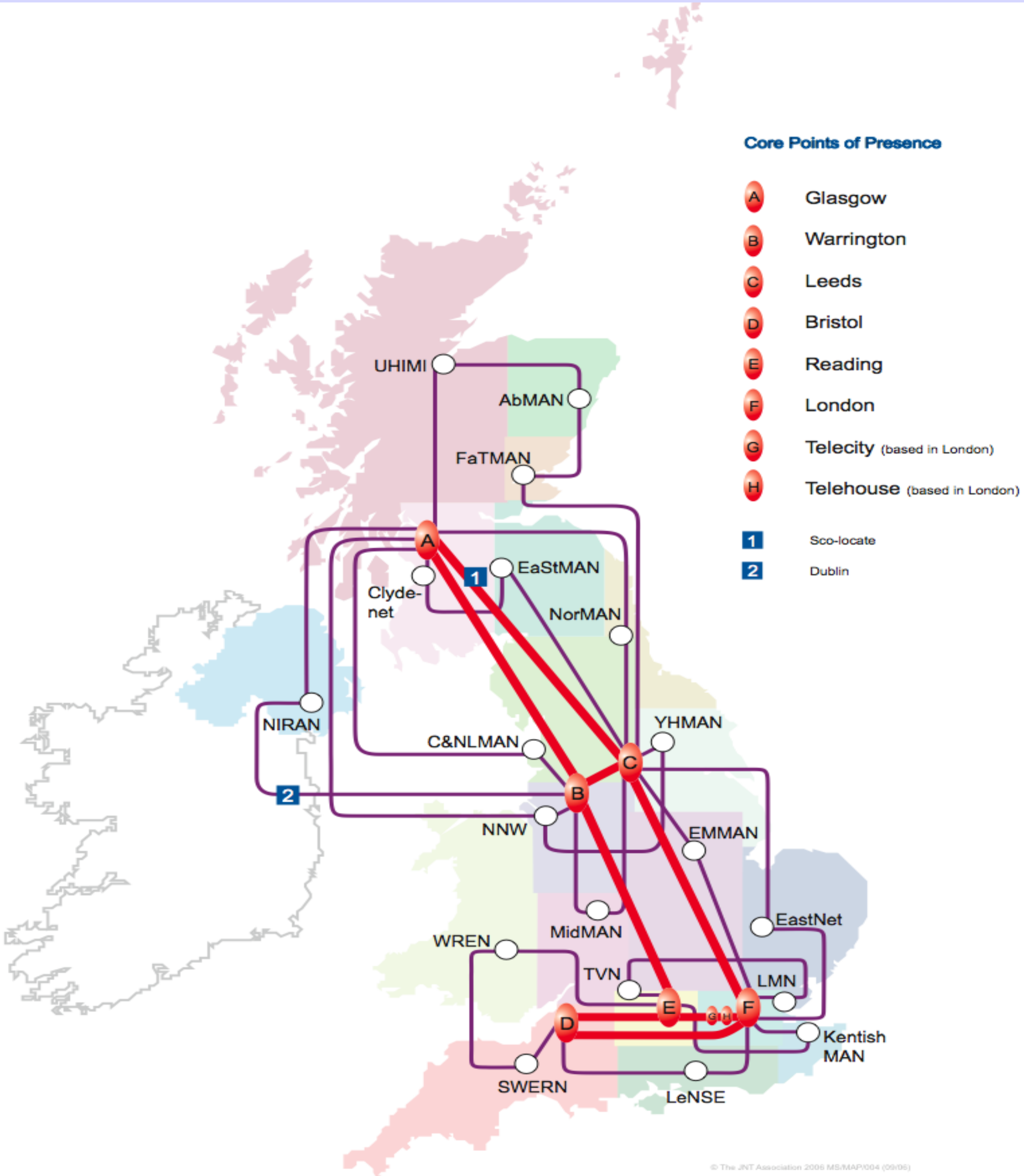
L11 : BGP
Lecture 14
2013

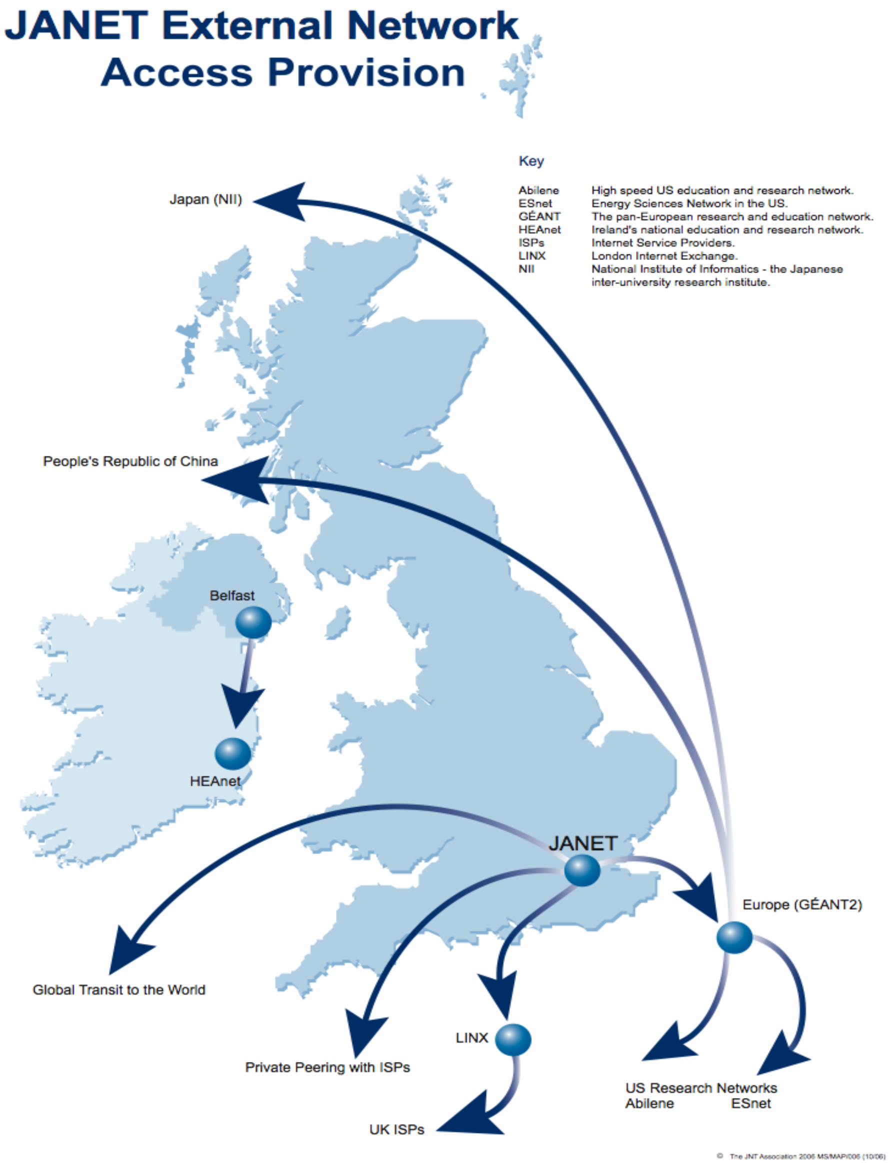
Timothy G. Griffin
Computer Lab
Cambridge UK

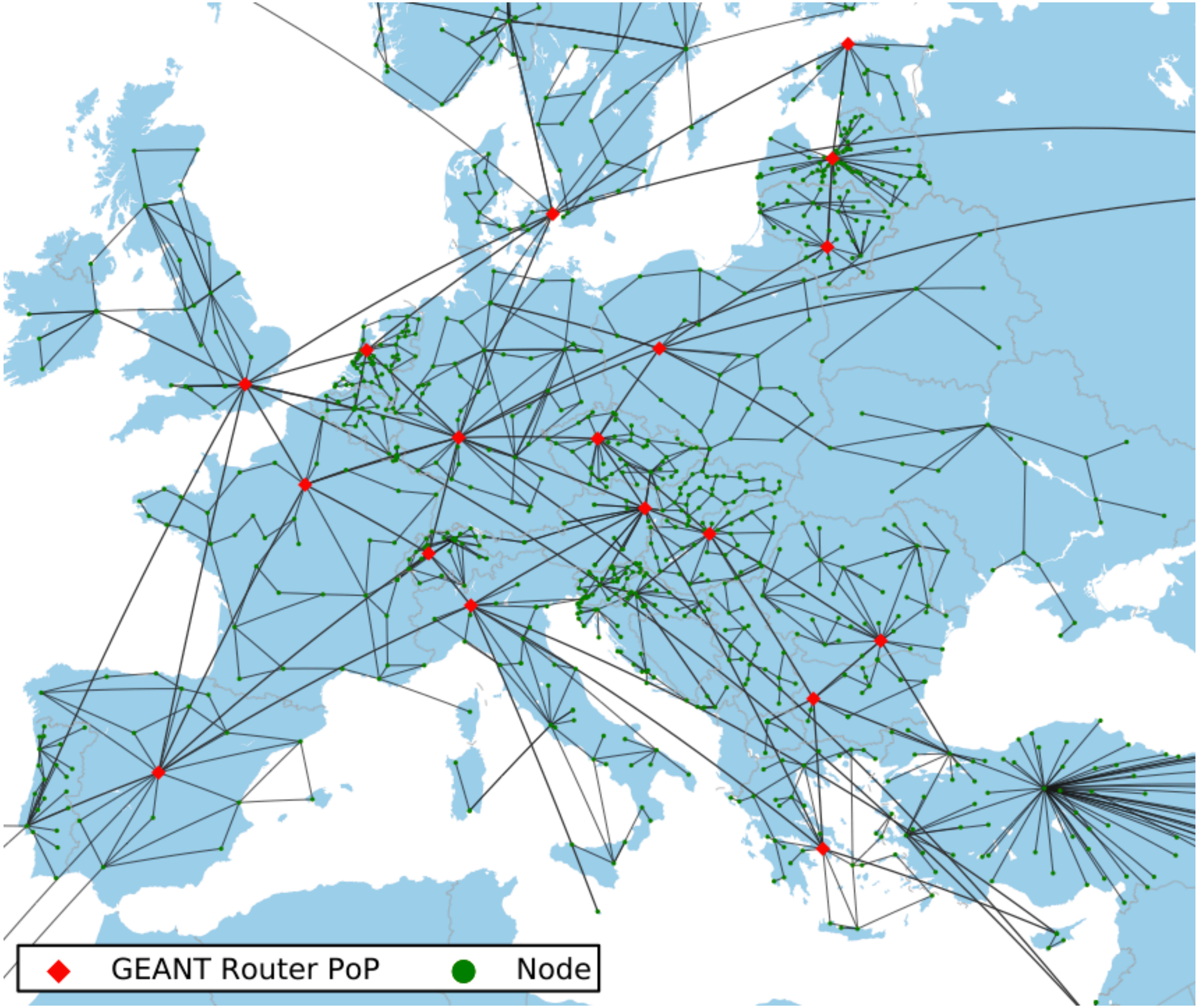
JANET



Core Points of Presence
A Glasgow
B Warrington
C Leeds
D Bristol
E Reading
F London
G Telecity (based in London)
H Telehouse (based in London)
1 Sco-locate
2 Dublin
UHIMI
AbMAN
FaTMAN
EaStMAN
Clyde-net
NorMAN
NIRAN
C&NLMAN
YHMAN
NNW
EMMAN
EastNet
MidMAN
WREN
TVN
LMN
Kentish MAN
SWERN
LeNSE

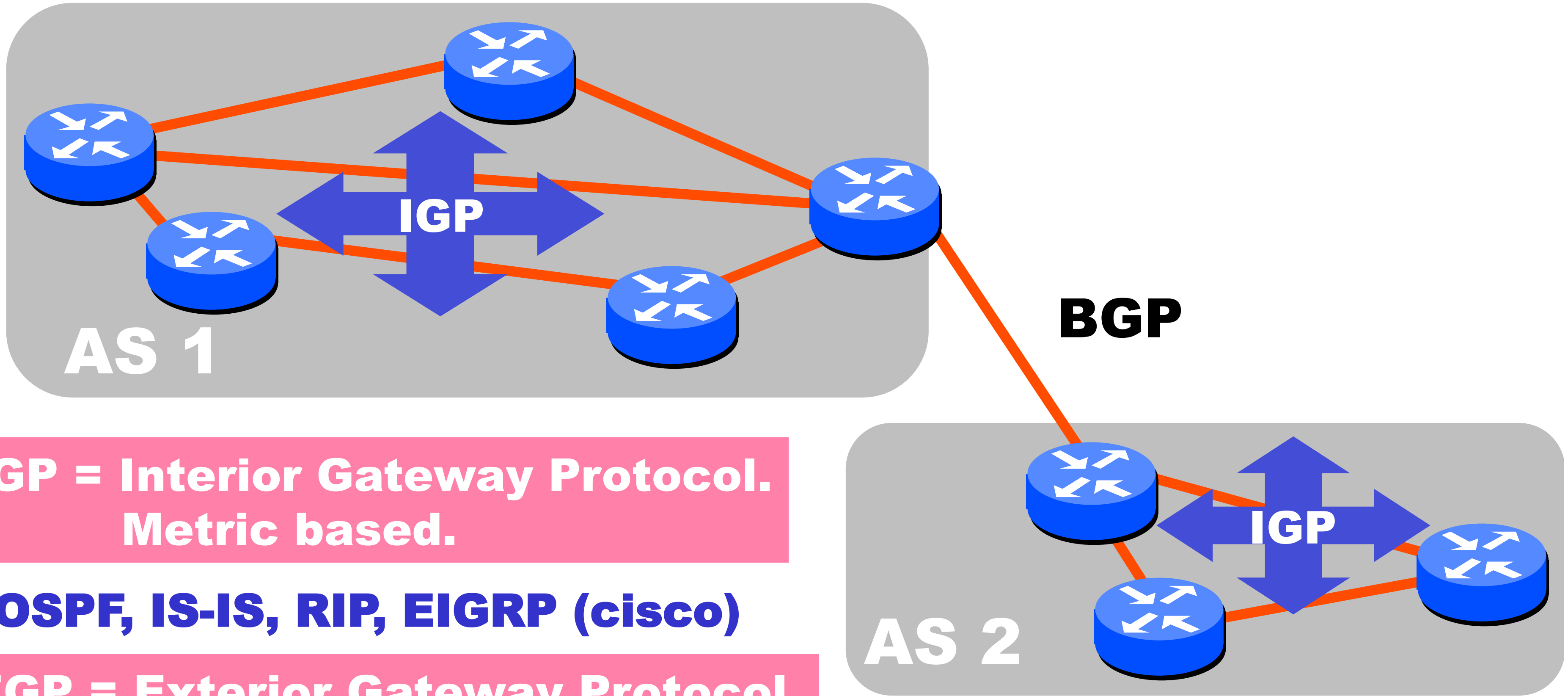
JANET and the Internet





http://www.topology-zoo.org/publications/eu_nren_tech/eu_nren_tech.html

Architecture of Dynamic Routing



IGP = Interior Gateway Protocol.
Metric based.

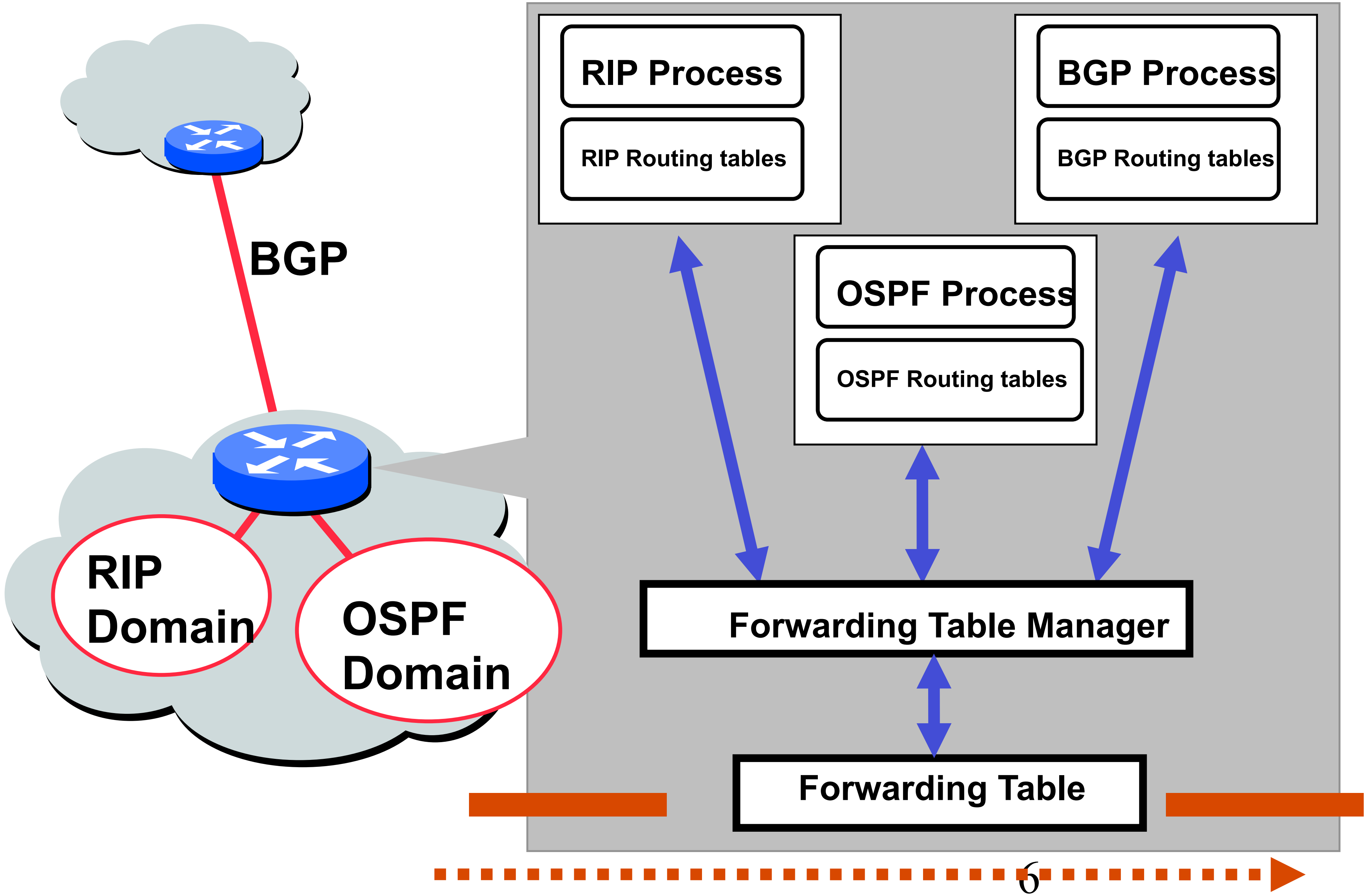
OSPF, IS-IS, RIP, EIGRP (cisco)

EGP = Exterior Gateway Protocol.
Policy Based.

Only one: BGP

The Routing Domain of BGP is the entire Internet

Happy Packets: The Internet Does Not Exist Only to Populated Routing Tables



Autonomous Routing Domains

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.

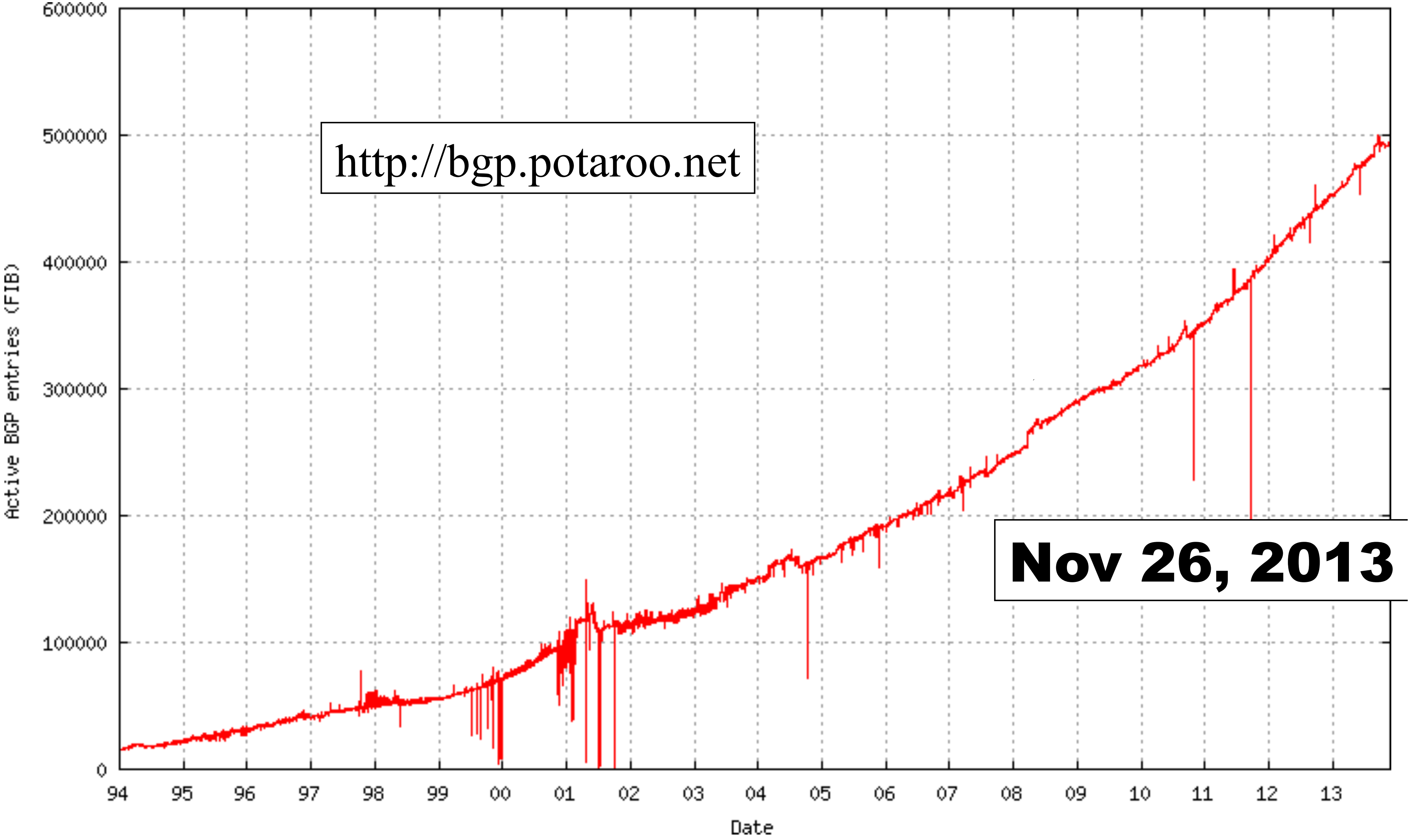
RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System

AS Numbers (ASNs)

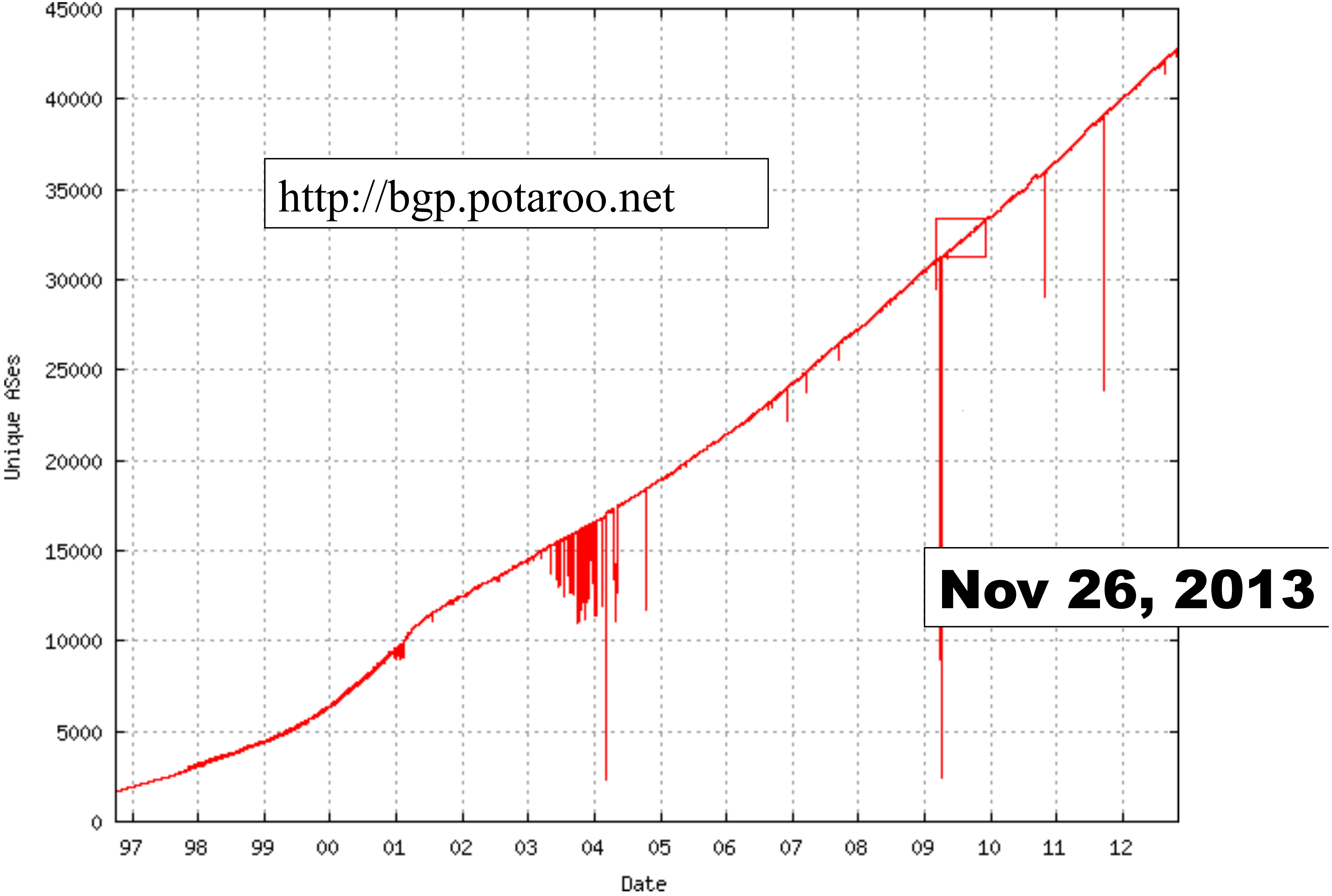
- JANET: 786
- MIT: 3
- Harvard: 11
- UC San Diego: 7377
- AT&T: 7018, 6341, 5074, ...
- UUNET: 701, 702, 284, 12199, ...
- Sprint: 1239, 1240, 6211, 6242, ...
- ...

ASNs represent units of routing policy

How many prefixes are used today?

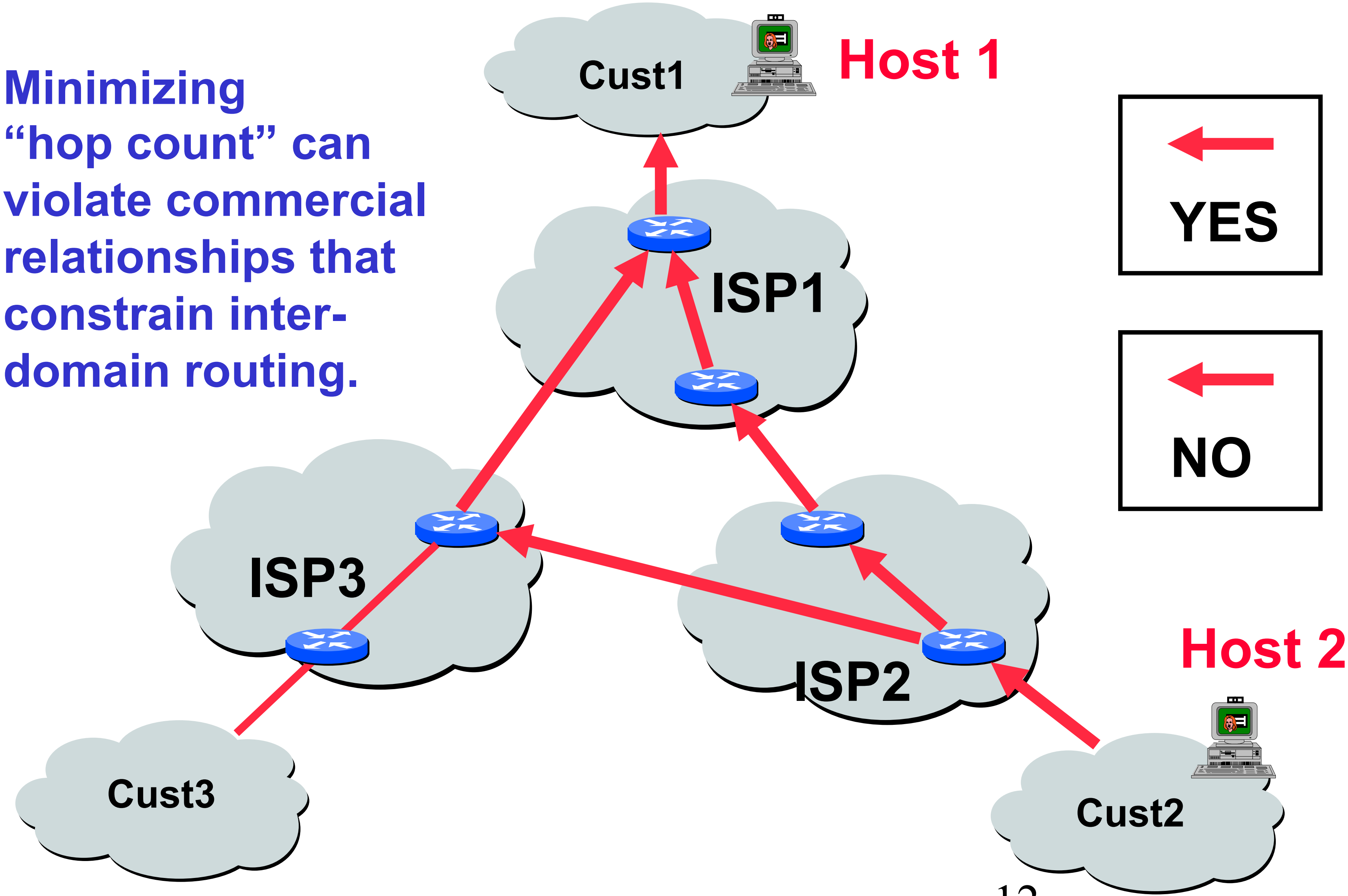


How many ASNs are used today?

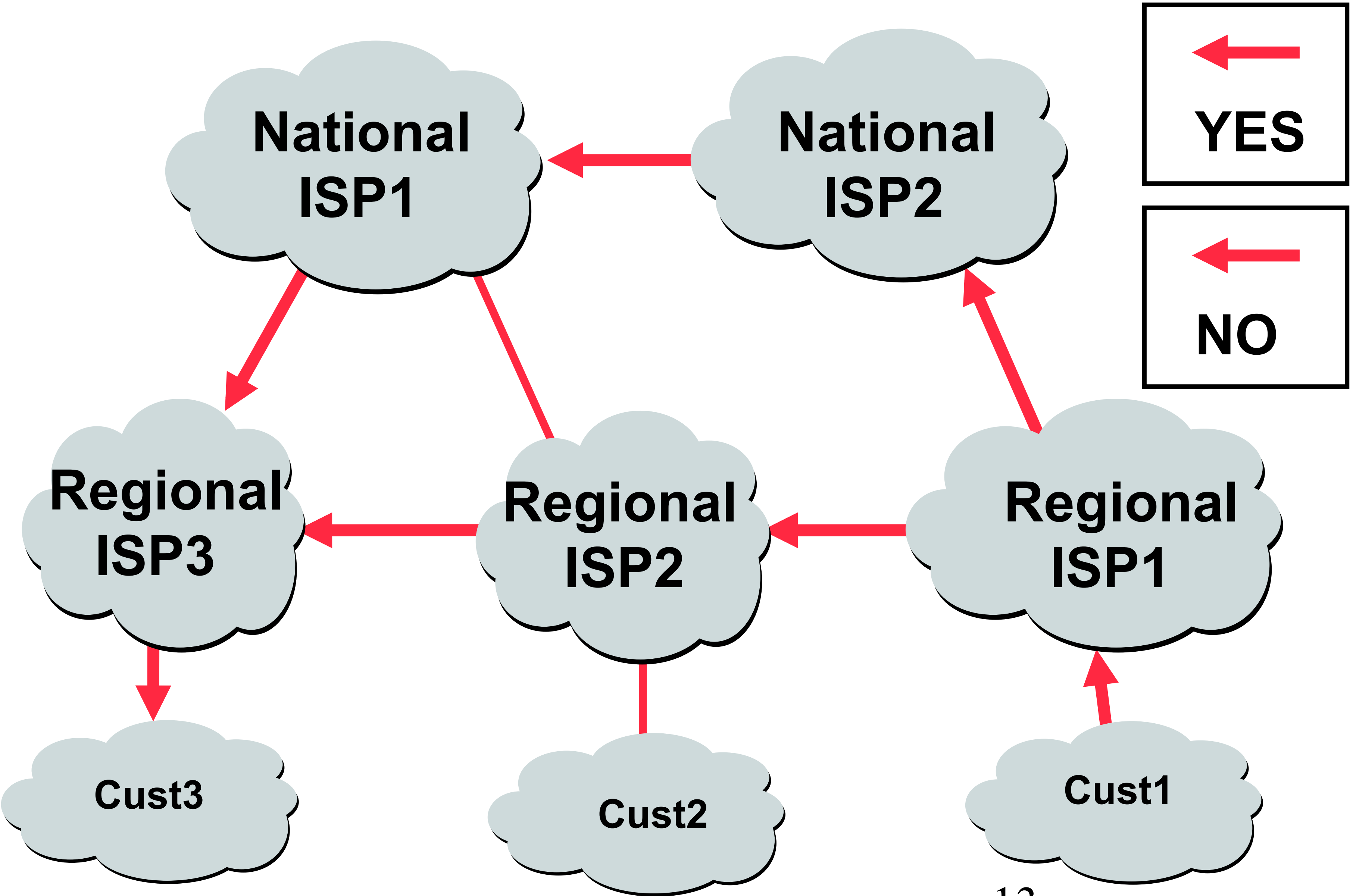


Policy-Based vs. Distance-Based Routing?

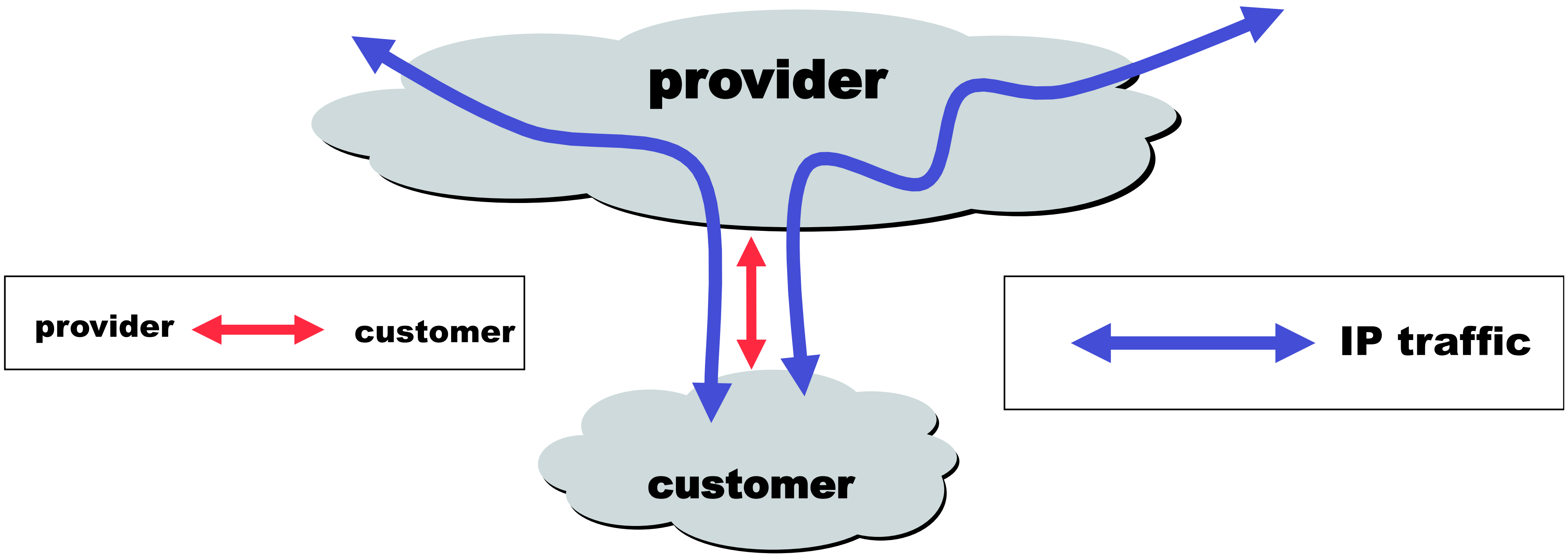
Minimizing "hop count" can violate commercial relationships that constrain inter-domain routing.



Why not minimize "AS hop count"?

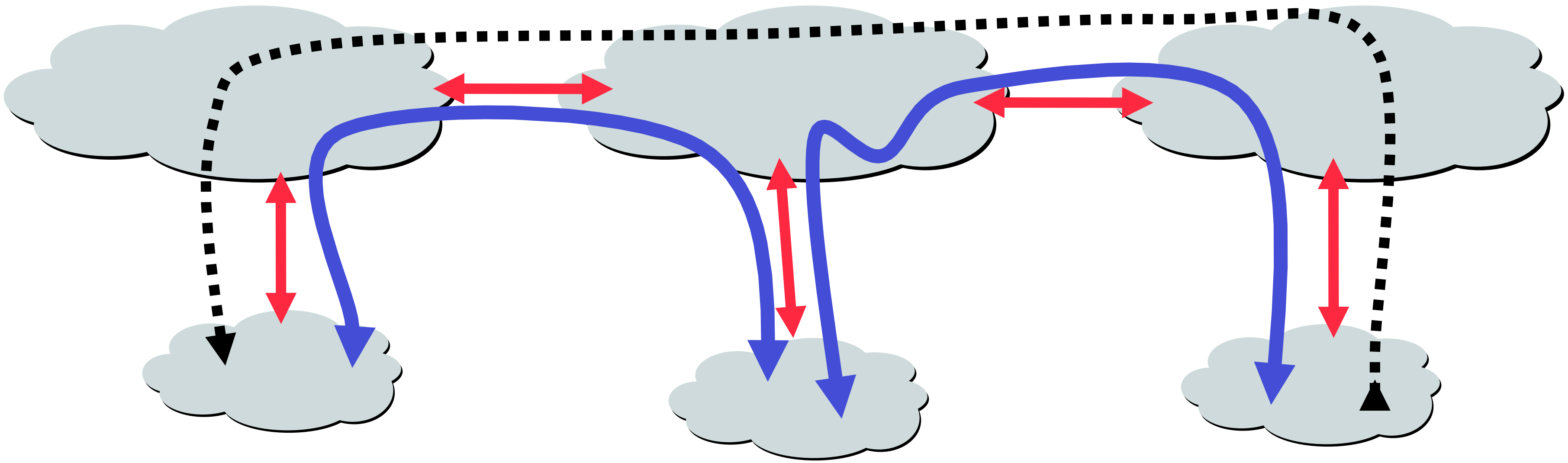


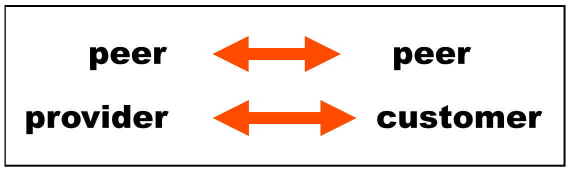
Customers and Providers



Customer pays provider for access to the Internet

The "Peering" Relationship





traffic allowed

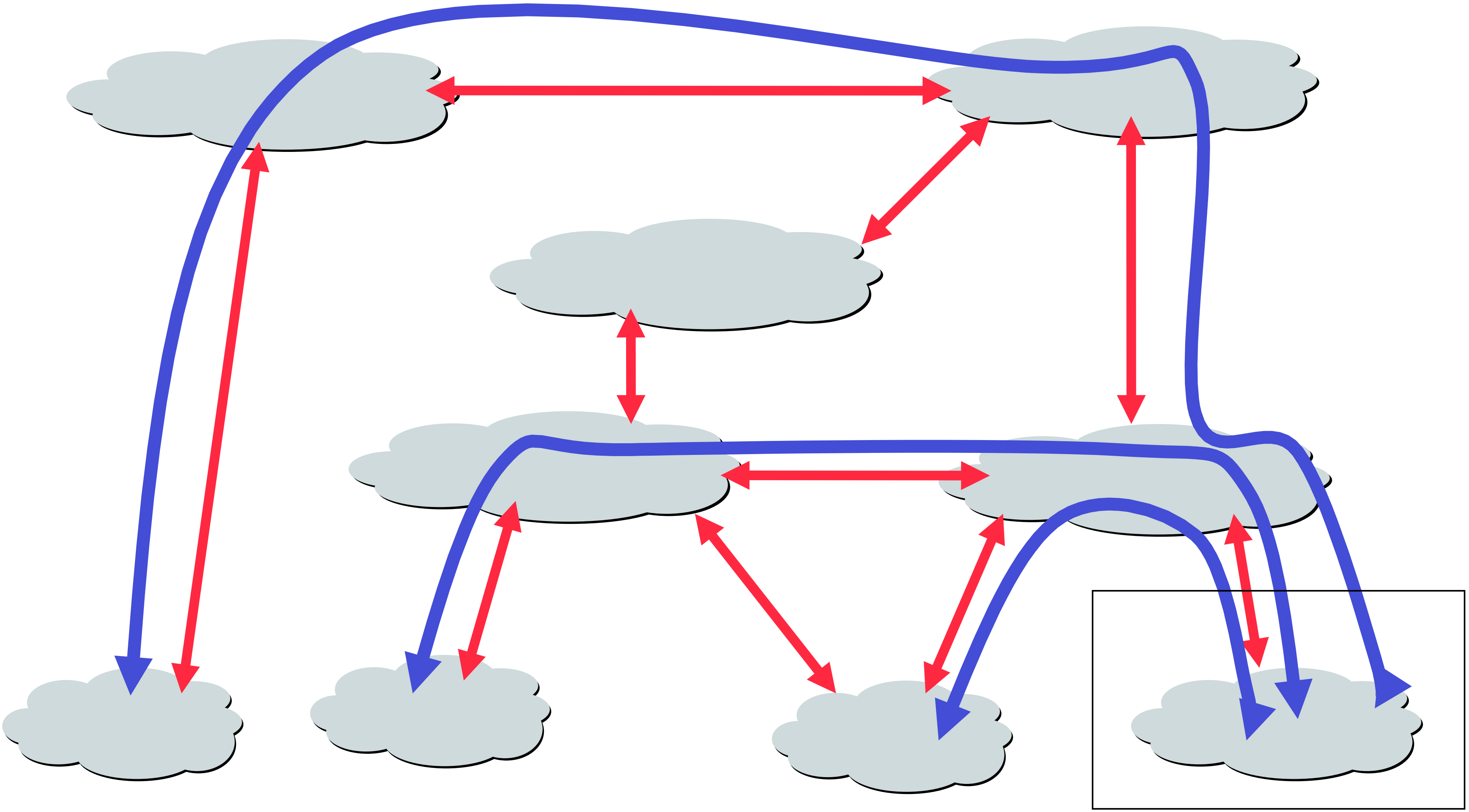
traffic NOT allowed

Peers provide transit between their respective customers

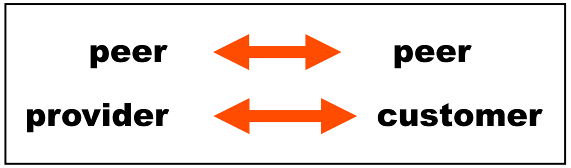
<u>Peers do not provide transit between peers</u>

Peers (often) do not exchange $$$

Peering Provides Shortcuts



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

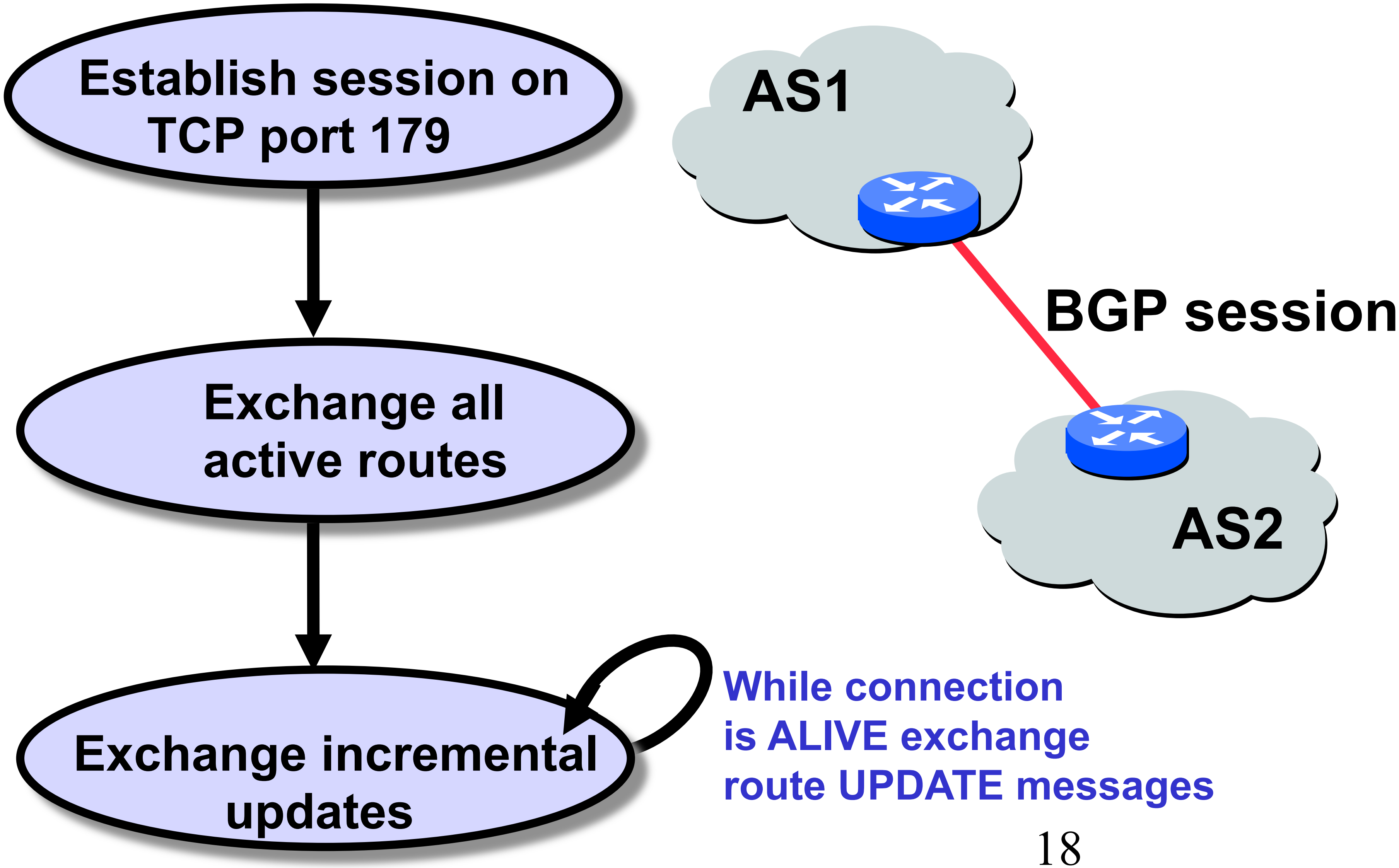


BGP-4

- **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.

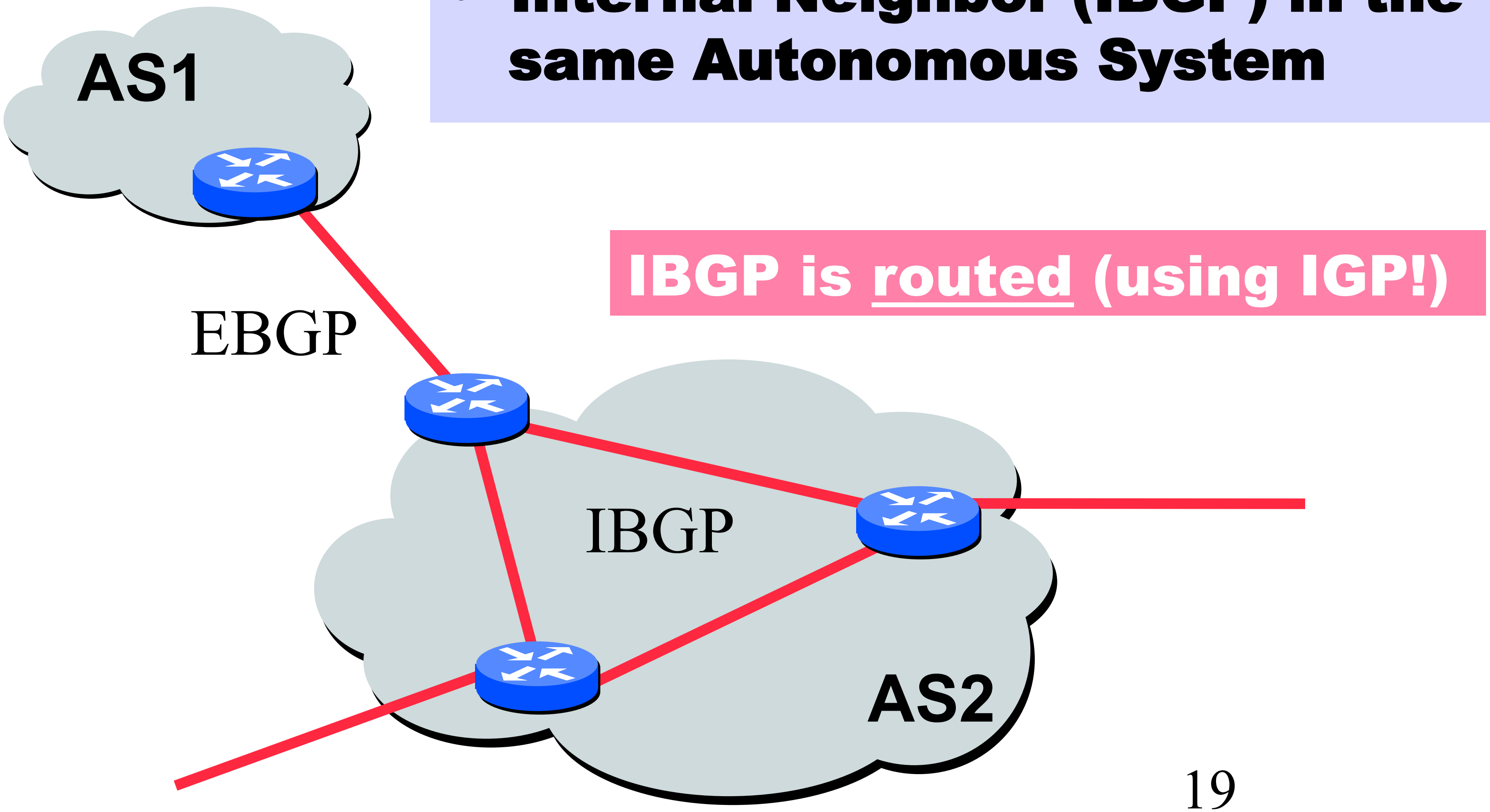
- **1989 : BGP-1 [RFC 1105]**
  - **Replacement for EGP (1984, RFC 904)**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771]**
  - **Support for Classless Interdomain Routing (CIDR)**
- **2006 : BGP-4 [RFC 4271]**

BGP Operations (Simplified)

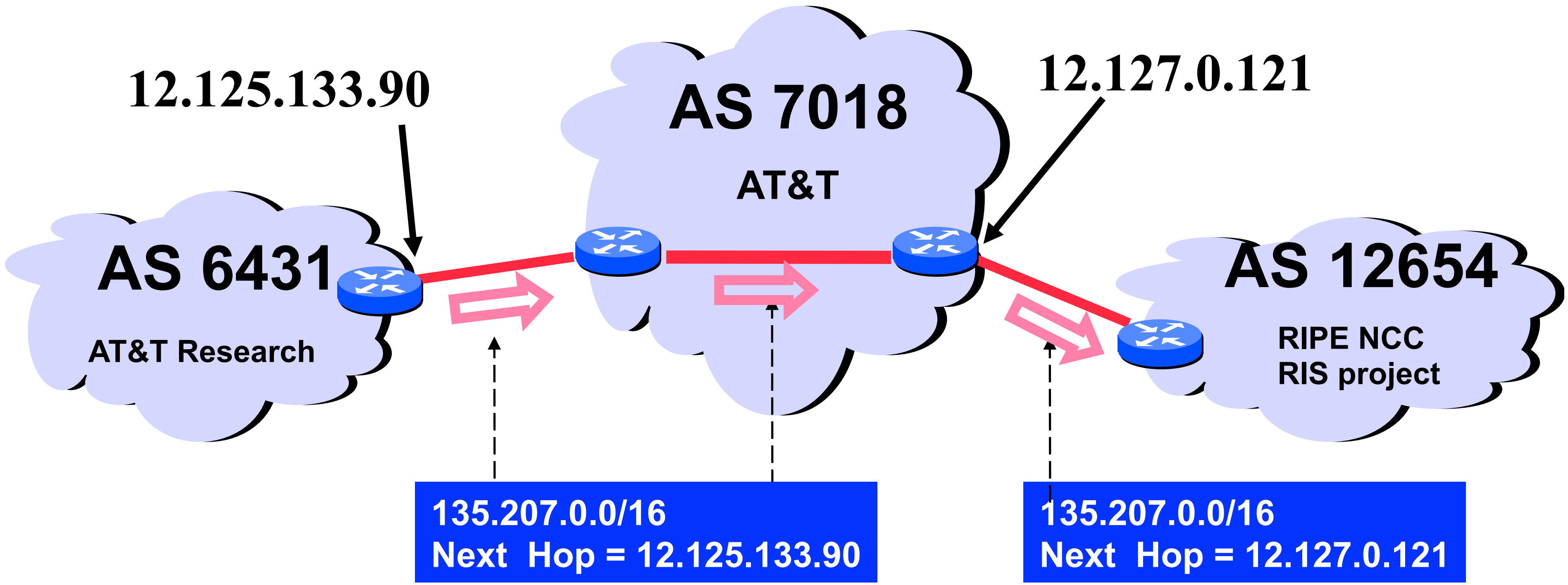


Two Types of BGP Sessions

- **External Neighbor (EBGP) in a different Autonomous Systems**
- **Internal Neighbor (IBGP) in the same Autonomous System**

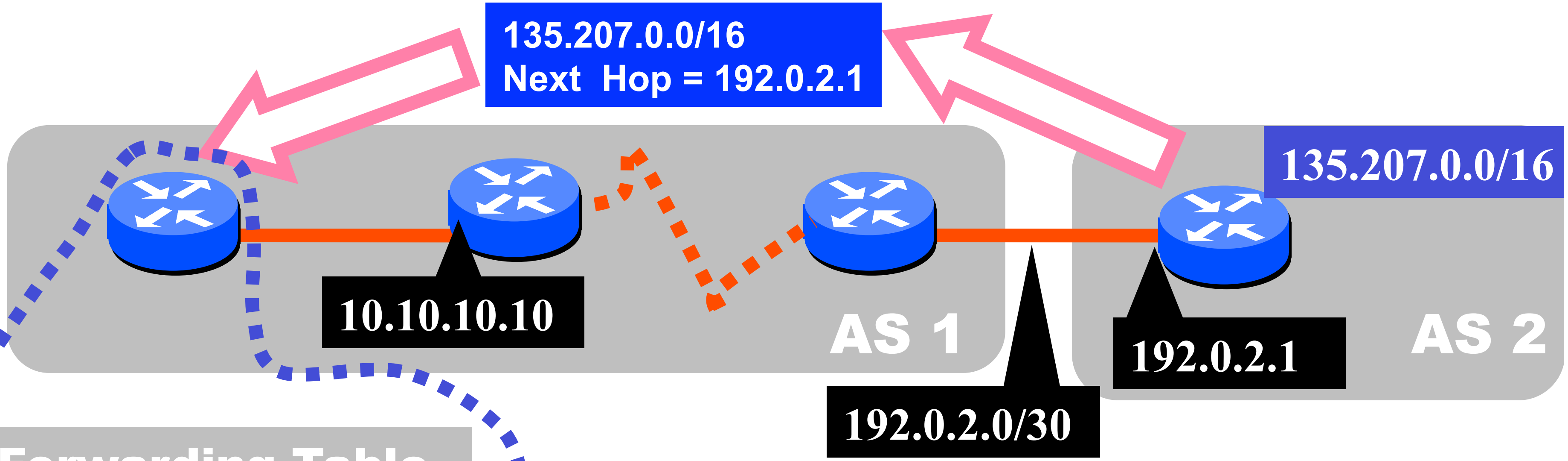


BGP Next Hop Attribute



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

+

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

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

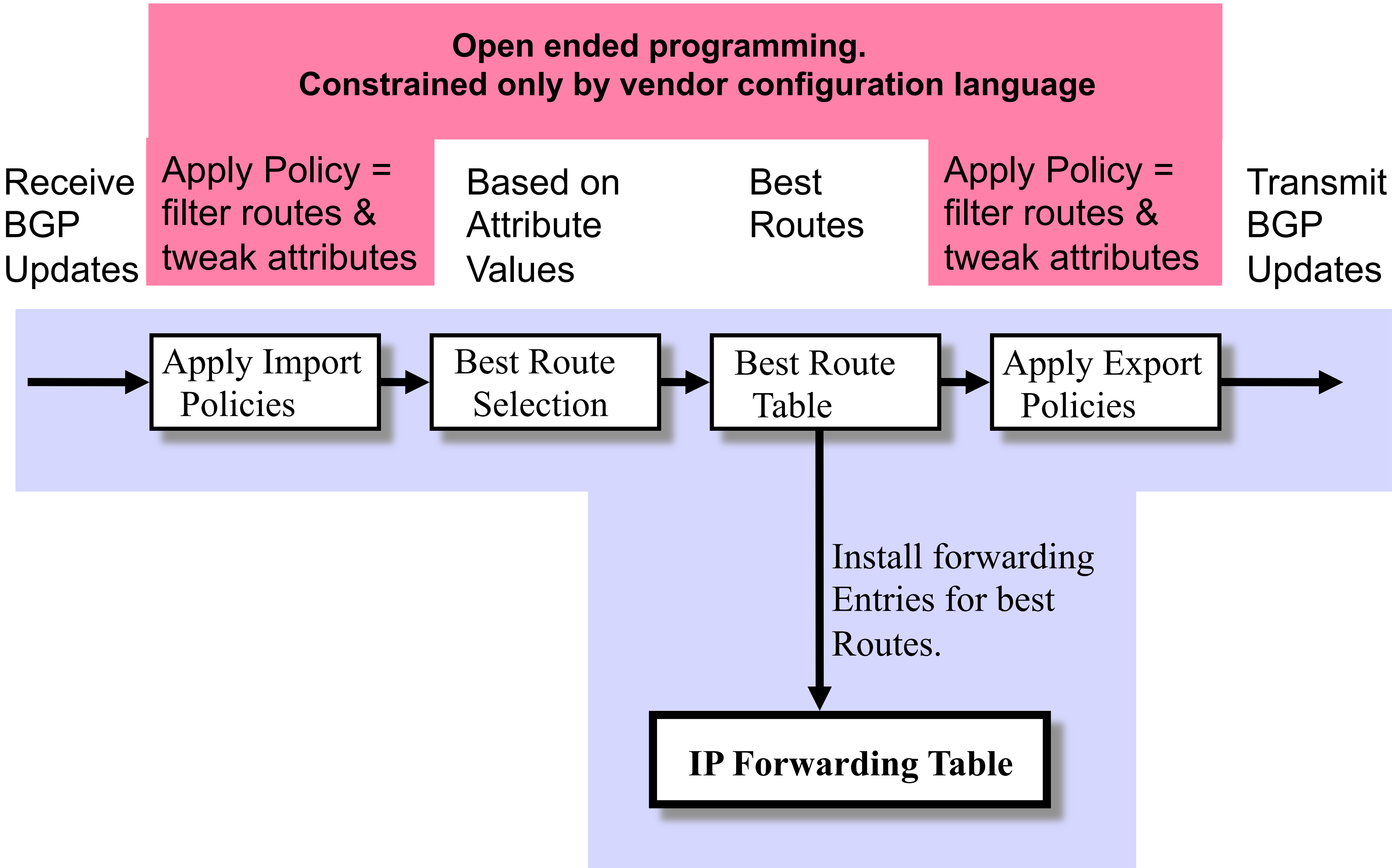
lue	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH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

Most important attributes

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

Not all attributes need to be present in every announcement

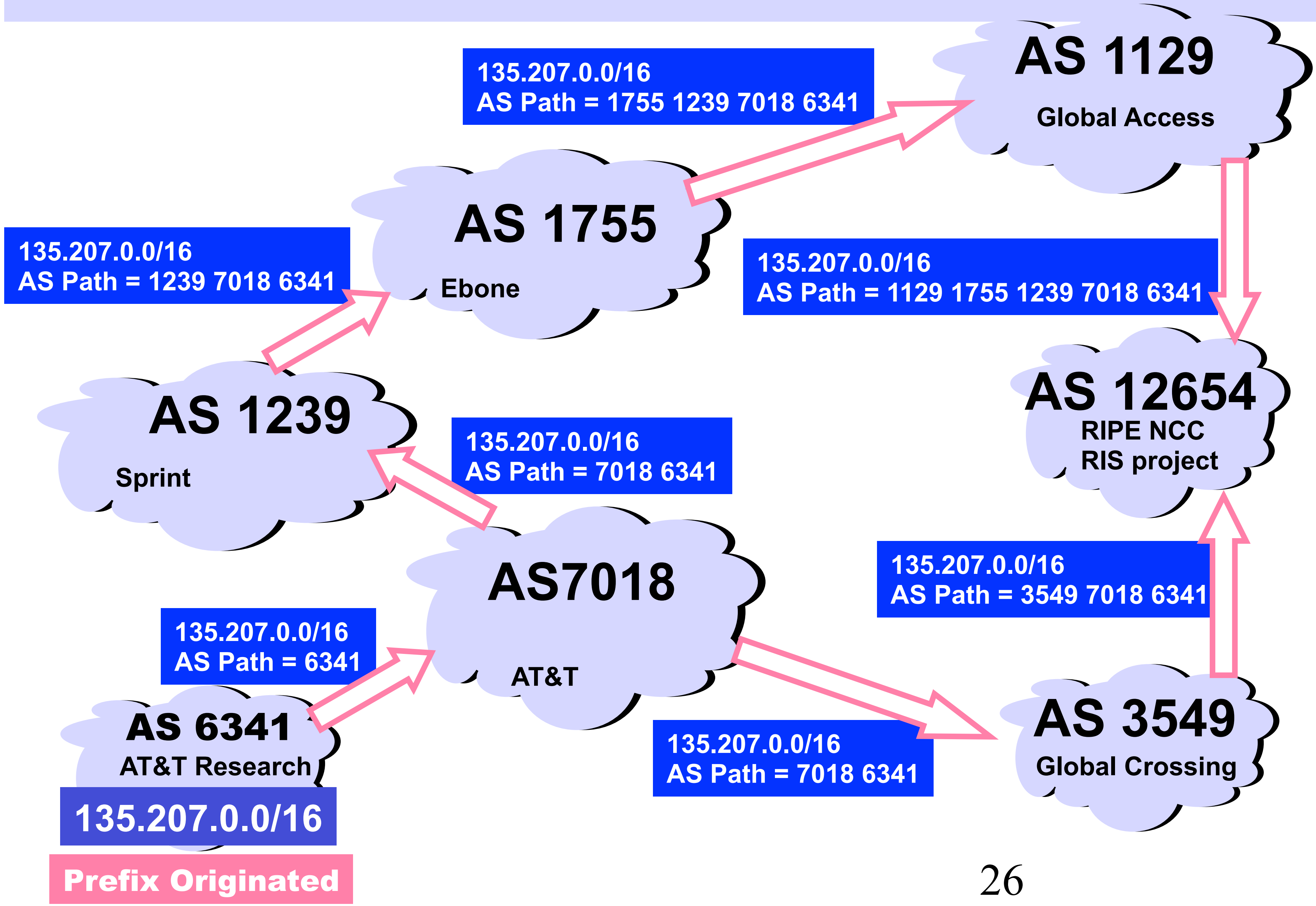
BGP Route Processing



Route Selection Summary

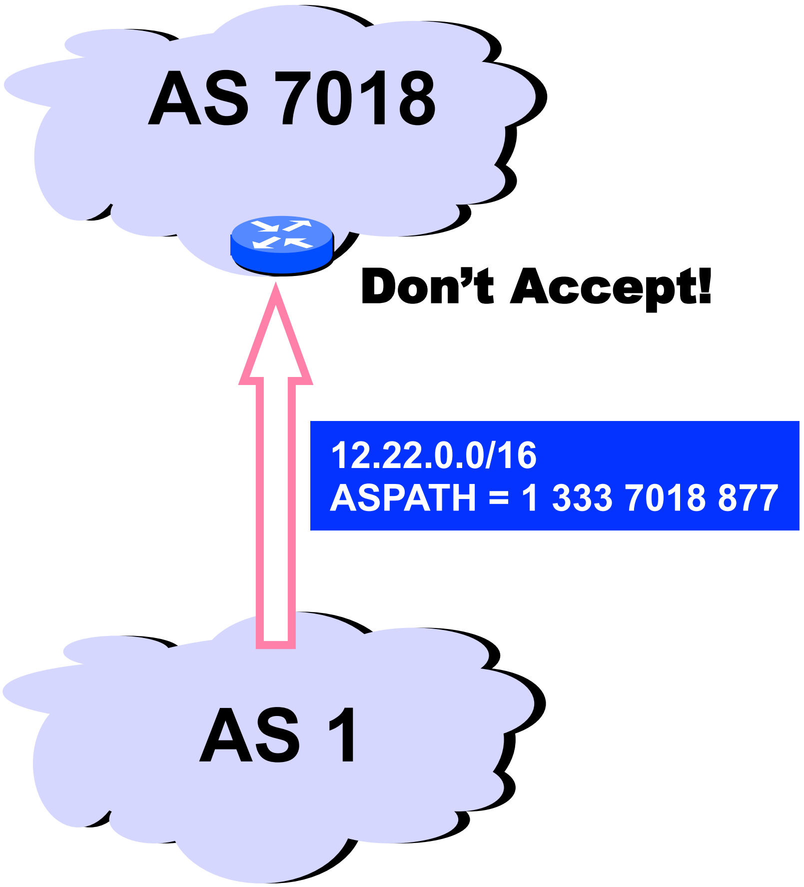
Highest Local Preference	Enforce relationships
Shortest ASPATH Lowest MED i-BGP < e-BGP Lowest IGP cost to BGP egress	traffic engineering
Lowest router ID	Throw up hands and break ties

ASPATH Attribute



Interdomain Loop Prevention

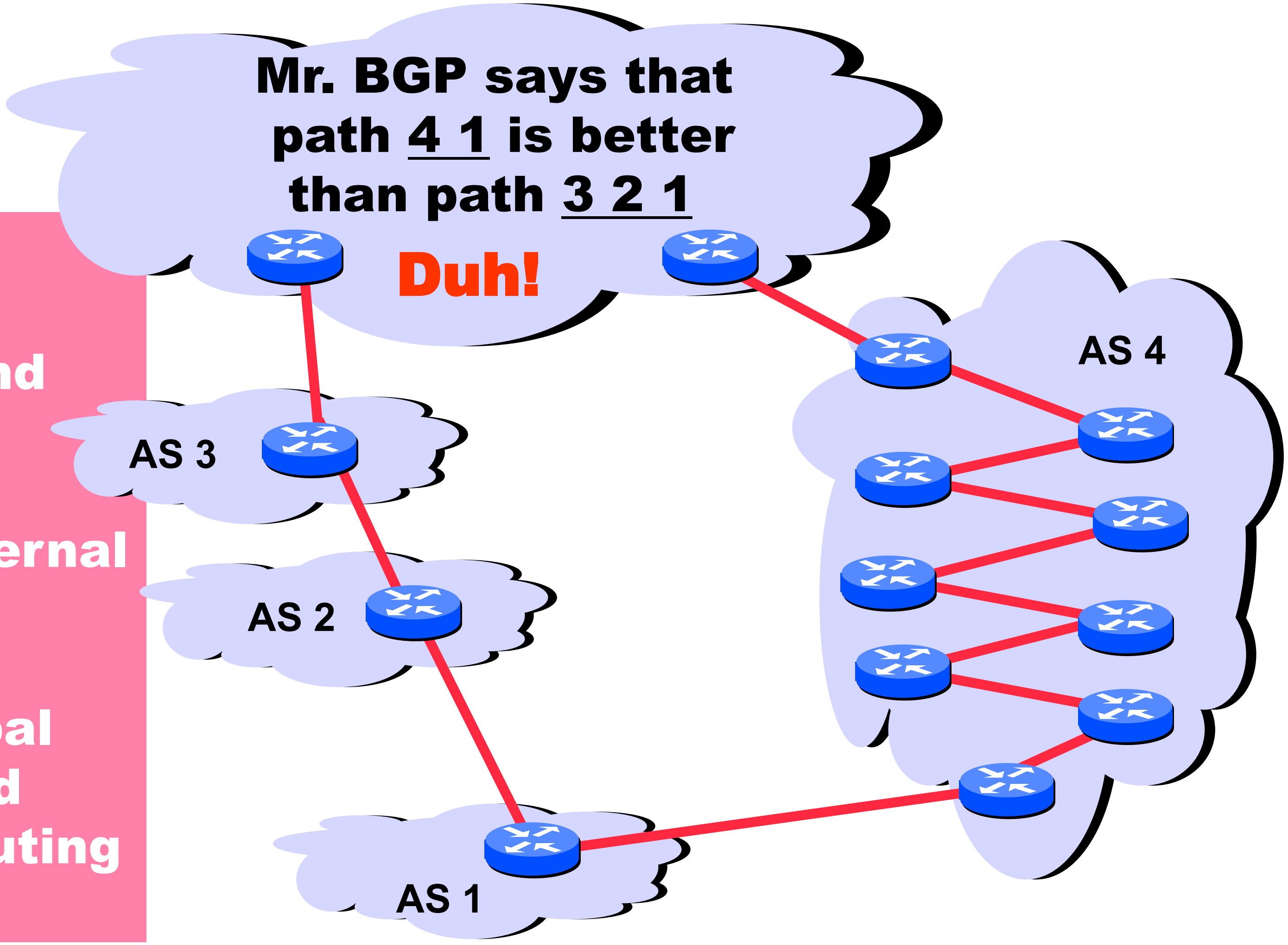
BGP at AS YYY will never accept a route with ASPATH containing YYY.



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

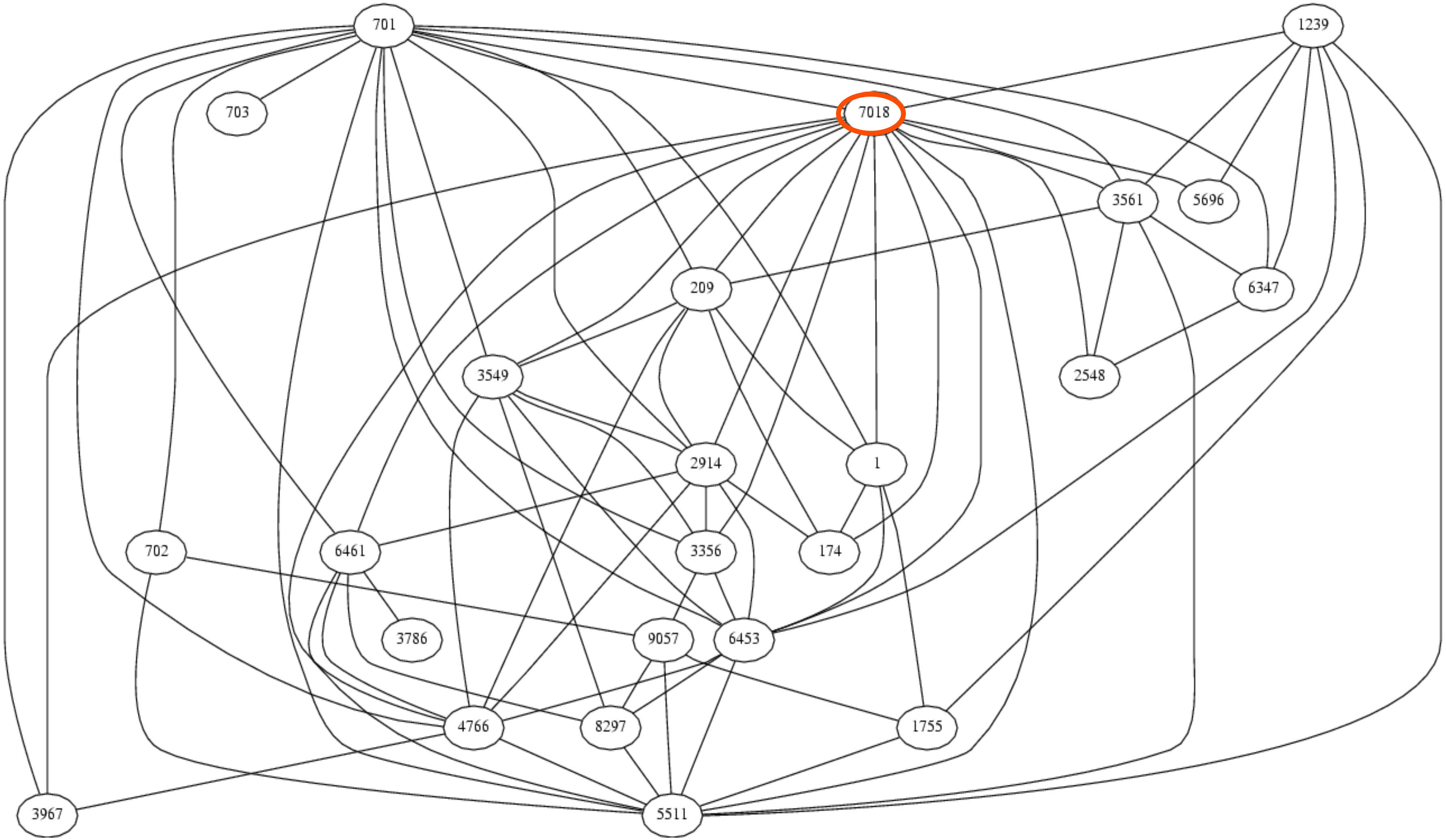


BGP Routing Tables

Thanks to Geoff Huston.
http://bgp.potaroo.net on Feb 1, 2008

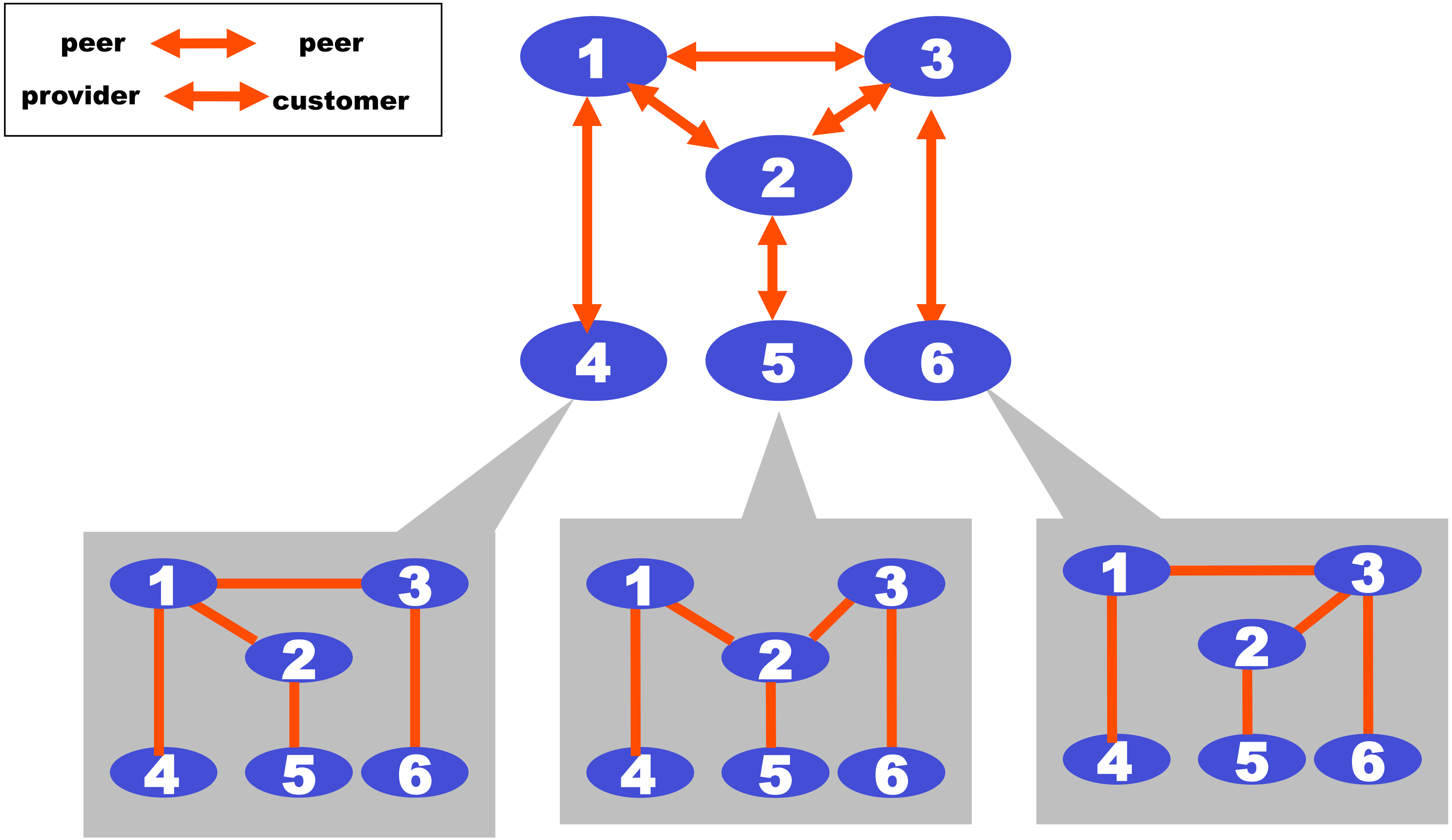
```
show ip bgp
BGP table version is 0, local router ID is 203.119.0.116
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale, R Removed
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 0.0.0.0          193.0.4.28                             0 12654 34225 1299 i
*  3.0.0.0          193.0.4.28                             0 12654 7018 701 703 80 i
*>                  203.50.0.33                            0 65056 4637 703 80 i
*                   202.12.29.79                           0 4608 1221 4637 703 80 i
*  4.0.0.0          193.0.4.28                             0 12654 7018 3356 i
*>                  203.50.0.33                            0 65056 4637 3356 i
*                   202.12.29.79                           0 4608 1221 4637 3356 i
*  4.0.0.0/9        193.0.4.28                             0 12654 7018 3356 i
*>                  203.50.0.33                            0 65056 4637 3356 i
*                   202.12.29.79                           0 4608 1221 4637 3356 i
*  4.23.112.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*  4.23.113.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*  4.23.114.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*  4.36.116.0/23    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*  4.36.116.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*  4.36.117.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*  4.36.118.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.78.22.0/23     193.0.4.28                             0 12654 3257 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909 i
*                   203.50.0.33                            0 65056 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909
*                   202.12.29.79                           0 4608 1221 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 1
*> 4.78.56.0/23     193.0.4.28                             0 12654 3257 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909 i
*                   203.50.0.33                            0 65056 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909
*                   202.12.29.79                           0 4608 1221 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 1
*  4.79.181.0/24    193.0.4.28                             0 12654 3741 10310 14780 i
*>                  203.50.0.33                            0 65056 4637 10310 14780 i
*                   202.12.29.79                           0 4608 1221 4637 10310 14780 i
```

AS Graphs Can Be Fun



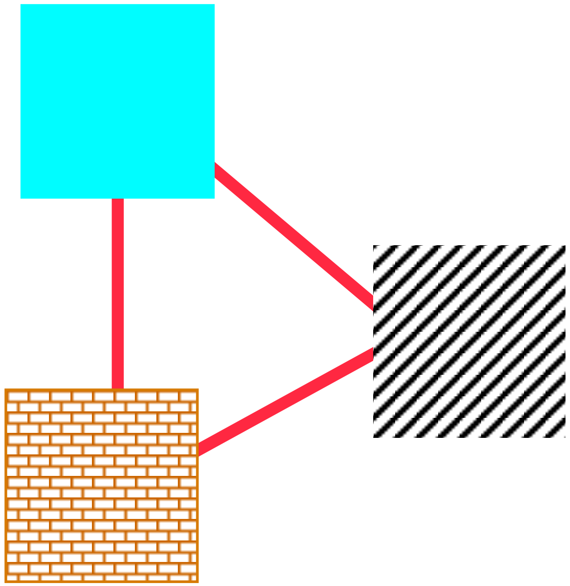
The <u>subgraph</u> showing all ASes that have more than 100 neighbors in full graph of 11,158 nodes. July 6, 2001. **Point of view: AT&T route-server**

AS Graphs Depend on Point of View

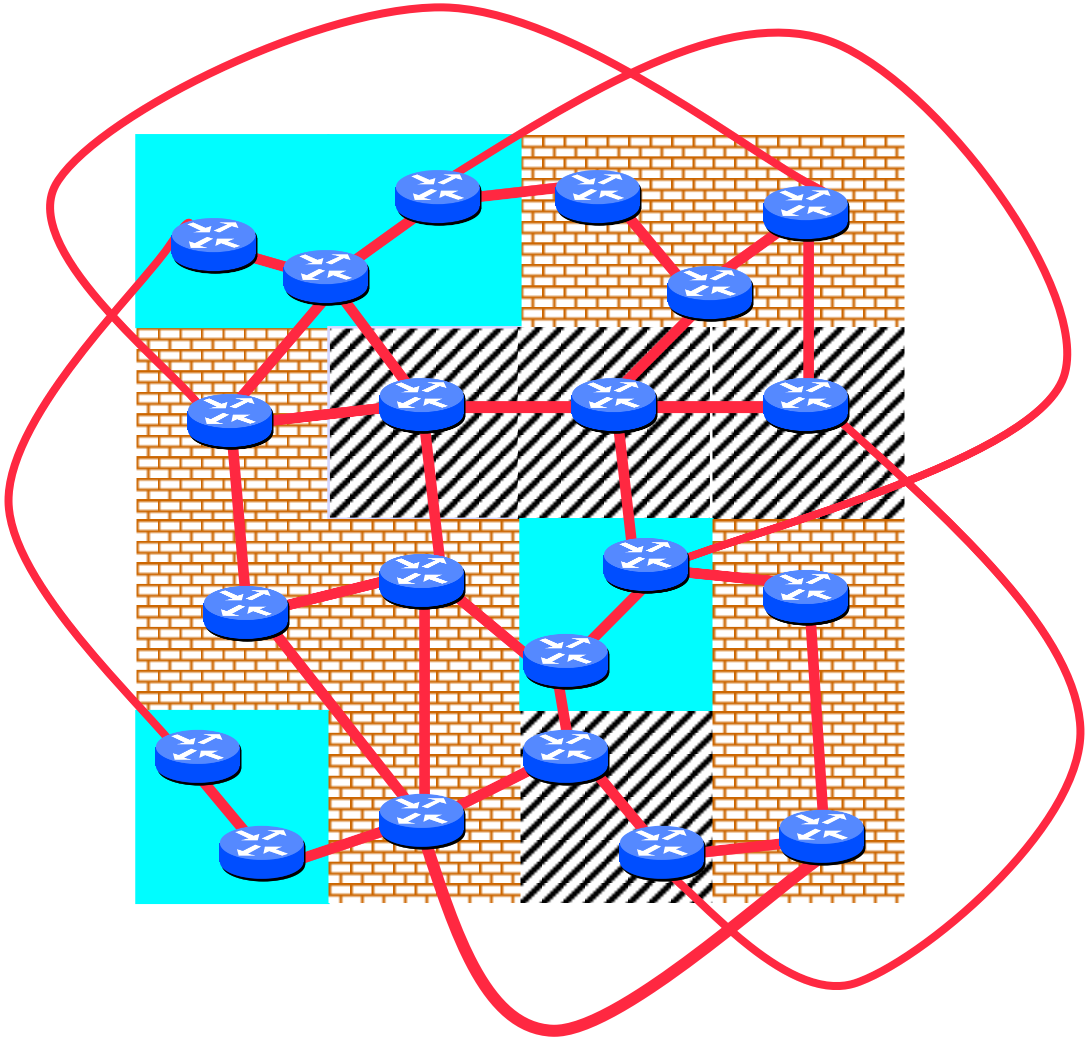


AS Graphs Do Not Show "Topology"!

BGP was designed to throw away information!



The AS graph may look like this.



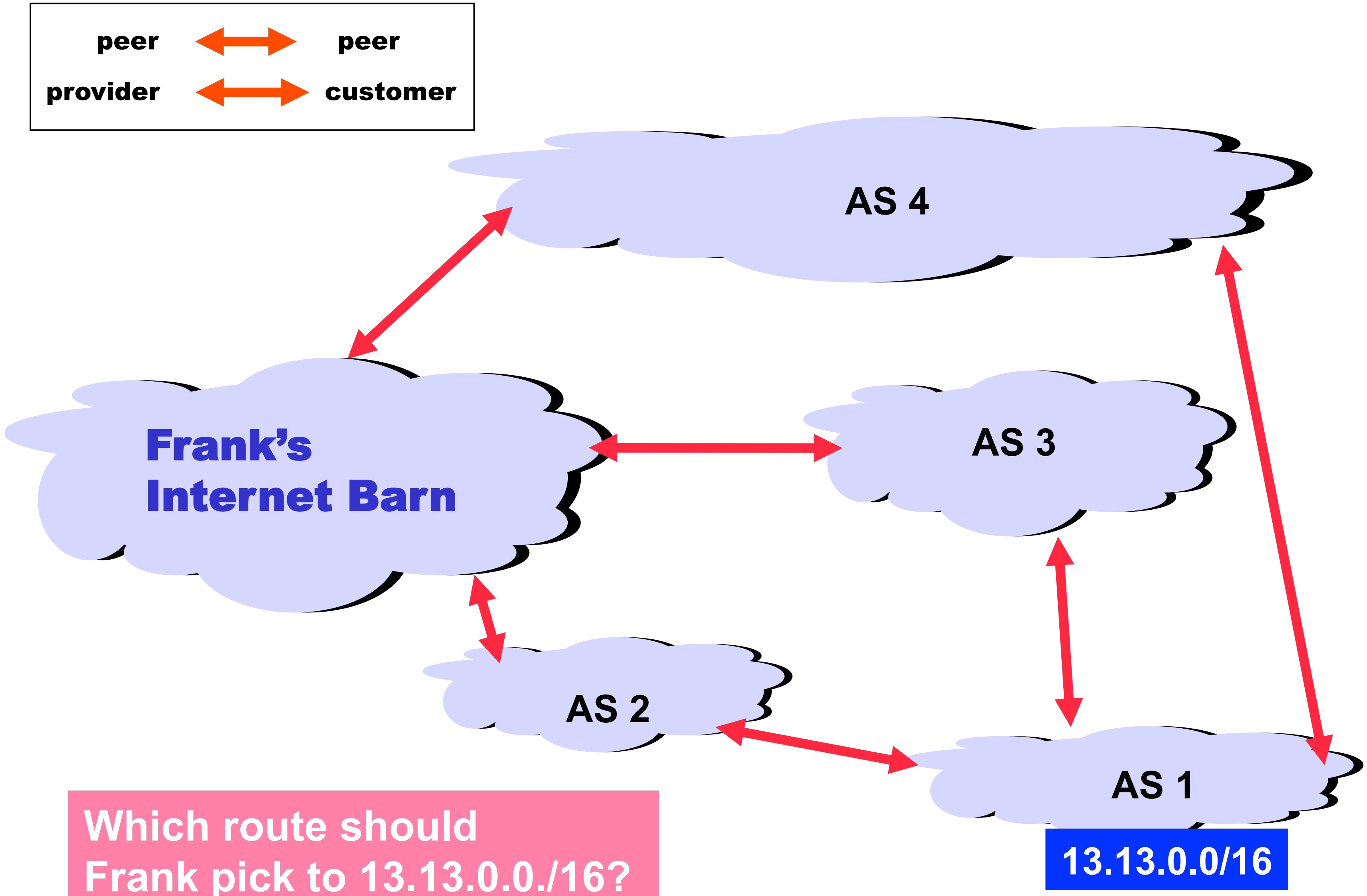
Reality may be closer to this...

Implementing Customer/Provider and Peer/Peer relationships

Two parts:

- Enforce transit relationships
  - Export all (best) routes to customers
  - Send only own and customer routes to all others
- Enforce order of route preference
  - provider < peer < customer

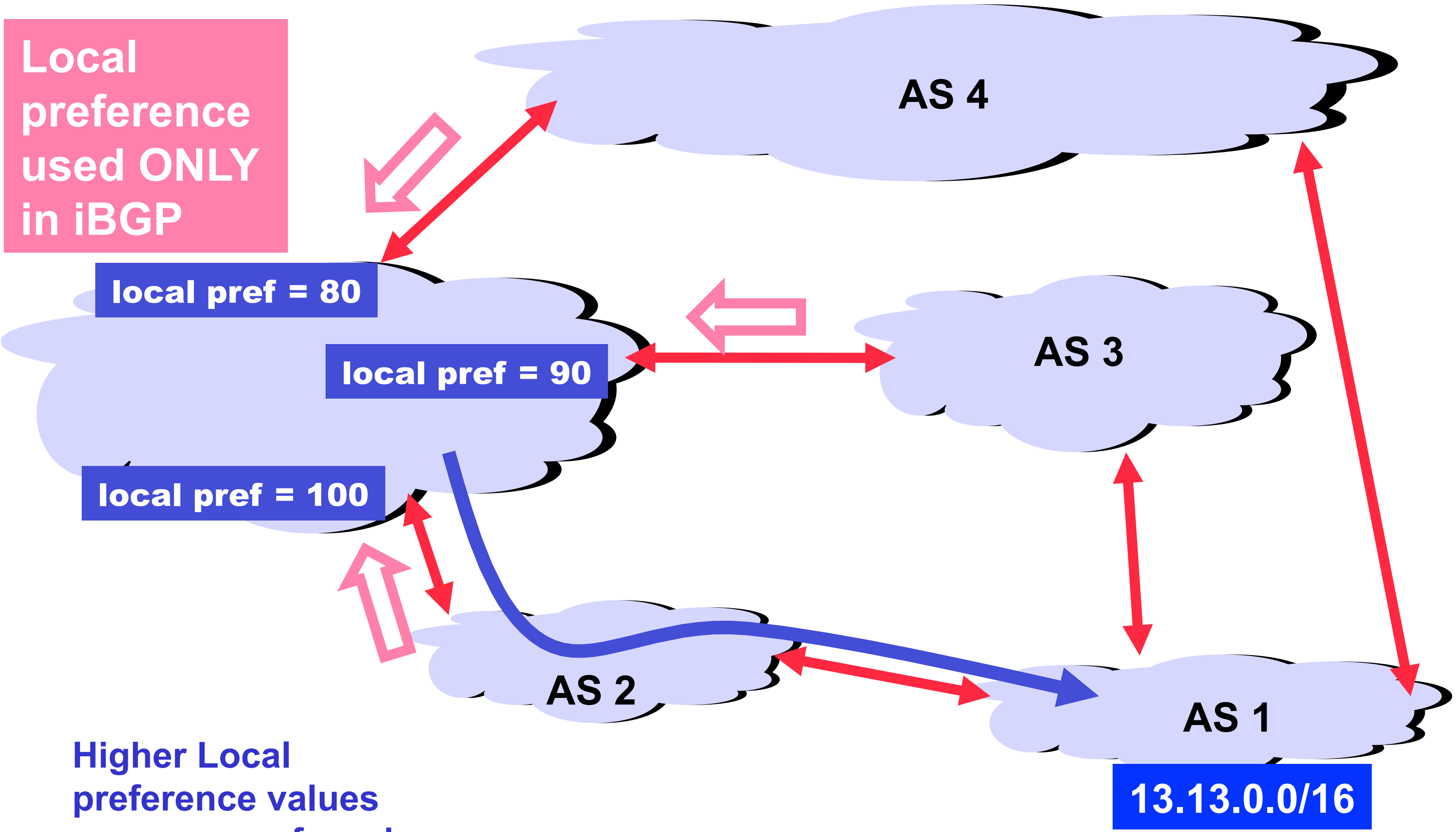
So Many Choices



Which route should Frank pick to 13.13.0.0./16?

LOCAL PREFERENCE

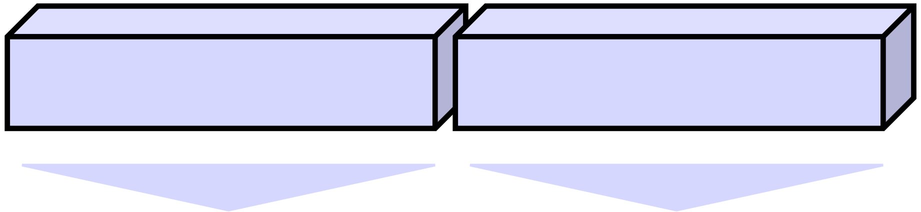
Local preference used ONLY in iBGP



Higher Local preference values are more preferred

How Can Routes be Classified? BGP Communities!

A community value is 32 bits



By convention, first 16 bits is ASN indicating who is giving it an interpretation

community number

Used for signally within and between ASes

Very powerful BECAUSE it has no (predefined) meaning

Community Attribute = a list of community values. (So one route can belong to multiple communities)

RFC 1997 (August 1996)

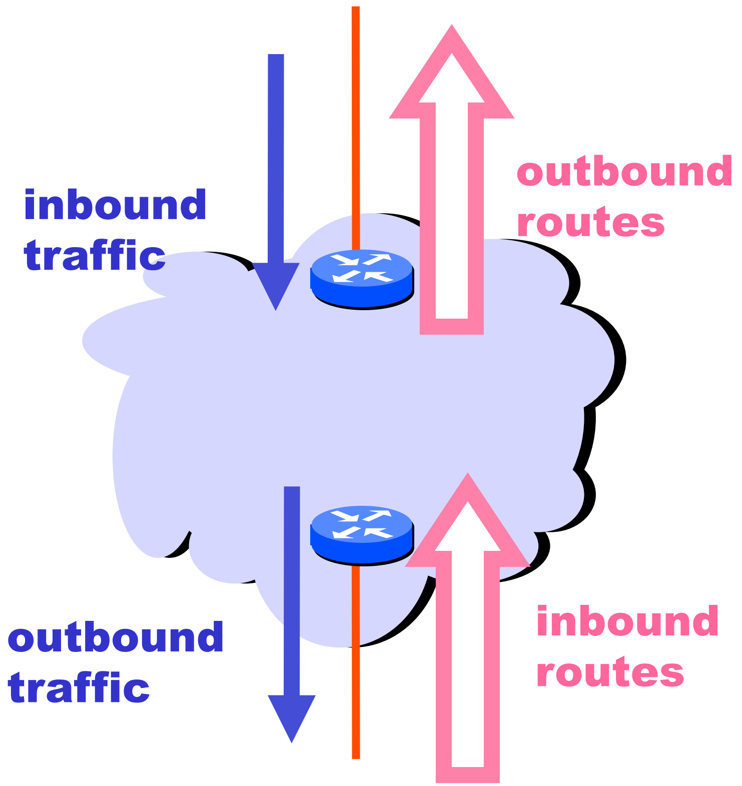
Reserved communities

no_export = 0xFFFFFF01: don’t export out of AS

no_advertise 0xFFFFFF02: don’t pass to BGP neighbors

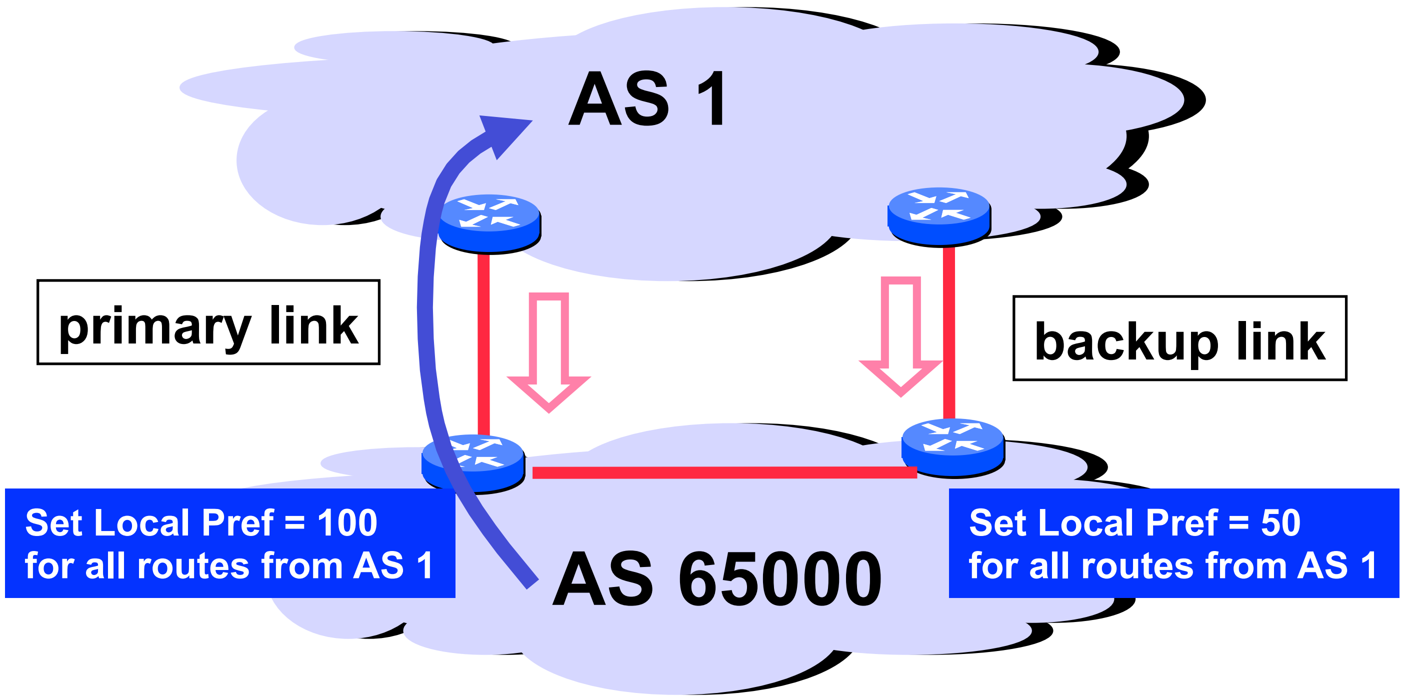
Tweak Tweak Tweak (TE)

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



In general, an AS has more control over outbound traffic

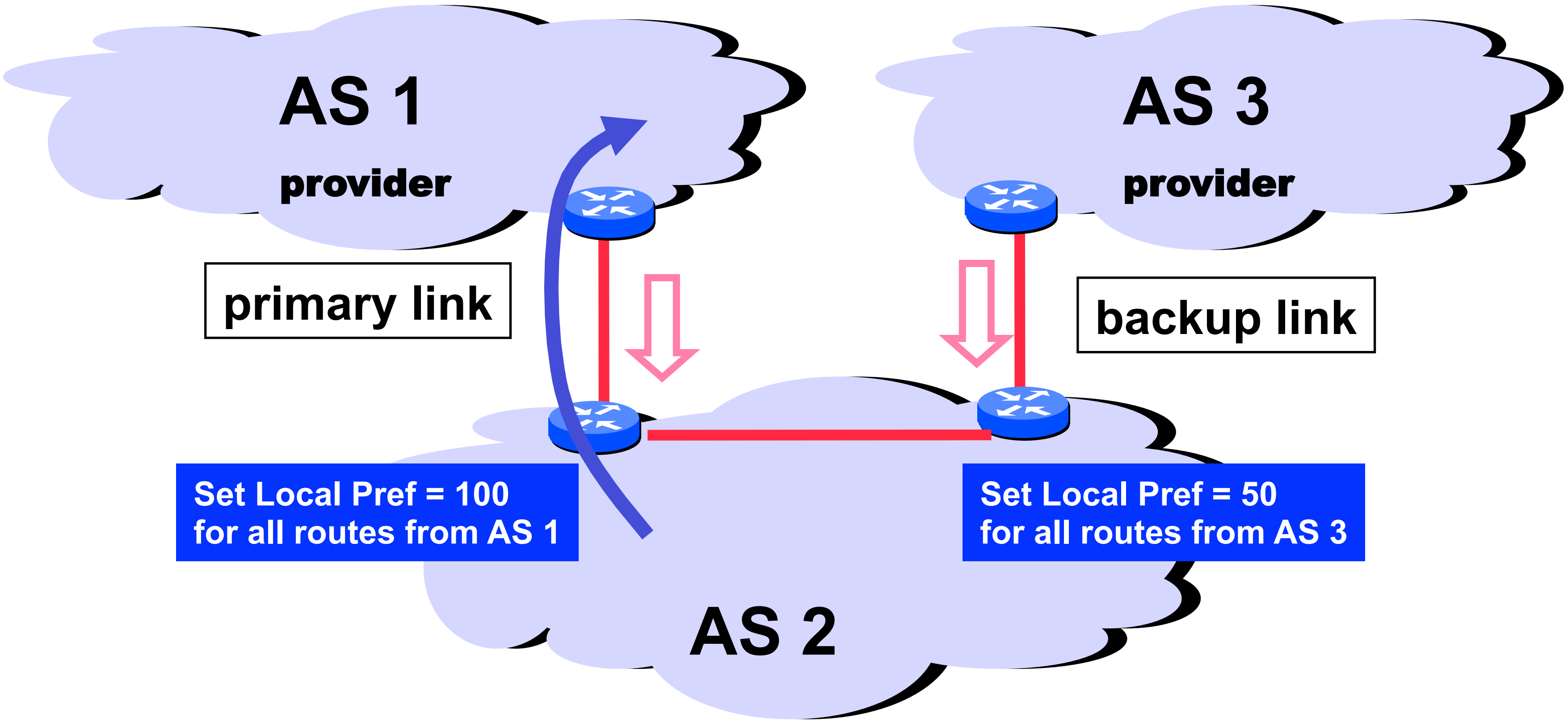
Implementing Backup Links with Local Preference (Outbound Traffic)



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

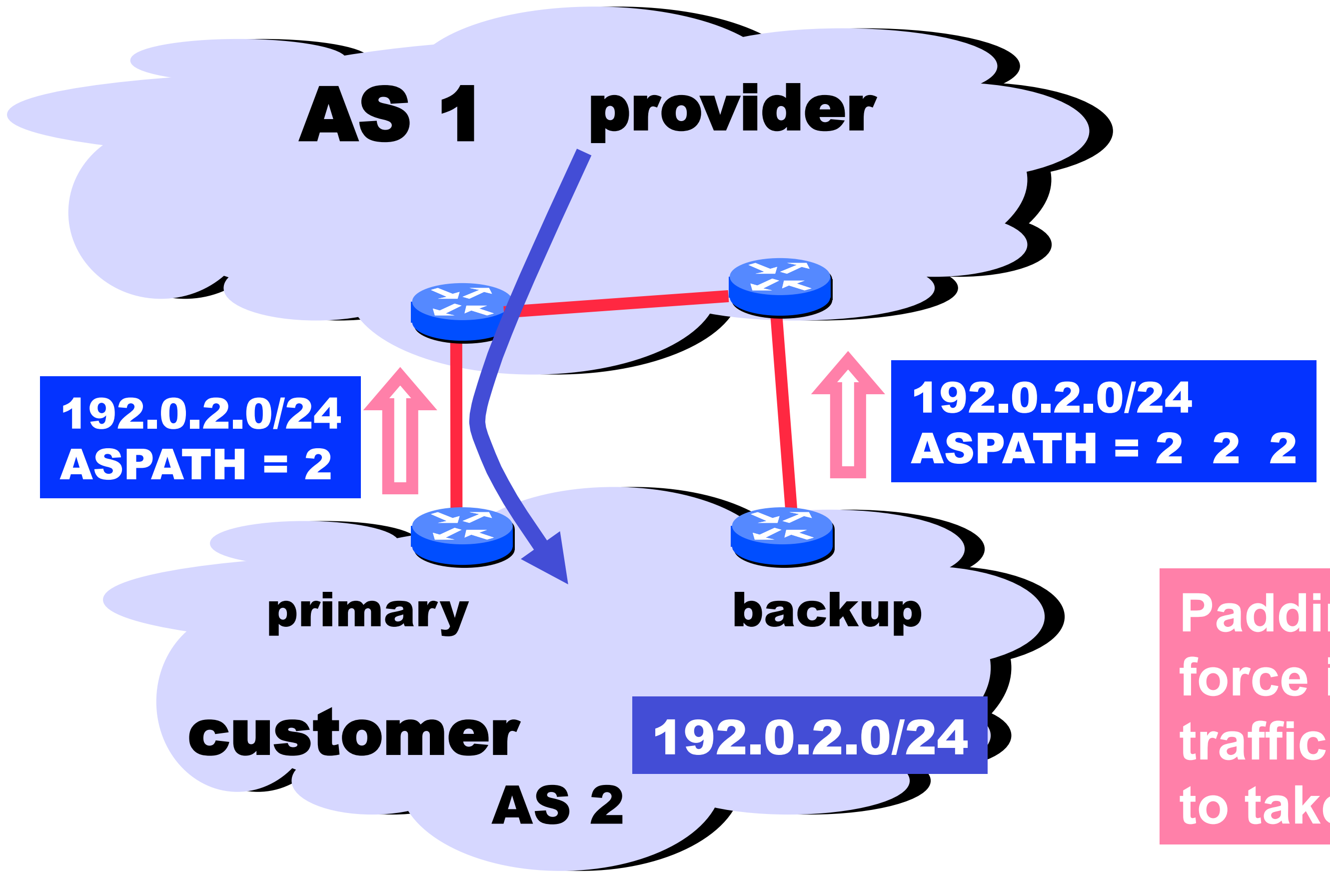
We'll talk about inbound traffic soon ...

Multihomed Backups (Outbound Traffic)



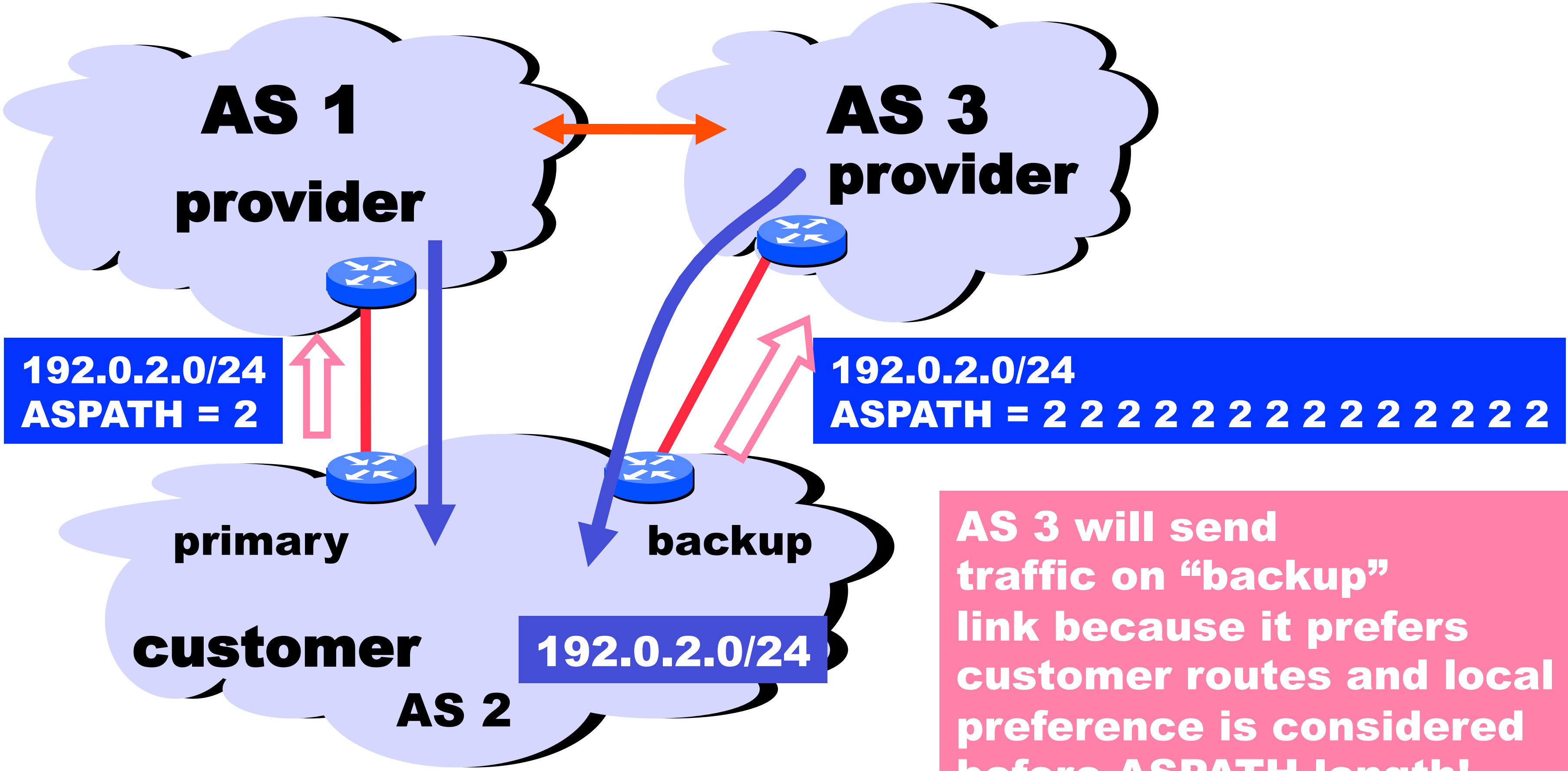
Forces outbound traffic to take primary link, unless link is down.

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

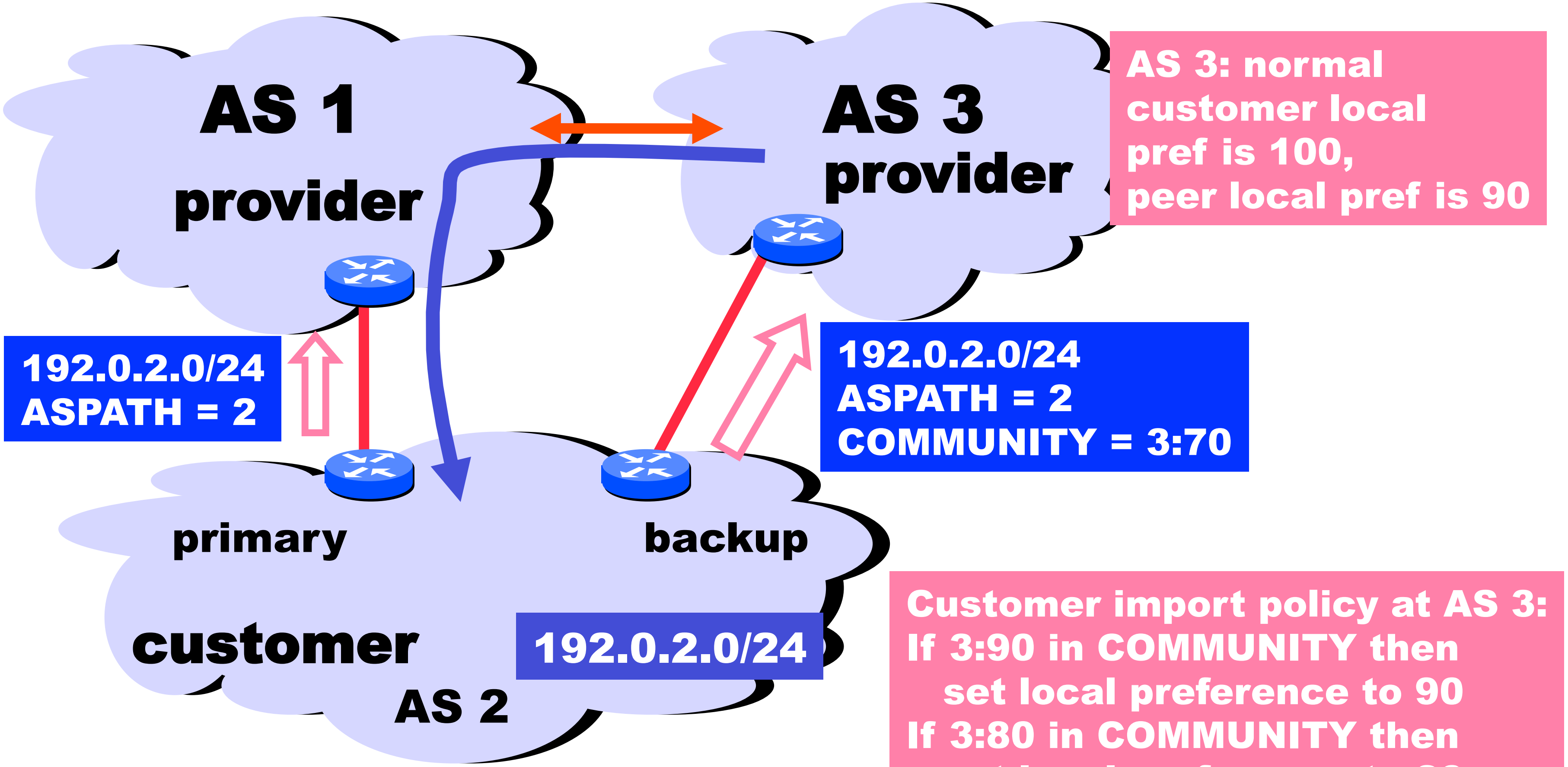
... But Padding Does Not Always Work



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

COMMUNITY Attribute to the Rescue!



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

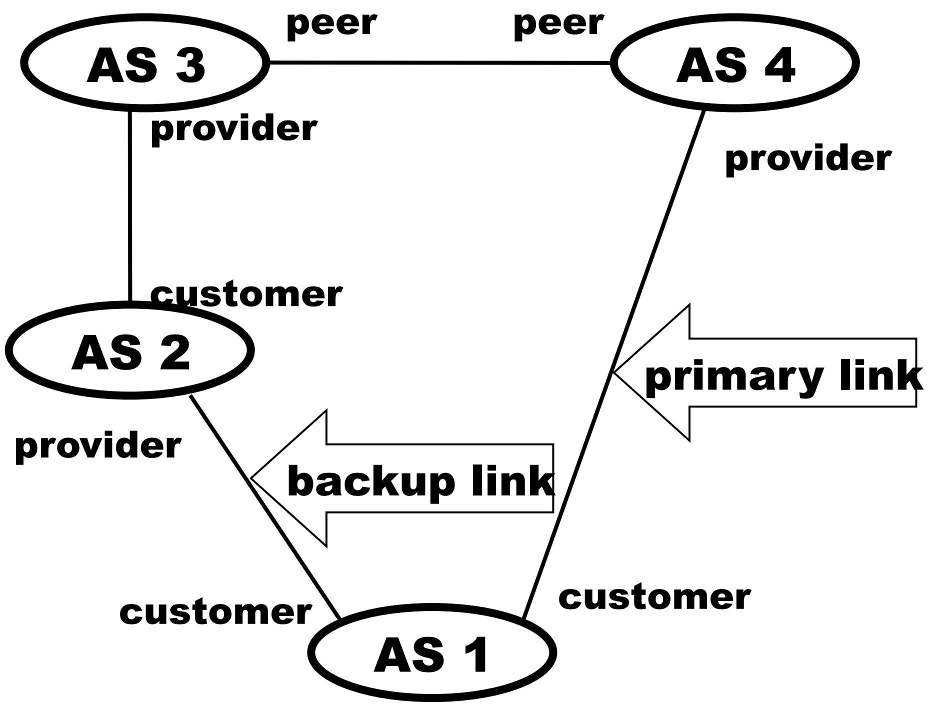
What is a BGP Wedgie (RFC 4264)?

full wedgie

¾ wedgie

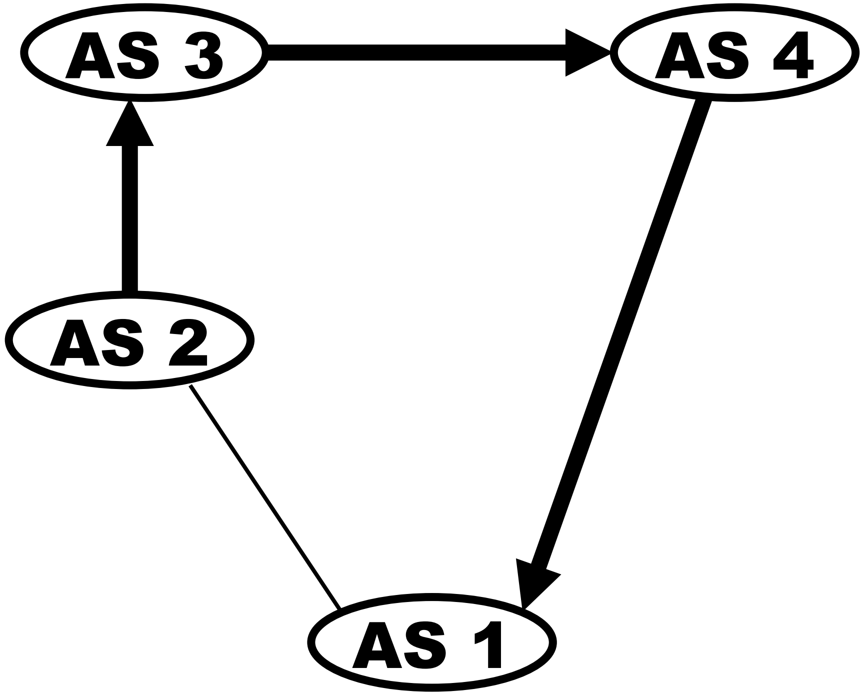
- BGP policies make sense locally
- Interaction of local policies allows multiple stable routings
- Some routings are consistent with intended policies, and some are not
  - If an unintended routing is installed (BGP is "wedged"), then manual intervention is needed to change to an intended routing
- When an unintended routing is installed, no single group of network operators has enough knowledge to debug the problem

¾ Wedgie Example



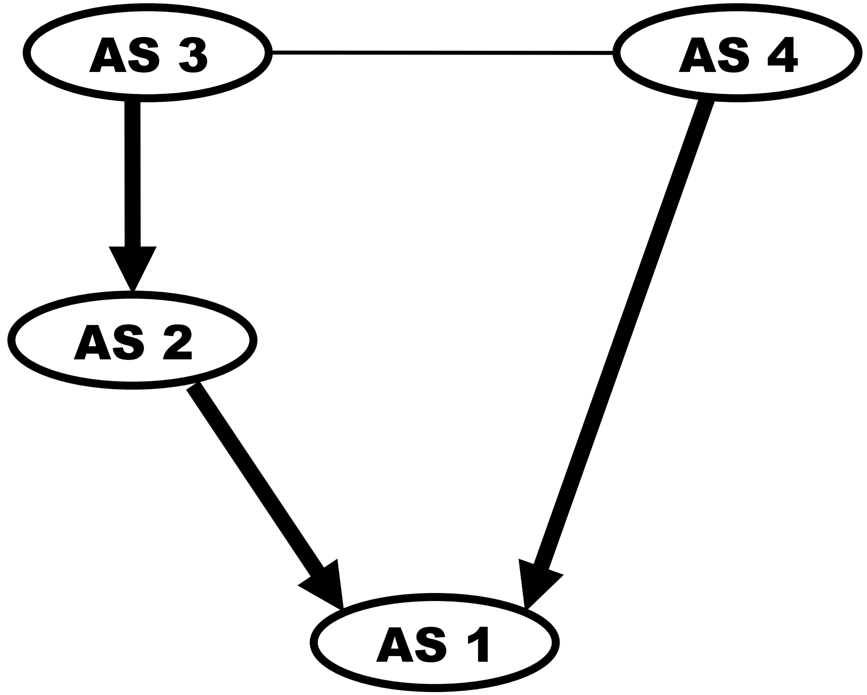
- AS 1 implements backup link by sending AS 2 a "depref me" community.
- AS 2 implements this community so that the resulting local pref is below that of routes from it's upstream provider (AS 3 routes)

And the Routings are…



Intended Routing

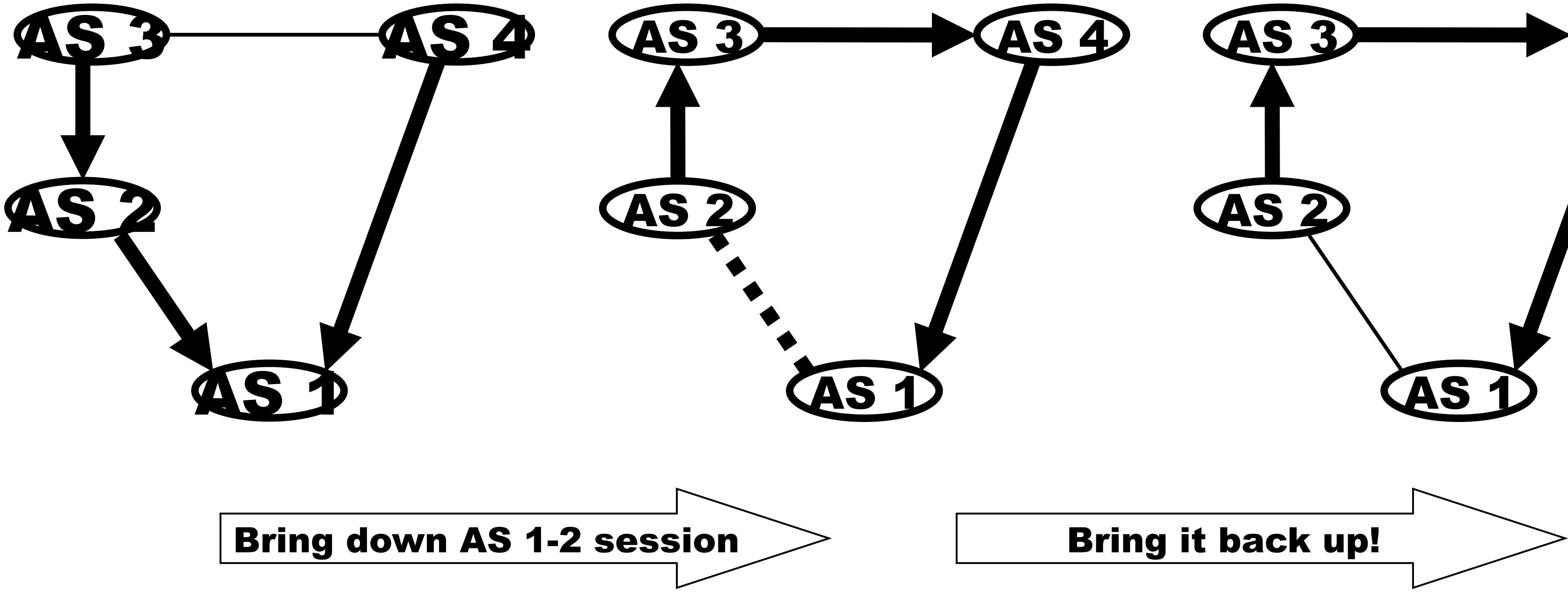
Note: this would be the ONLY routing if AS2 translated its “depref me” community to a “depref me” community of AS 3



Unintended Routing

Note: This is easy to reach from the intended routing just by “bouncing” the BGP session on the primary link.

Recovery

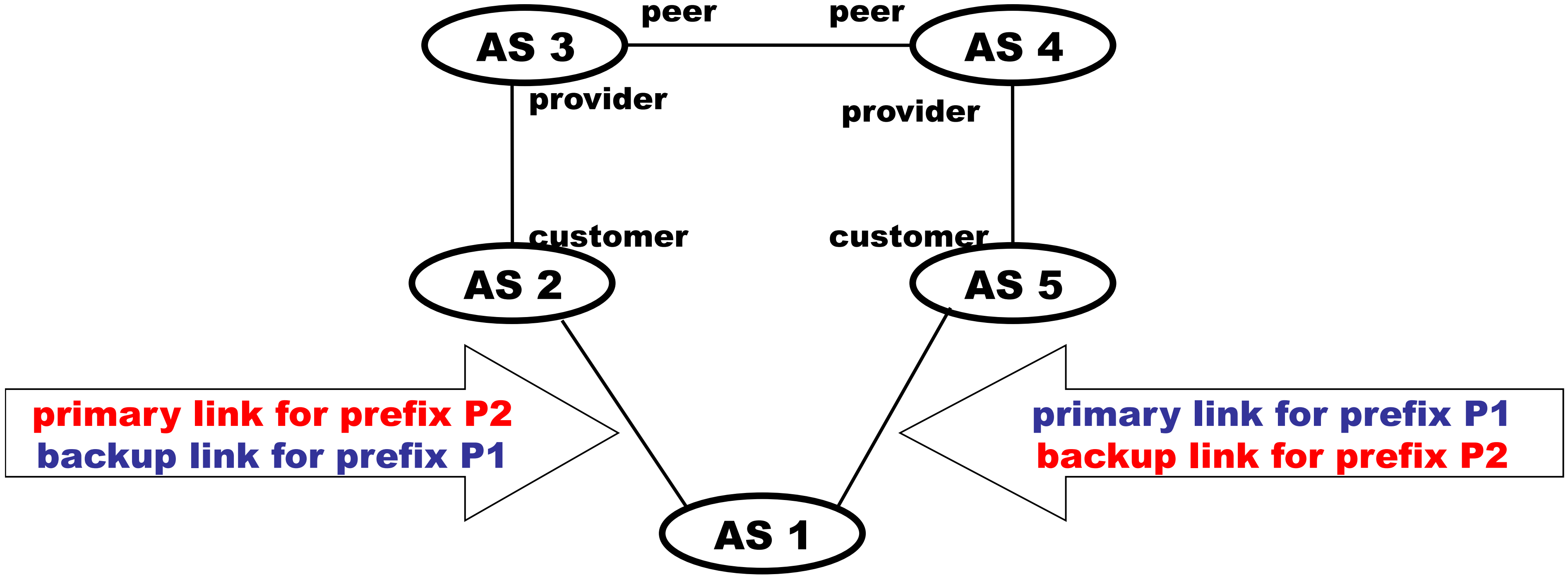


- Requires manual intervention
- Can be done in AS 1 or AS 2

What the heck is going on?

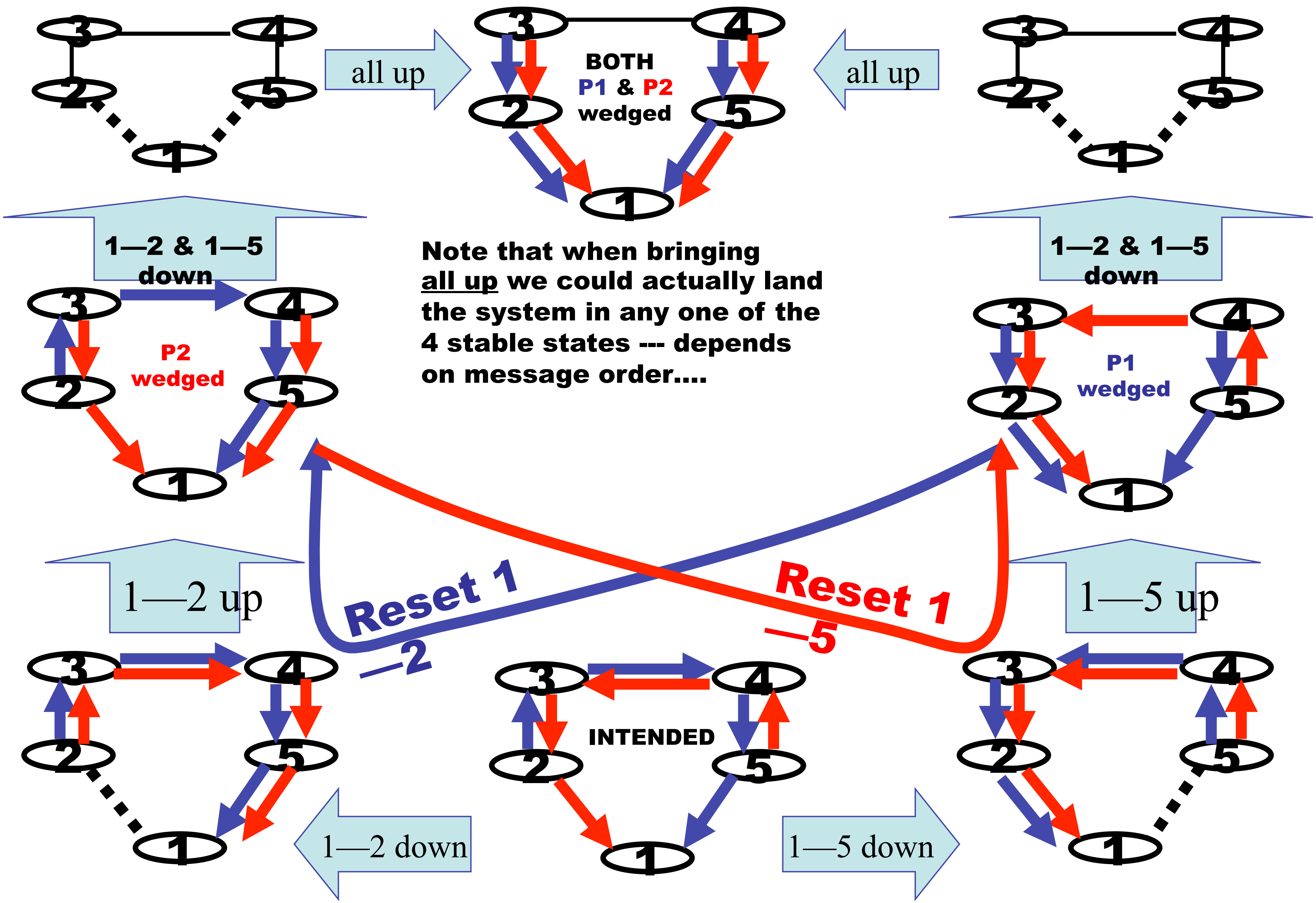
- There is no guarantee that a BGP configuration has a unique routing solution.
  - When multiple solutions exist, the (unpredictable) order of updates will determine which one is wins.
- There is no guarantee that a BGP configuration has any solution!
  - And checking configurations NP-Complete
  - Lab demonstrations of BGP configs never converging
- Complex policies (weights, communities setting preferences, and so on) increase chances of routing anomalies.
  - … yet this is the current trend!

Load Balancing Example

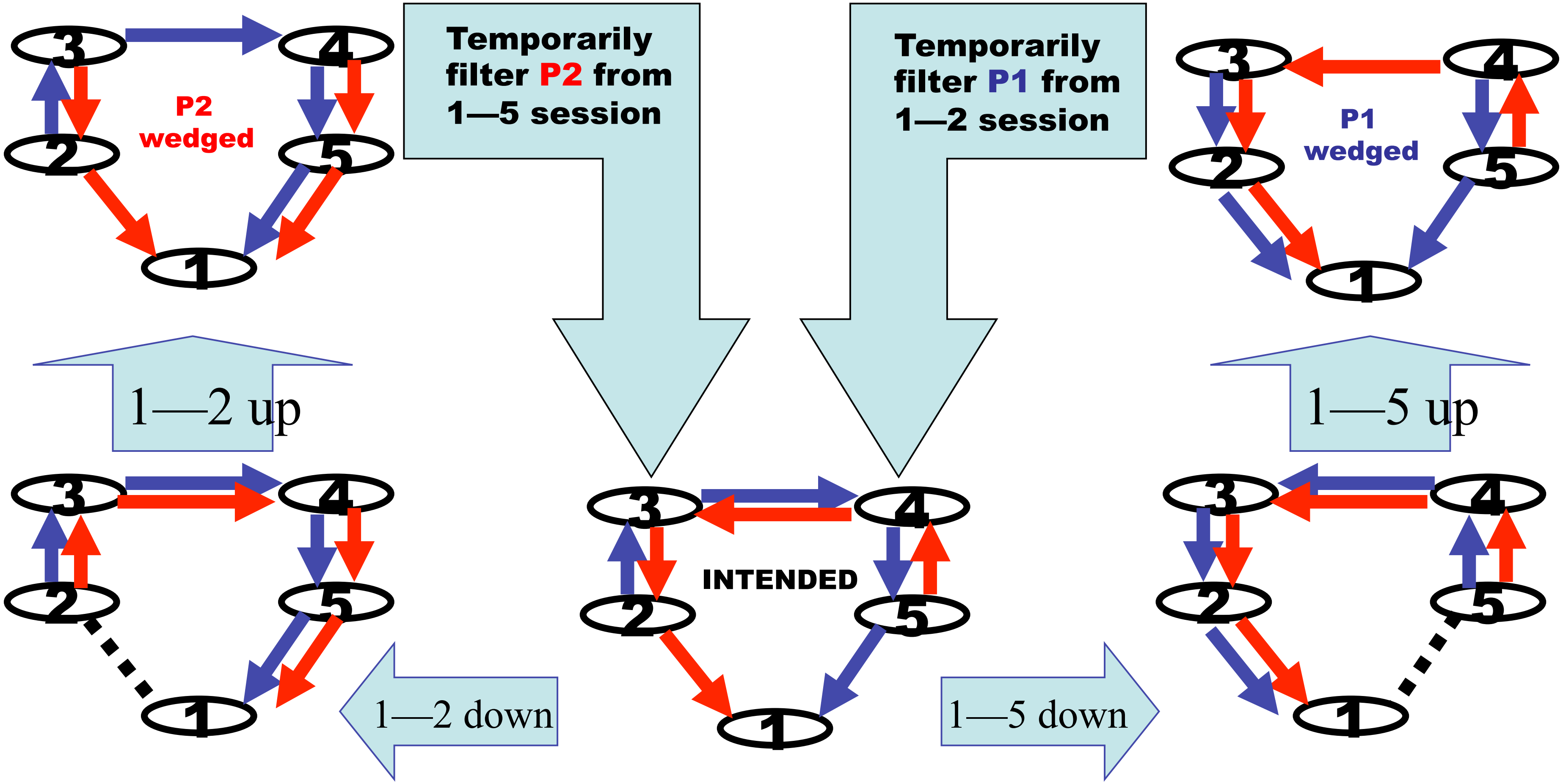


Simple session reset my not work!!

Can't un-wedge with session resets!

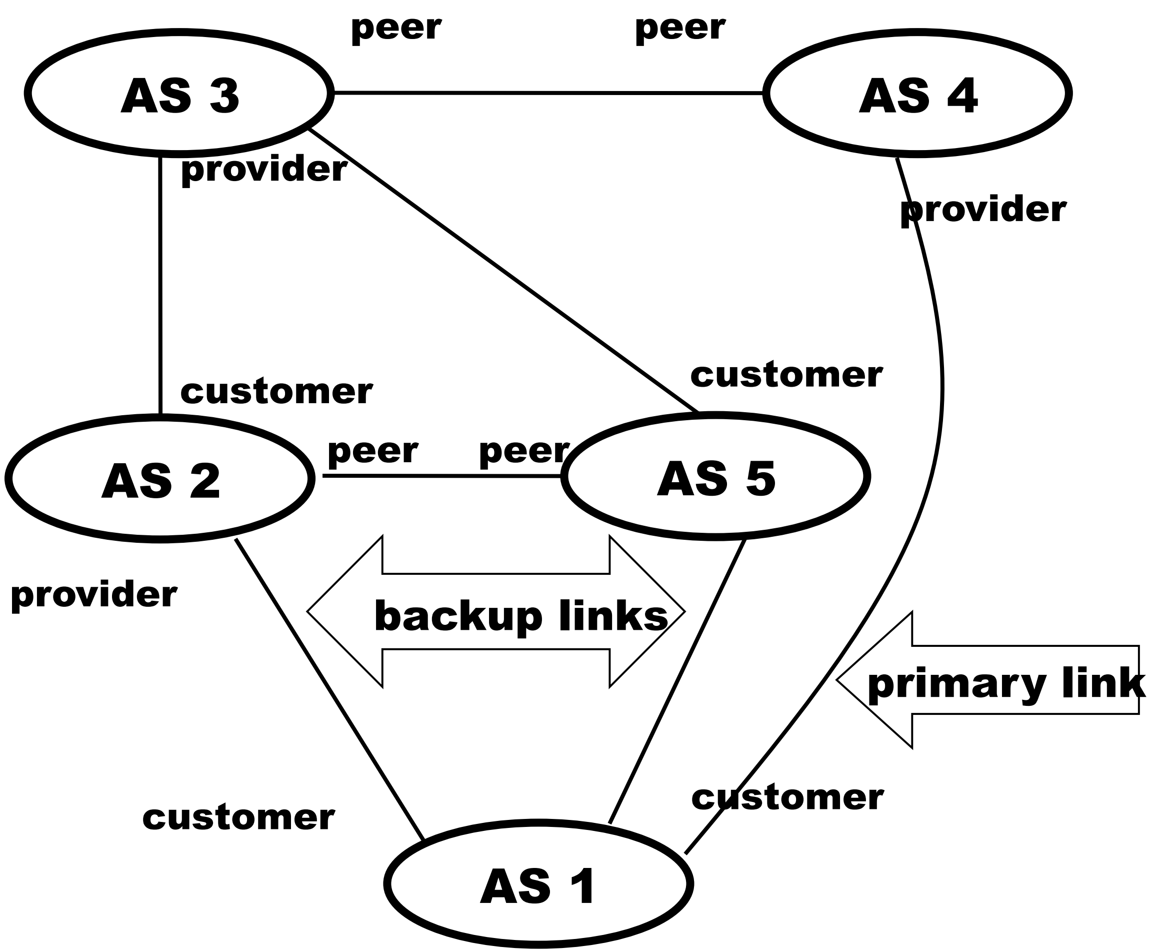


Recovery



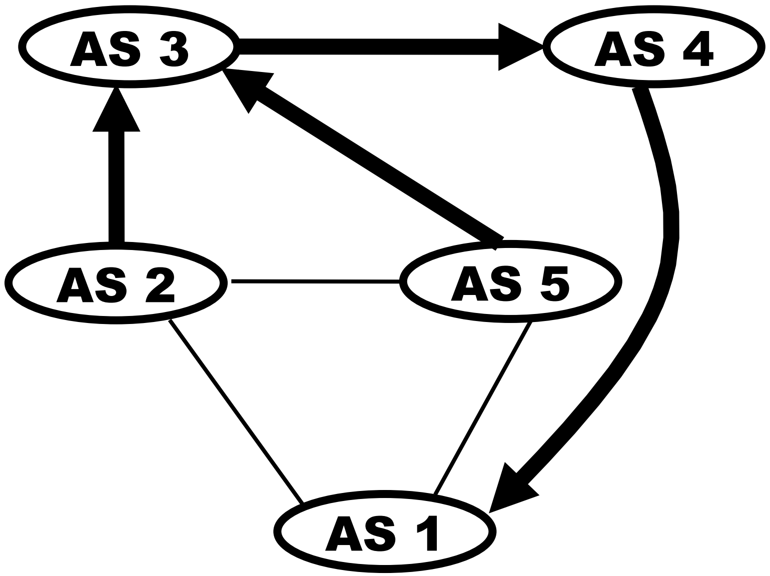
Who among us could figure this one out?
When 1—2 is in New York and 1—5 is in Tokyo?

Full Wedgie Example

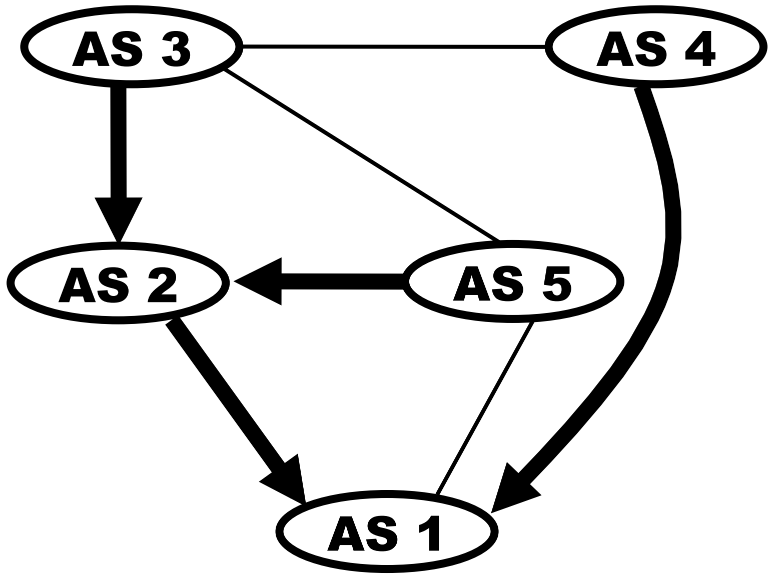


- AS 1 implements backup links by sending AS 2 and AS 3 a "depref me" communities.
- AS 2 implements its community so that the resulting local pref is below that of its upstream providers and it's peers (AS 3 and AS 5 routes)
- AS 5 implements its community so that the resulting local pref is below its peers (AS 2) but above that of its providers (AS 3)

And the Routings are...

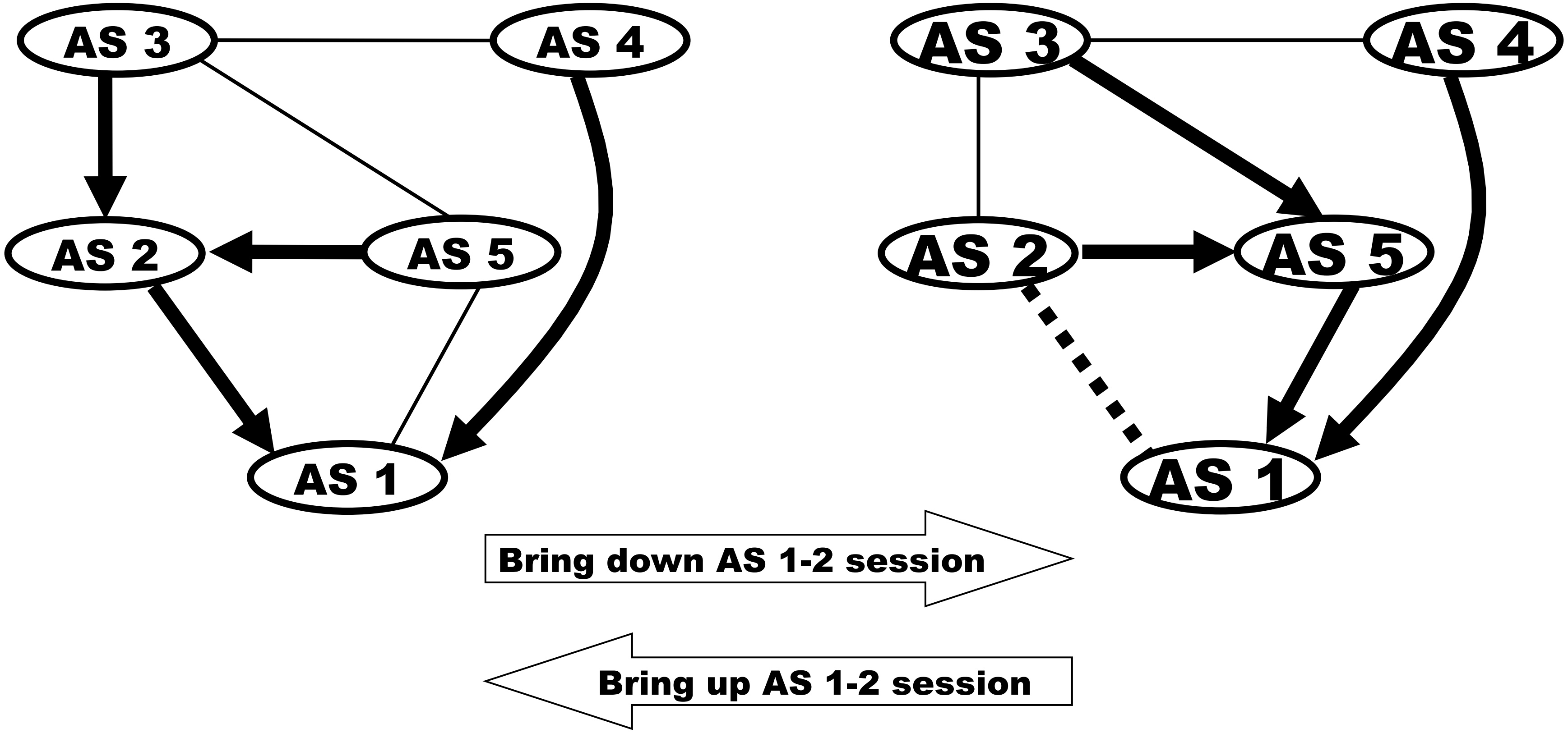


Intended Routing

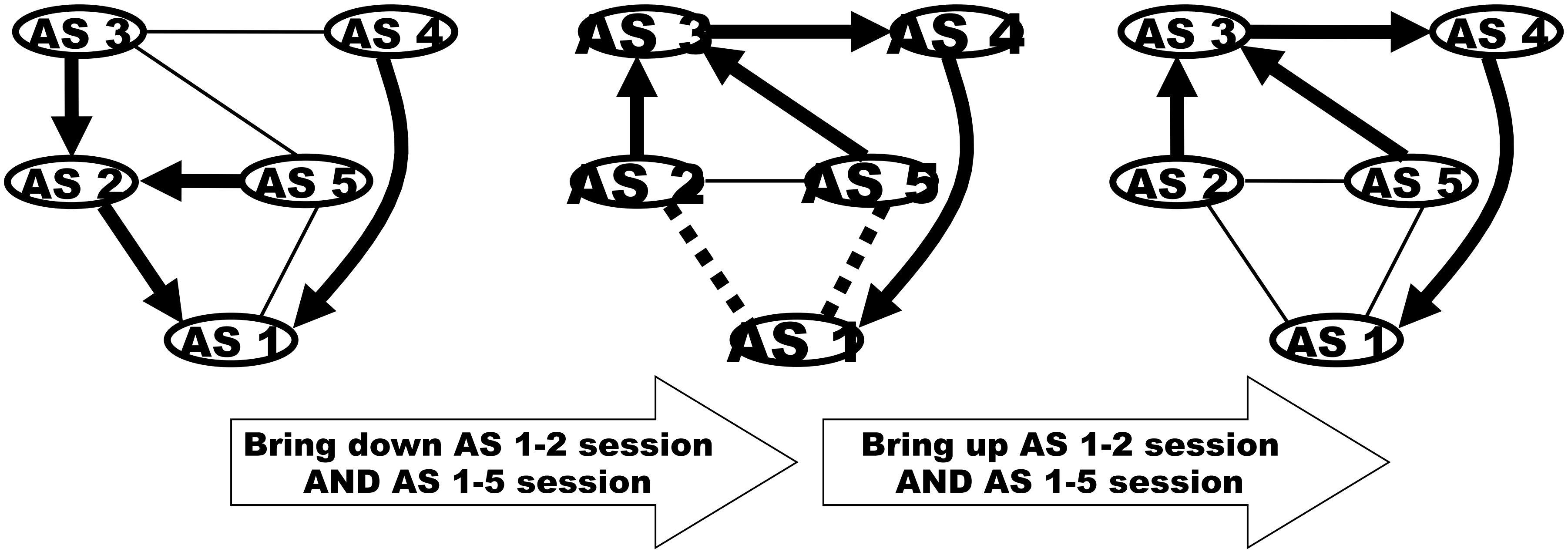


Unintended Routing

Resetting 1—2 does not help!!



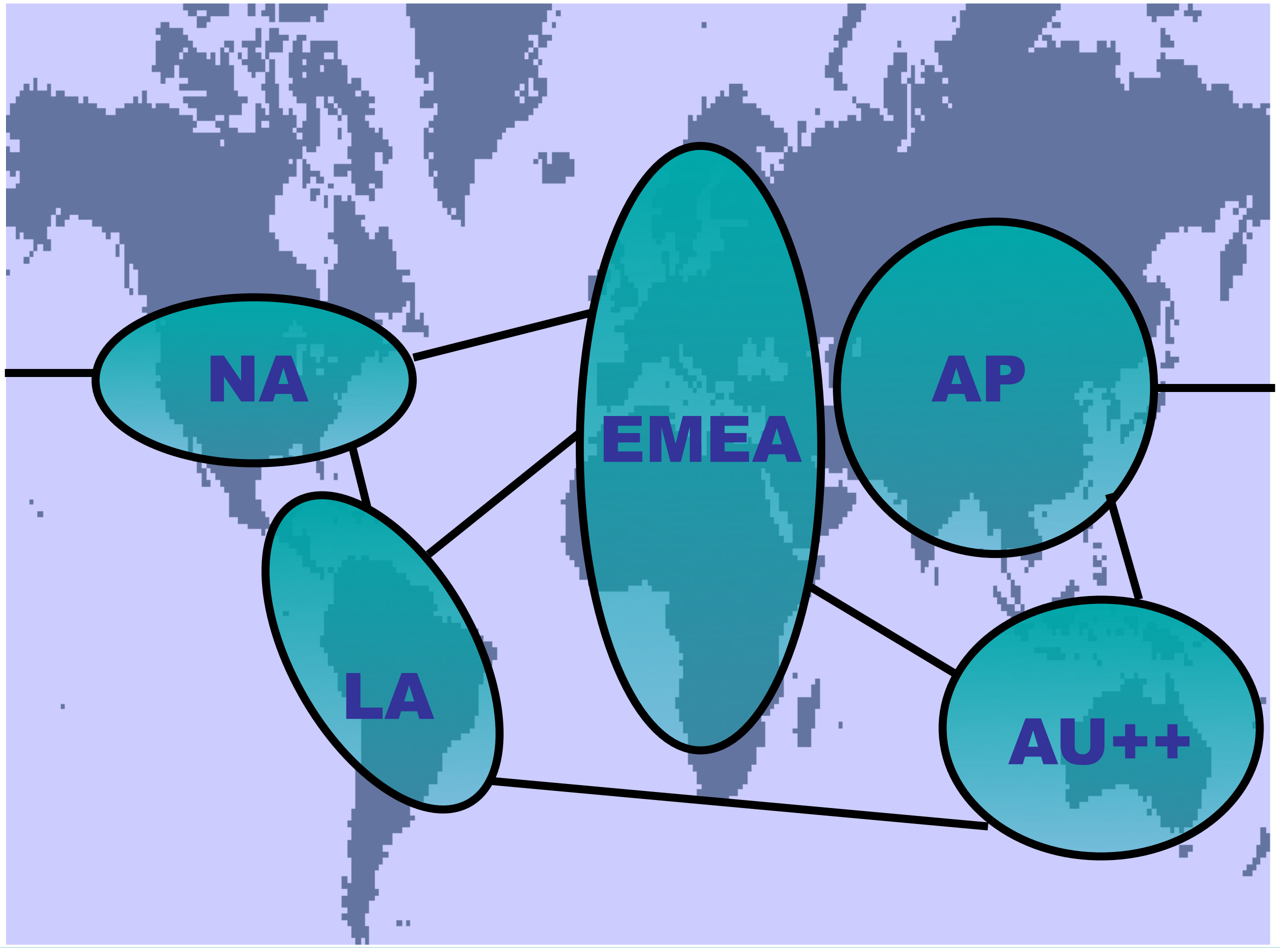
Recovery



A lot of “non-local” knowledge is required to arrive at this recovery strategy!

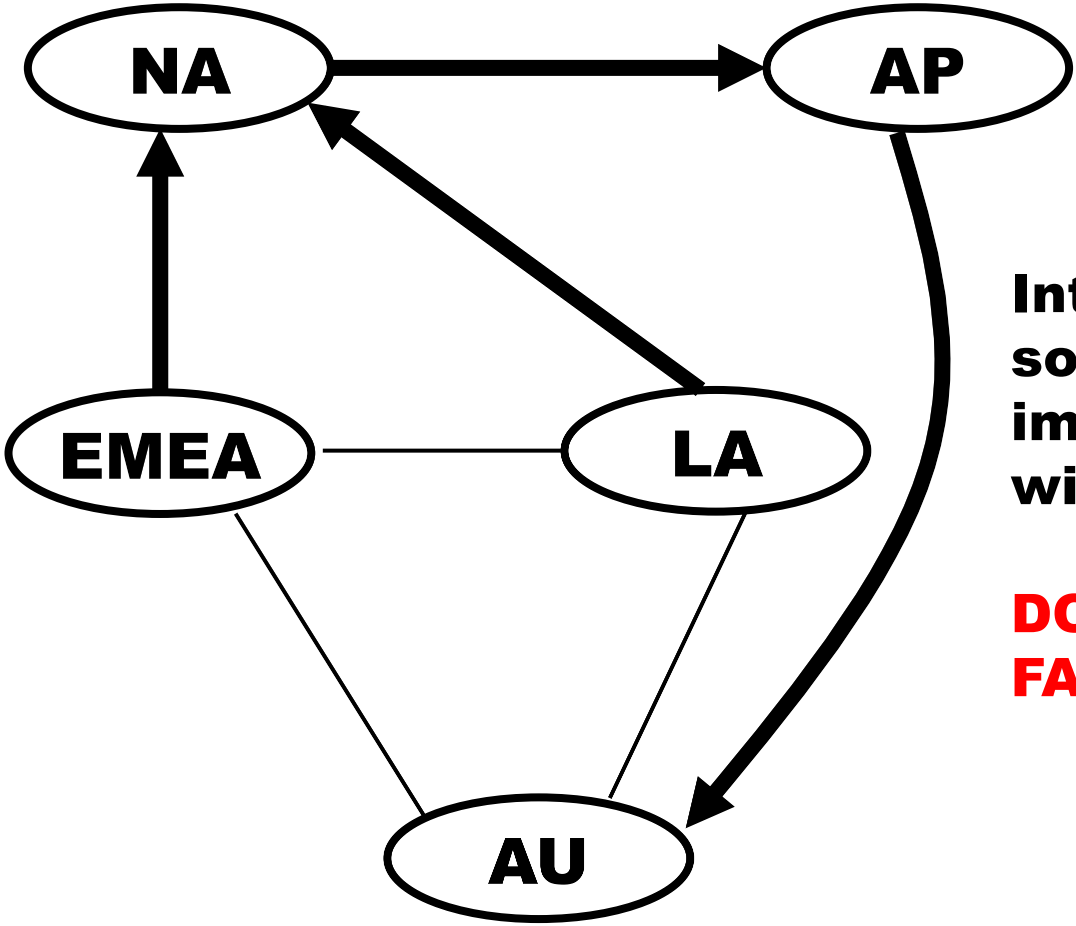
Try to convince AS 5 and AS 1 that their session has be reset (or filtered) even though it is not associated with an active route!

That Can't happen in MY network!!



An "normal" global global backbone (ISP or Corporate Intranet) implemented with 5 regional ASes

The Full Wedgie Example, in a new Guise



Intended Routing for some prefixes in AU, implemented with communities.

DOES THIS LOOK FAMILIAR??

Message: Same problems can arise with "traffic engineering" across regional networks.