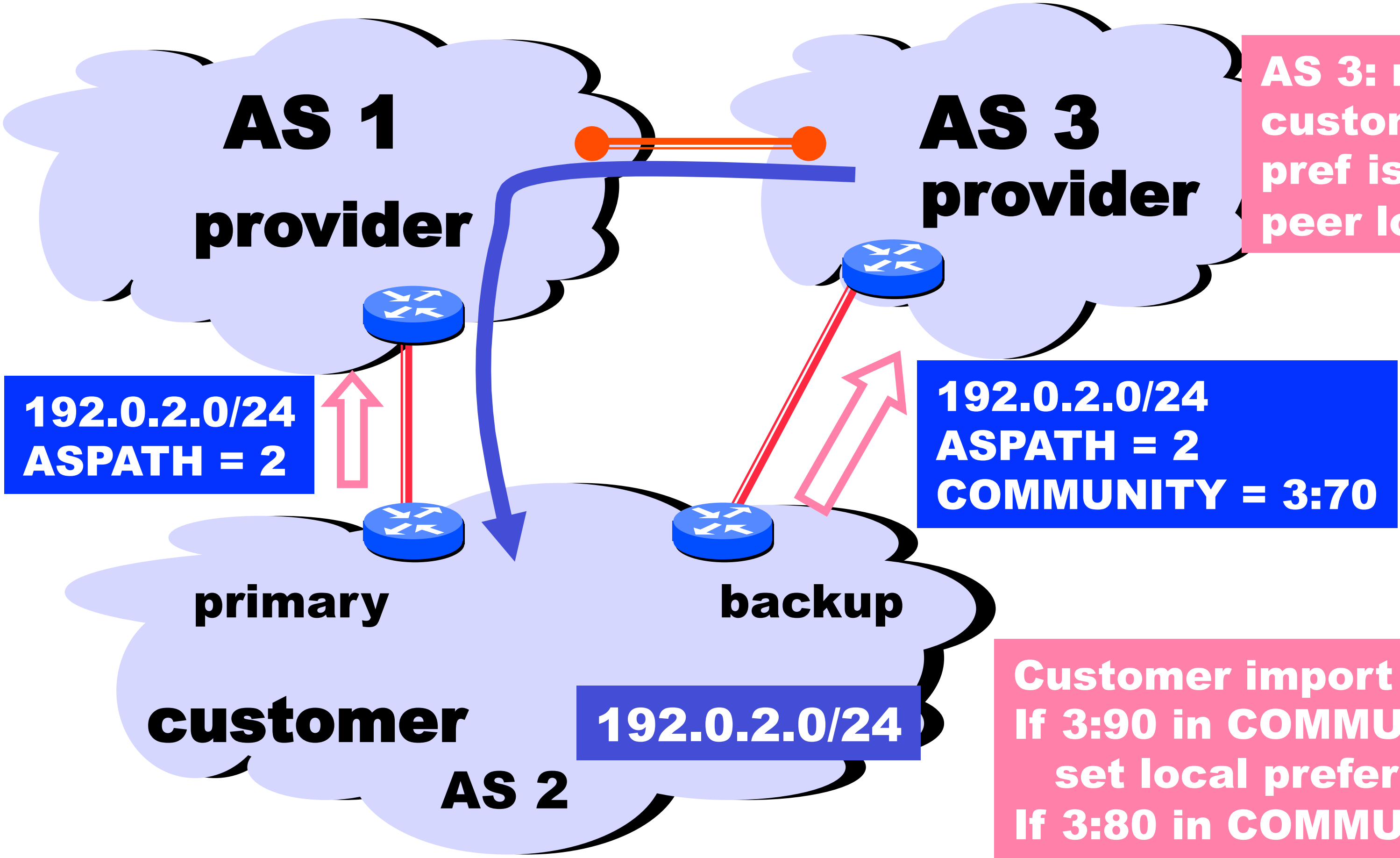
Padding May Not Shut Off All Traffic



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

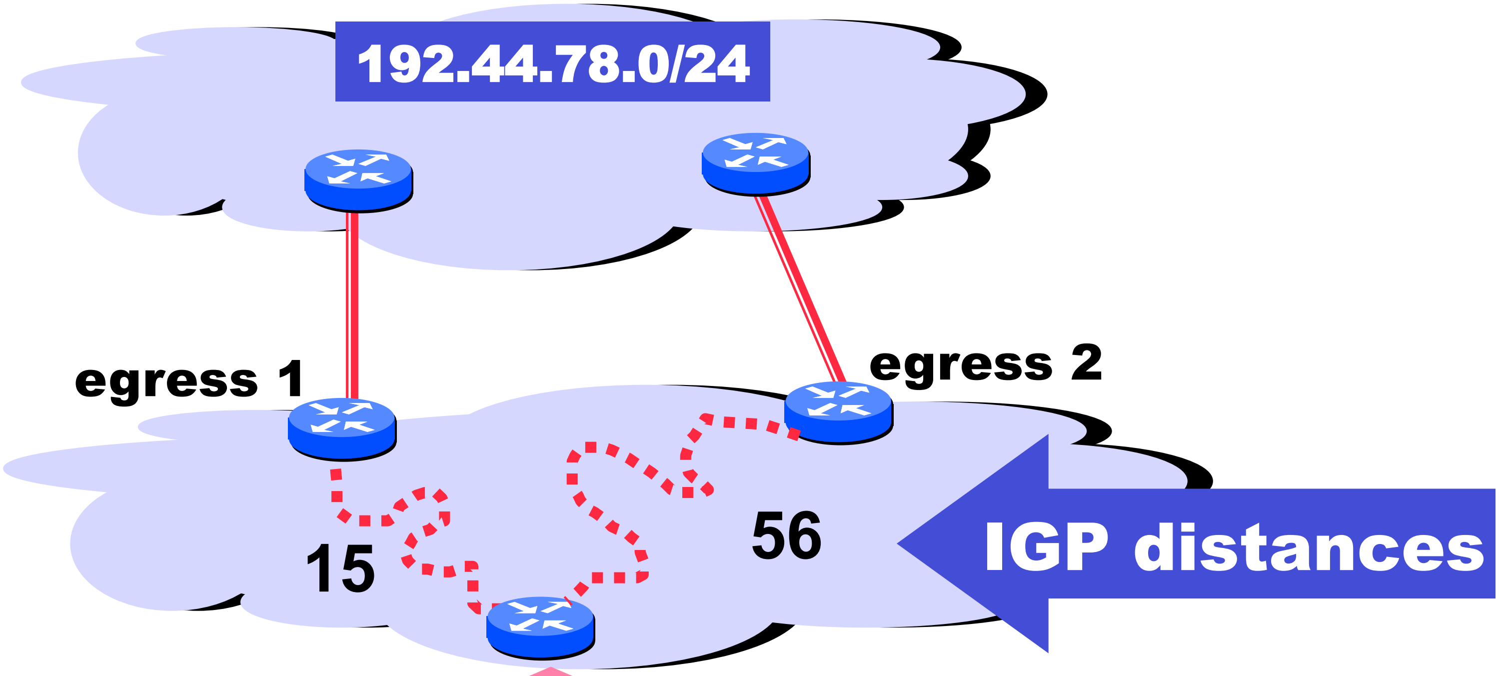
COMMUNITY Attribute to the Rescue!



AS 3: normal customer local pref is 100,
peer local pref is 90

Customer import policy at AS 3:
If 3:90 in COMMUNITY then
  set local preference to 90
If 3:80 in COMMUNITY then
  set local preference to 80
If 3:70 in COMMUNITY then
  set local preference to 70

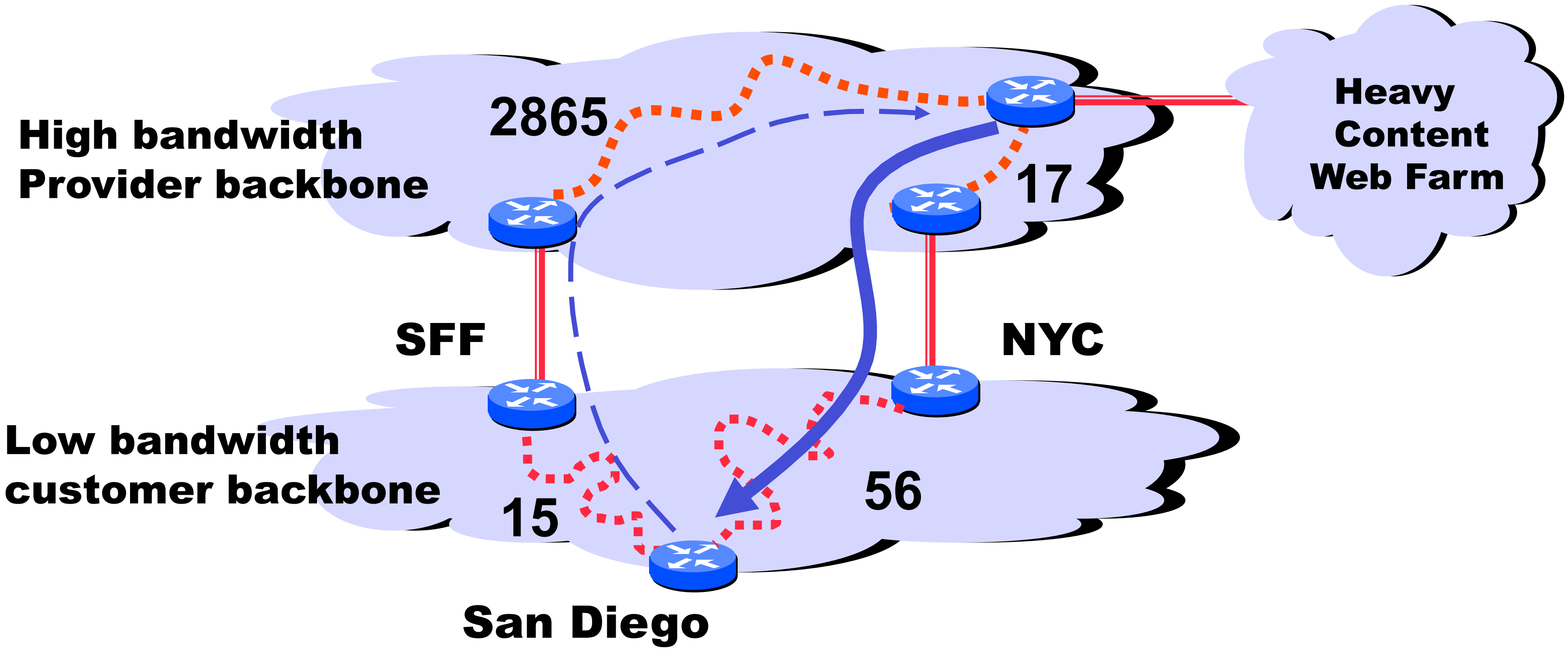
Hot Potato Routing: Go for the Closest Egress Point



This Router has two BGP routes to 192.44.78.0/24.

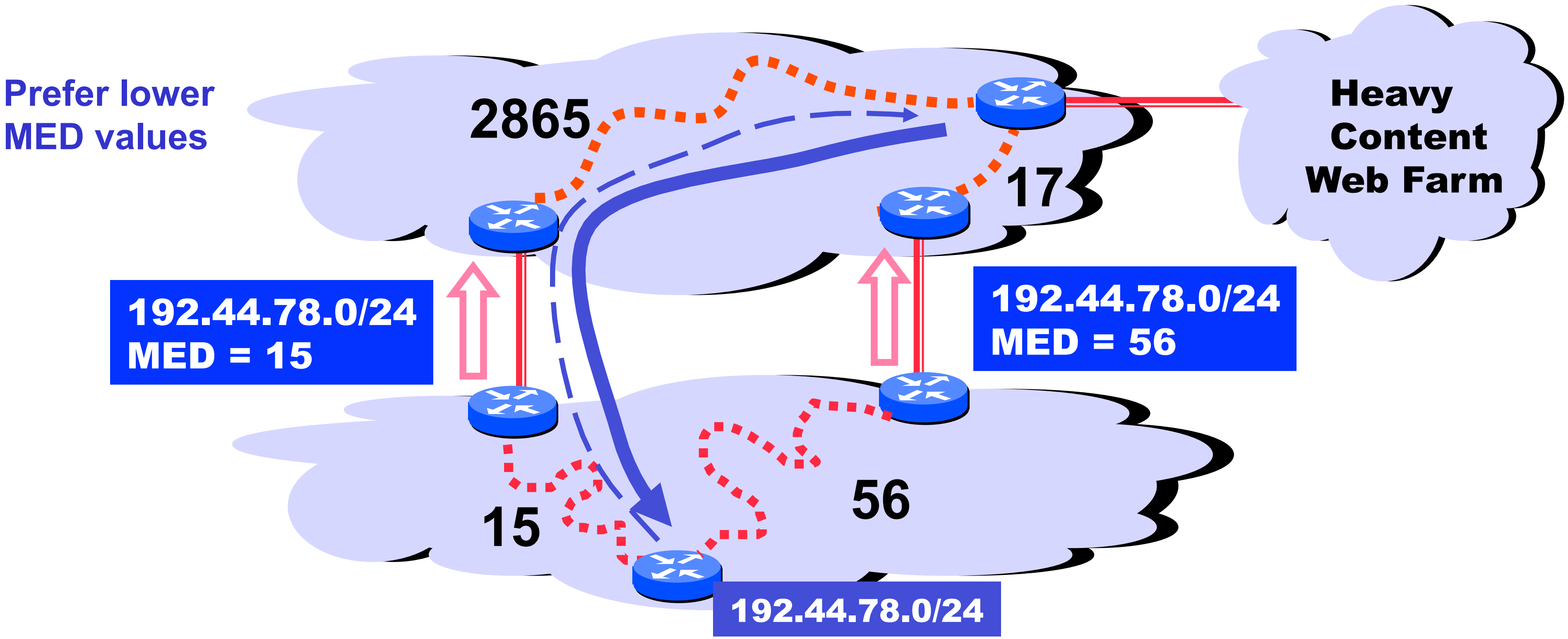
Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!

Getting Burned by the Hot Potato



Many customers want their provider to carry the bits!

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



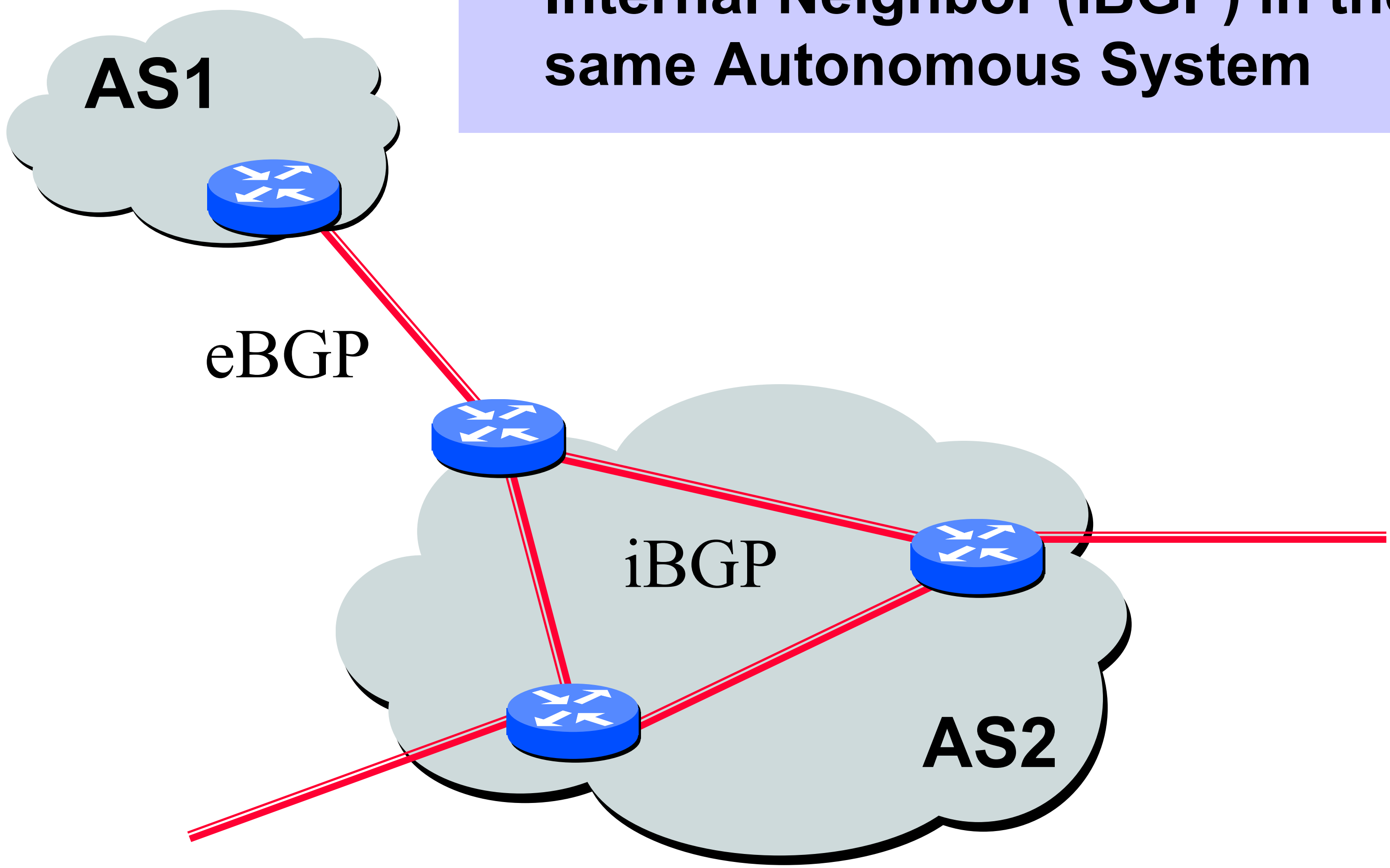
This means that MEDs must be considered BEFORE IGP distance!

Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

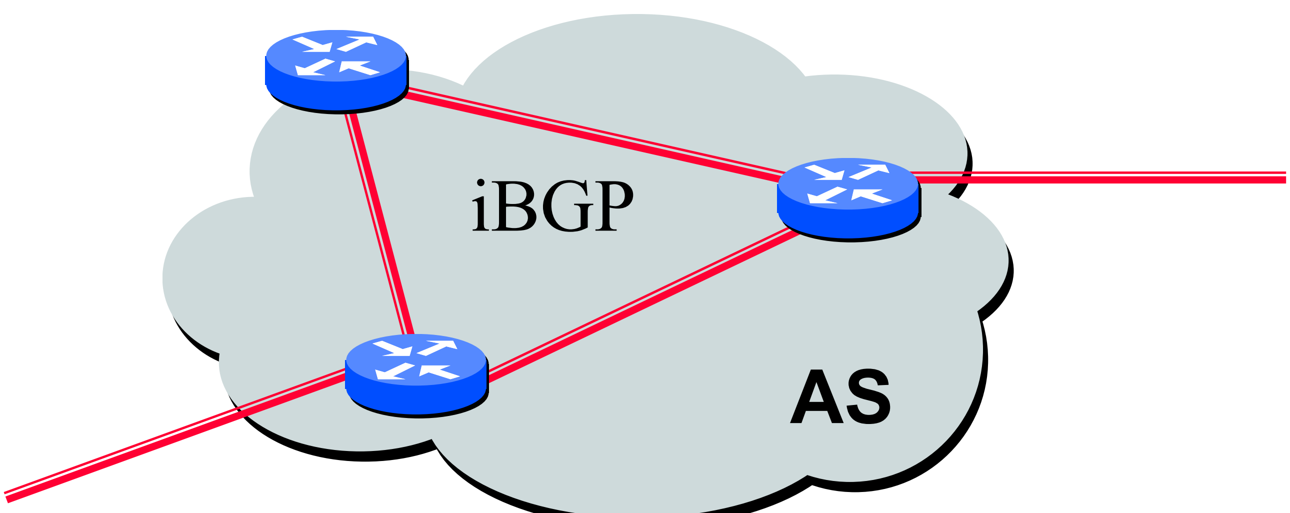
Two Types of BGP Neighbor Relationships

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



iBGP

- AS has more than one router participating in eBGP
- iBGP is run between BGP routers in the same AS to allow all of them to obtain a complete and consistent view of external routes



Internal BGP (iBGP)

- Same messages as eBGP
- Different rules about re-advertising prefixes:
  - Prefix learned from eBGP can be advertised to iBGP neighbor and vice-versa, but
  - Prefix learned from one iBGP neighbor cannot be advertised to another iBGP neighbor
    - Reason: no AS PATH within the same AS and thus danger of looping.

We learned

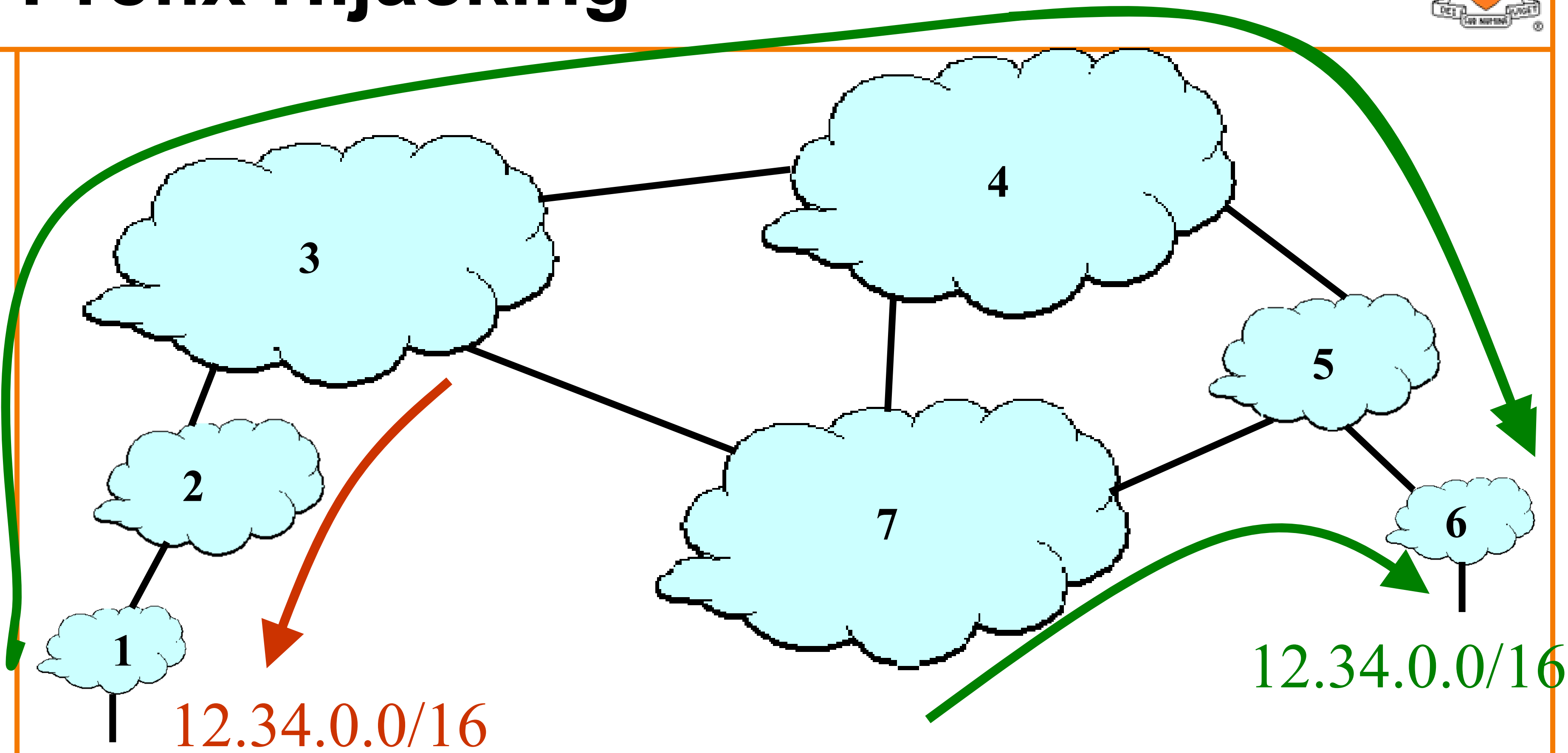
- Inter-domain routing uses policy
- As a result, routing is not a simple optimization of a single number which can be done using shortest path algorithms
- BGP is designed to route based on policies

Validity of the routing information: Origin authentication

IP Address Ownership and Hijacking

- IP address block assignment
  - Regional Internet Registries (ARIN, RIPE, APNIC)
  - Internet Service Providers
- Proper origination of a prefix into BGP
  - By the AS who owns the prefix
  - … or, by its upstream provider(s) in its behalf
- However, what's to stop someone else?
  - Prefix hijacking: another AS originates the prefix
  - BGP does not verify that the AS is authorized
  - Registries of prefix ownership are inaccurate

Prefix Hijacking

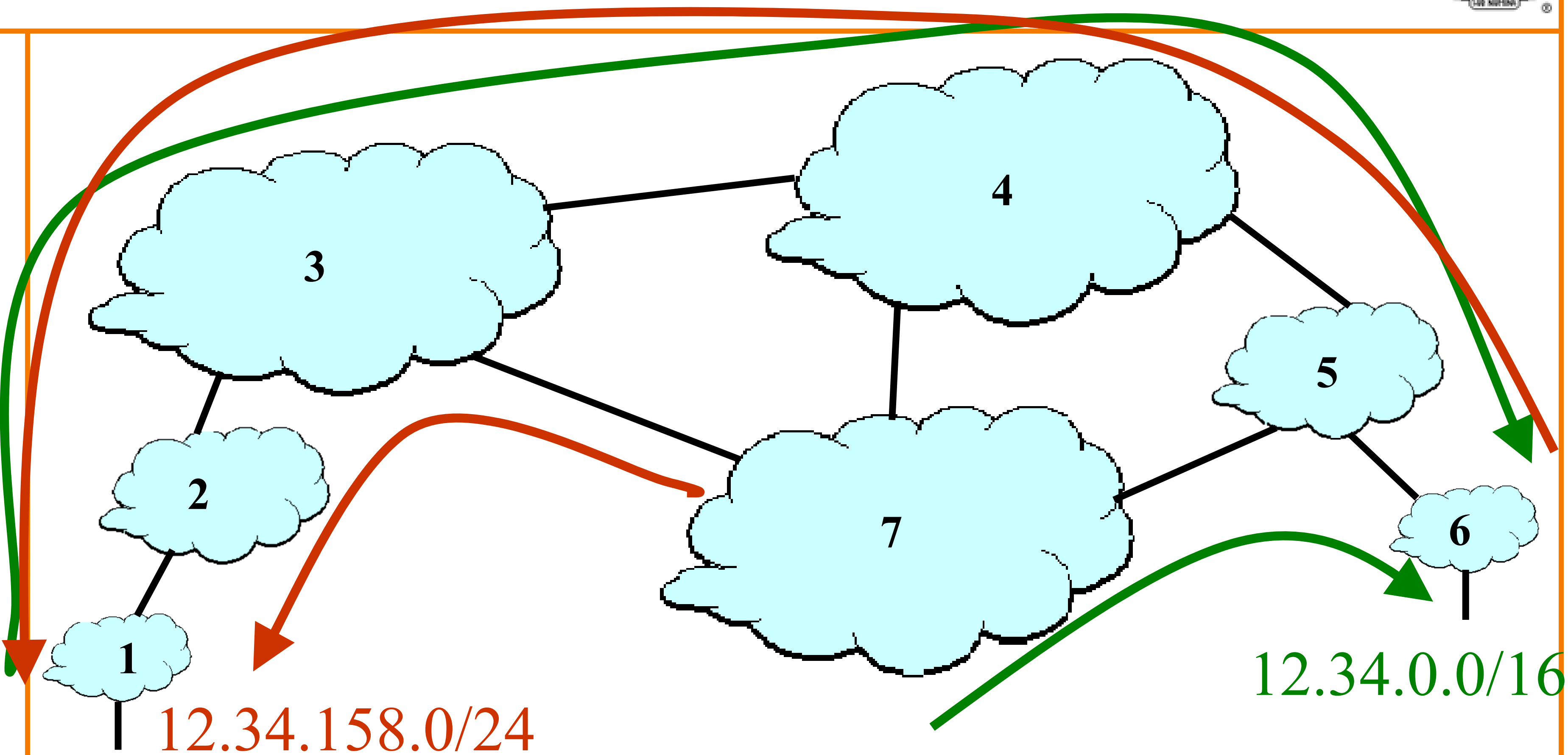


- Consequences for the affected ASes
  - Blackhole: data traffic is discarded
  - Snooping: data traffic is inspected, and then redirected
  - Impersonation: data traffic is sent to bogus destinations

Hijacking is Hard to Debug

- Real origin AS doesn't see the problem
  - Picks its own route
  - Might not even learn the bogus route
- May not cause loss of connectivity
  - E.g., if the bogus AS snoops and redirects
  - … may only cause performance degradation
- Or, loss of connectivity is isolated
  - E.g., only for sources in parts of the Internet
- Diagnosing prefix hijacking
  - Analyzing updates from many vantage points
  - Launching traceroute from many vantage points

Sub-Prefix Hijacking



- Originating a more-specific prefix
  - Every AS picks the bogus route for that prefix
  - Traffic follows the longest matching prefix

How to Hijack a Prefix

- The hijacking AS has
  - Router with eBGP session(s)
  - Configured to originate the prefix
- Getting access to the router
  - Network operator makes configuration mistake
  - Disgruntled operator launches an attack
  - Outsider breaks in to the router and reconfigures
- Getting other ASes to believe bogus route
  - Neighbor ASes not filtering the routes
  - … e.g., by allowing only expected prefixes
  - But, specifying filters on *peering* links is hard