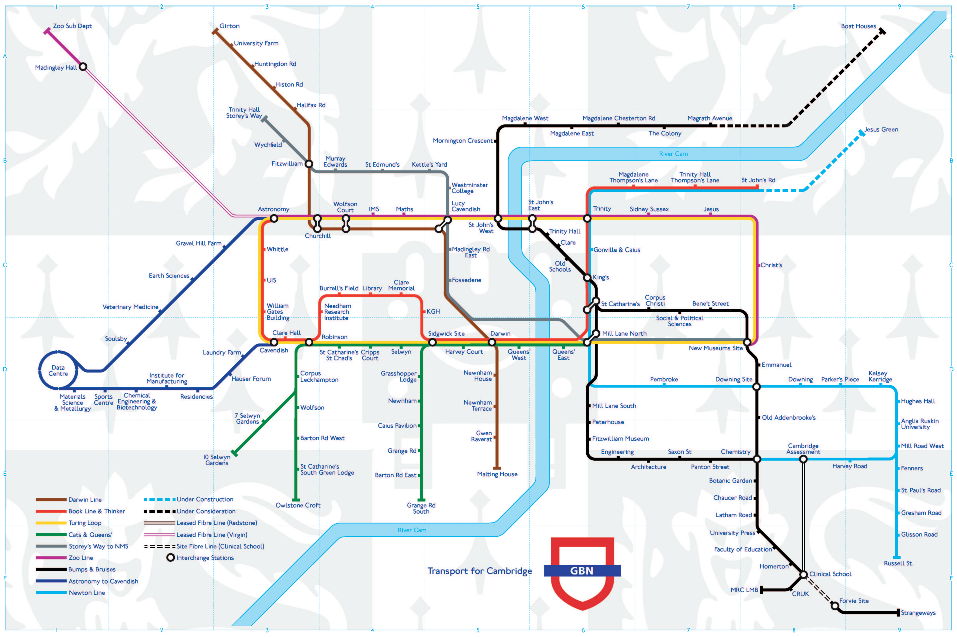
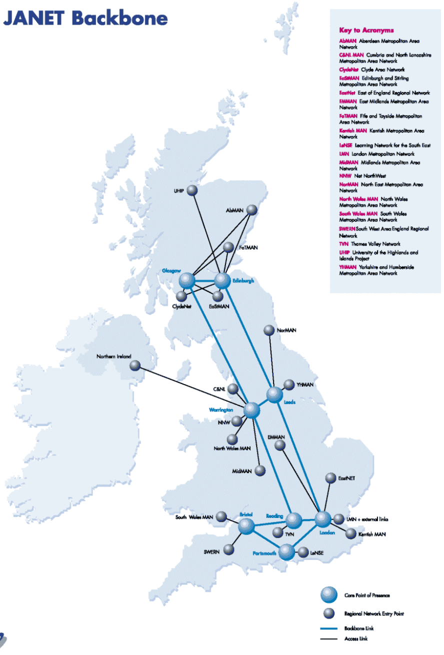
L11 : Inter-domain Routing with BGP Lecture14 Michaelmas, 2016

Timothy G. Griffin
Computer Lab
Cambridge UK



The JANET Backbone

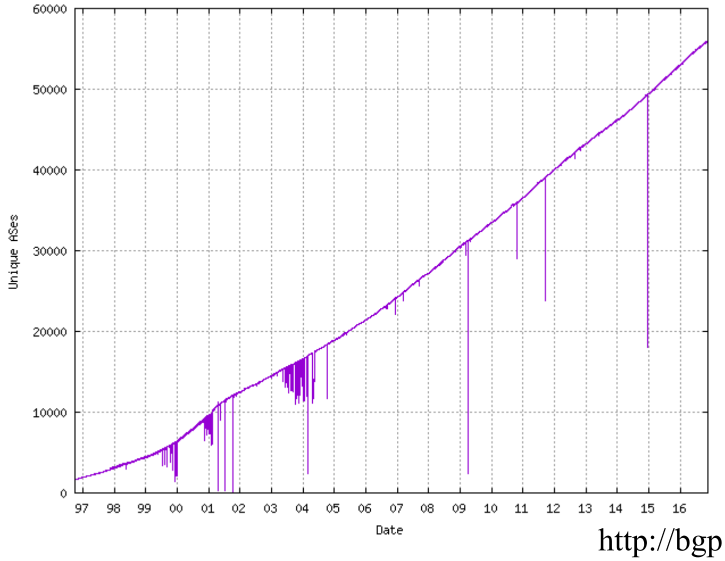


Key to Acronyms

AbMAN	Aberdeen Metropolitan Area Network
C&NL MAN	Cumbria and North Lancashire Metropolitan Area Network
ClydeNet	Clyde Area Network
EaStMAN	Edinburgh and Stirling Metropolitan Area Network
EastNet	East of England Regional Network
EMMAN	East Midlands Metropolitan Area Network
FaTMAN	Fife and Tayside Metropolitan Area Network
Kentish MAN	Kentish Metropolitan Area Network
LeNSE	Learning Network for the South East
LMN	London Metropolitan Network
MidMAN	Midlands Metropolitan Area Network
NNW	Net NorthWest
NorMAN	North East Metropolitan Area Network
North Wales MAN	North Wales Metropolitan Area Network
South Wales MAN	South Wales Metropolitan Area Network
SWERN	South West Area England Regional Network
TVN	Thames Valley Network
UHIP	University of the Highlands and Islands Project
YHMAN	Yorkshire and Humberside Metropolitan Area Network

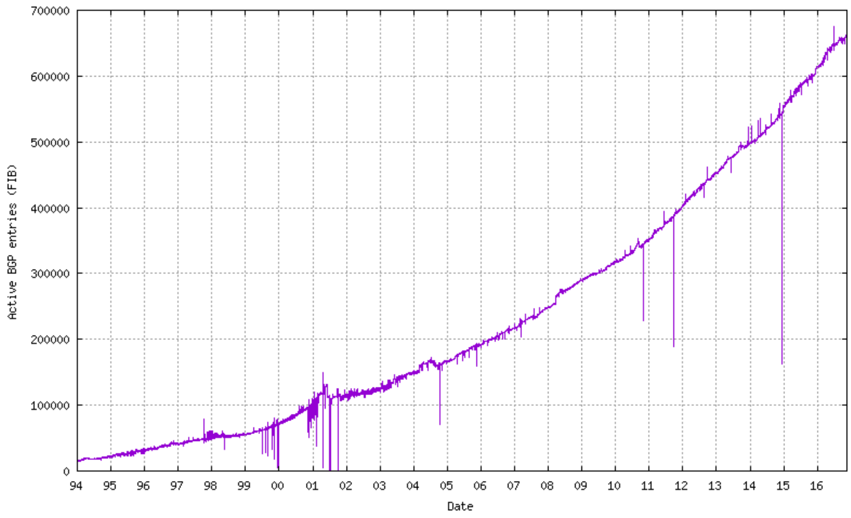
JANET

How many ASNs today (17 November, 2016)?



http://bgp.potaroo.net/

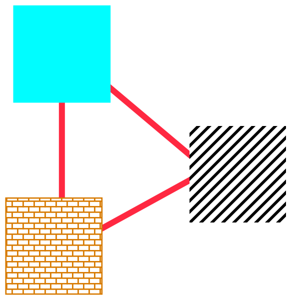
How many prefixes today (17 November, 2016)?

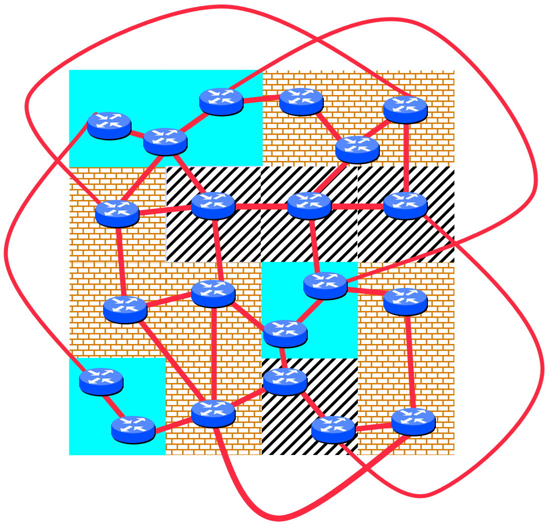


http://bgp.potaroo.net/

AS Graph != Internet Topology

BGP was designed to throw away information!





ICNP 2002

BGP Route 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	

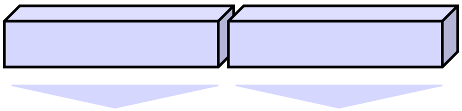
Most important attributes

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

Not all attributes need to be present in every announcement

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

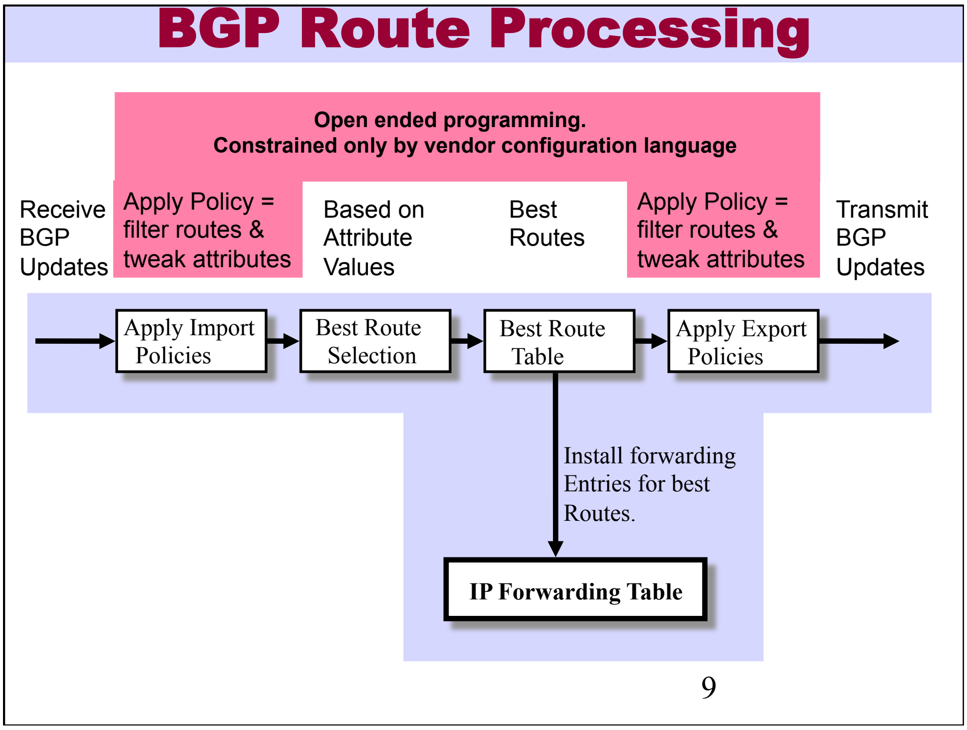
Community Attribute = a list of community values.
(So one route can belong to multiple communities)

RFC 1997 (August 1996)

Reserved communities

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

no_advertise 0xFFFFFF02: don't pass to BGP neighbors



BGP Route Processing

Open ended programming.
Constrained only by vendor configuration language

Receive BGP Updates

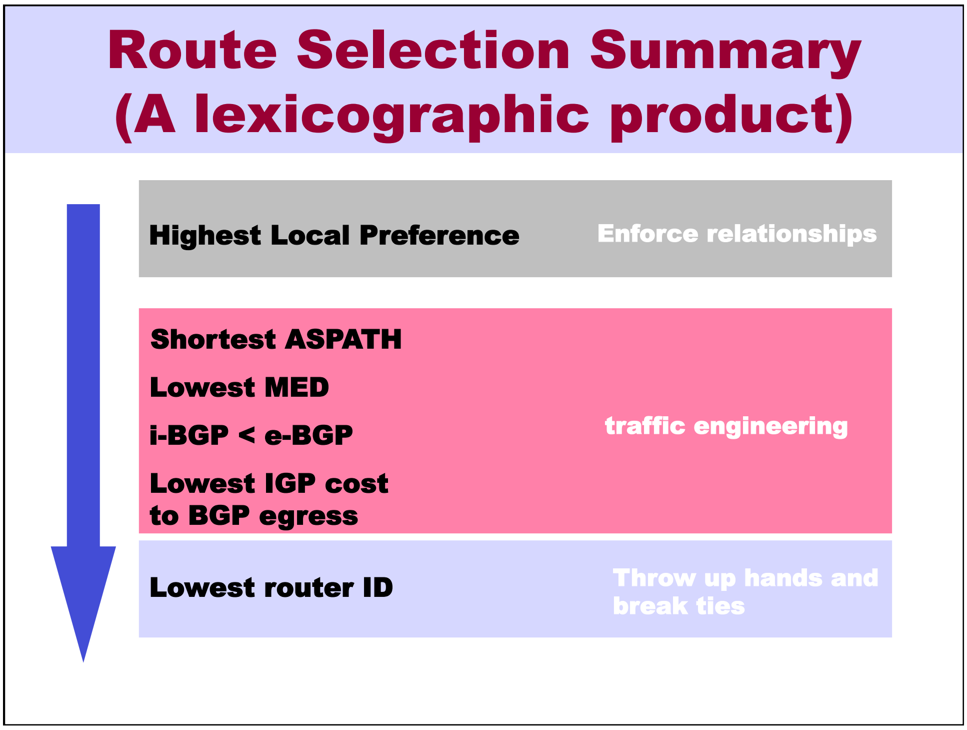
Apply Policy = filter routes & tweak attributes

Based on Attribute Values

Best Routes

Apply Policy = filter routes & tweak attributes

Transmit BGP Updates

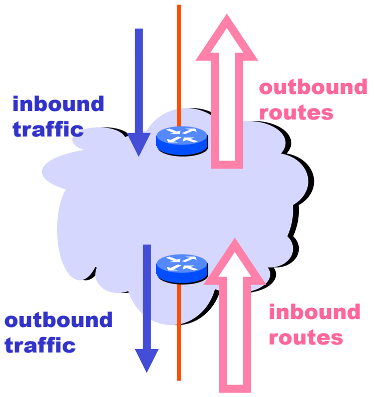


Route Selection Summary (A lexicographic product)

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**

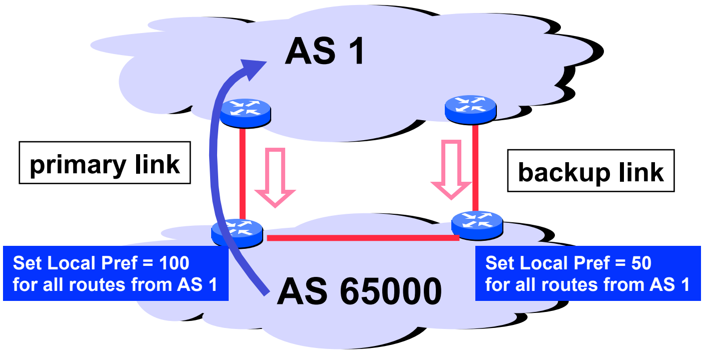
Traffic Engineering

- 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)



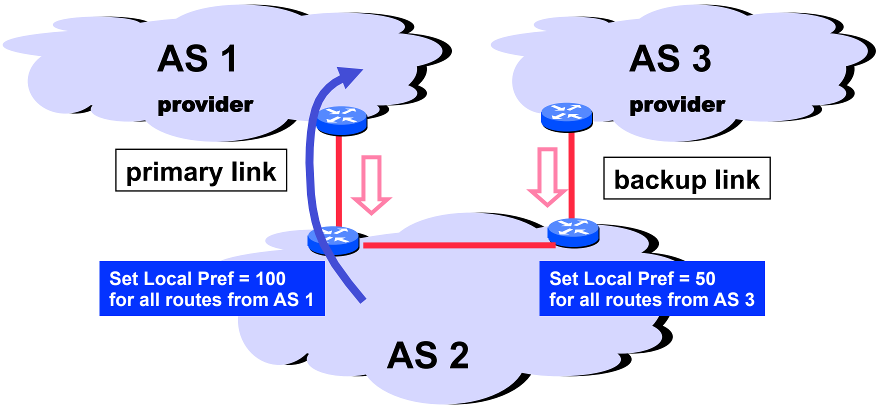
Set Local Pref = 100 for all routes from AS 1

Set Local Pref = 50 for all routes from AS 1

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

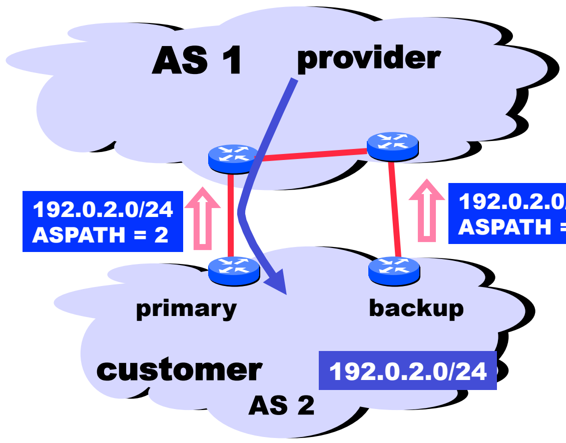
We'll talk about inbound traffic soon …

Multihomed Backups (Outbound Traffic)



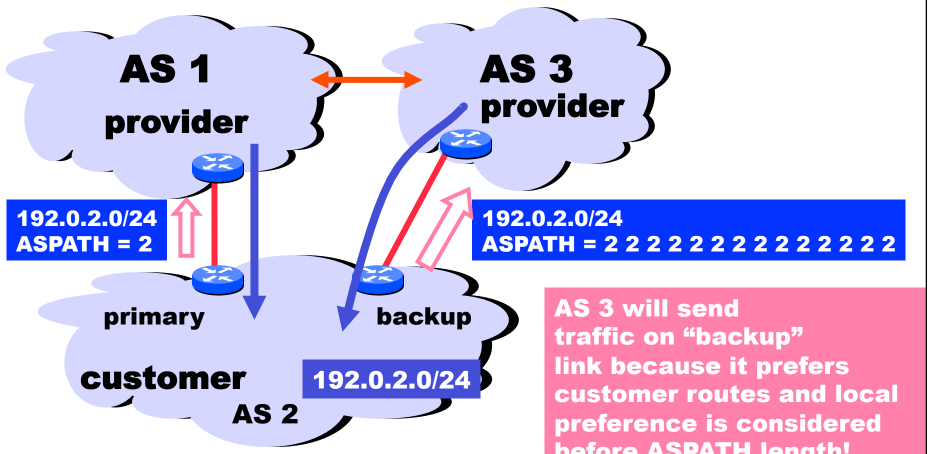
Forces outbound traffic to take primary link, unless link is down.

Shedding Inbound Traffic with ASPATH Padding. Yes, this is an ugly hack …



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

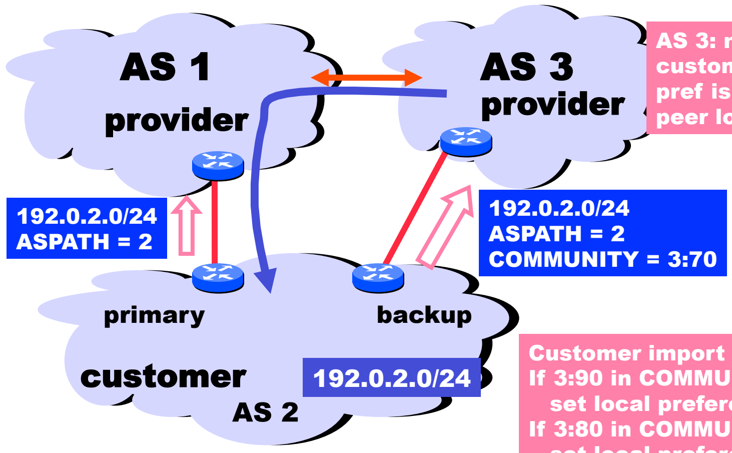
… But Padding Does Not Always Work



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!



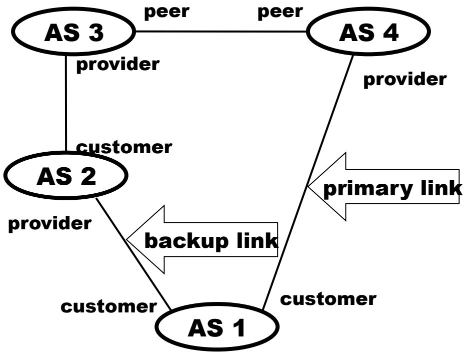
AS 3: normal customer local pref is 100, peer local pref is 90

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

Problem scenario with inter-domain routing in the Internet (BGP)

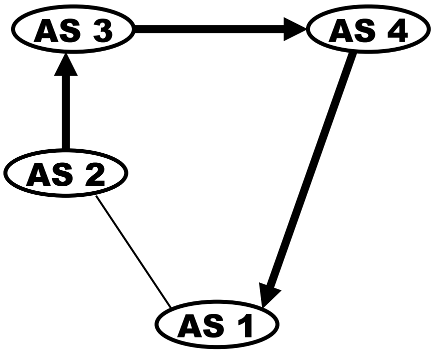
- BGP policies make sense locally
- Interaction of local policies allows multiple stable routings
- Some routings are consistent with intended policies, and some are not
  - If an unintended routing is installed (BGP is “wedged”), then manual intervention is needed to change to an intended routing
-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

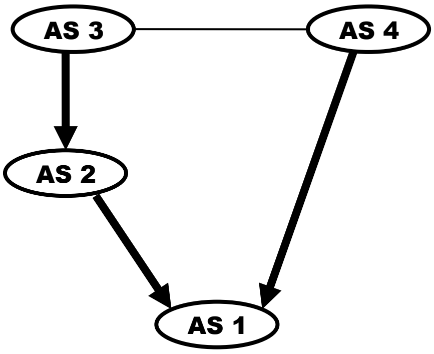
Simple Example



- 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…





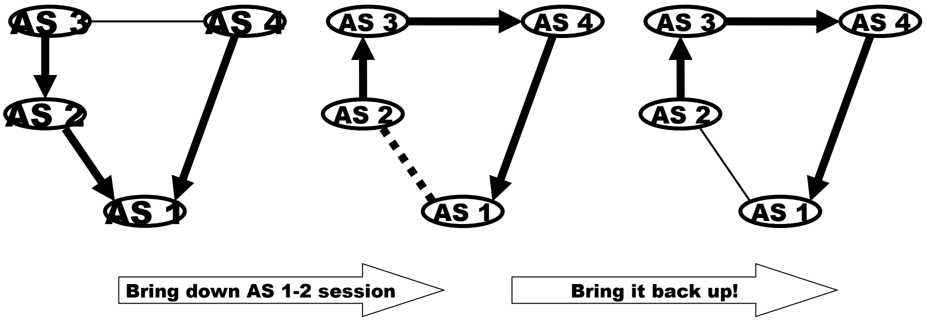
Intended Routing

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

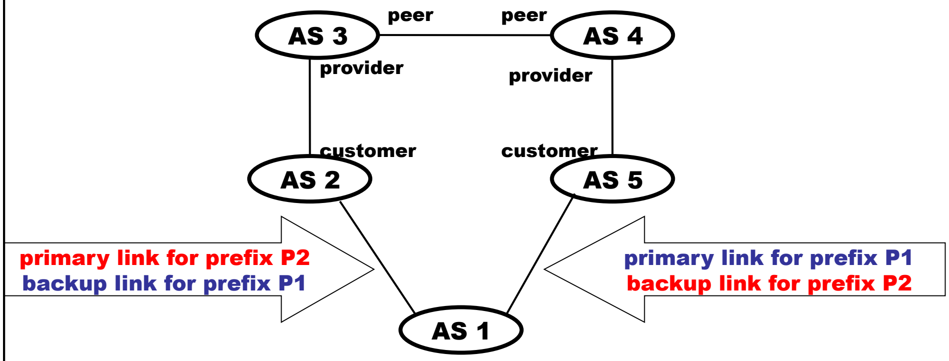


- Requires manual intervention
- Can be done in AS 1 or AS 2

What is going on?

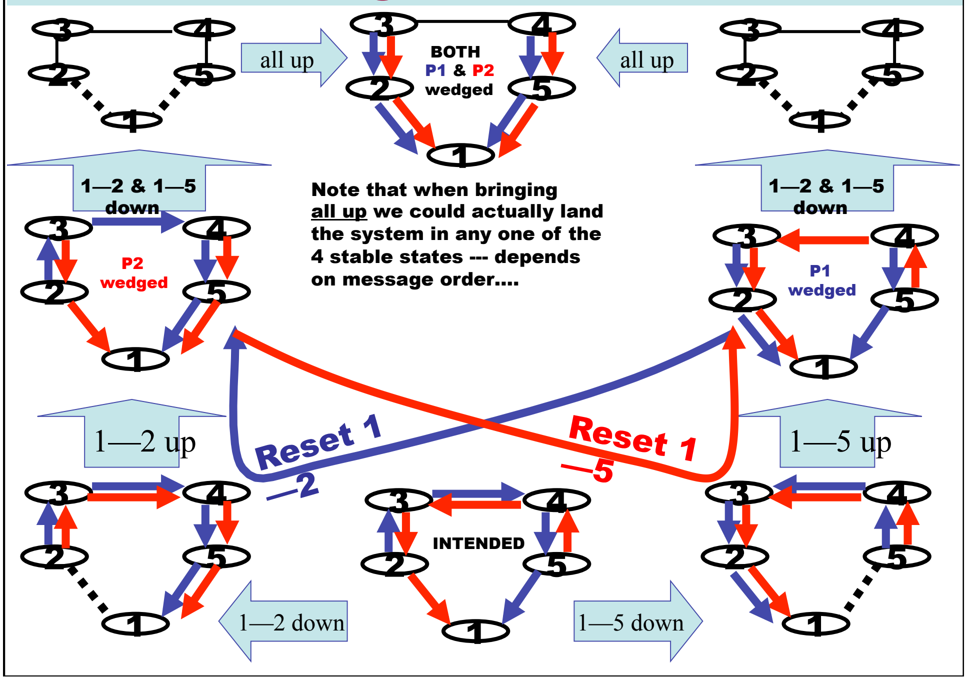
- There is no guarantee that a BGP configuration has a unique routing solution.
  - When multiple solutions exist, the (unpredictable) order of updates will determine which one is wins.
- There is no guarantee that a BGP configuration has any solution!
  - And checking configurations NP-Complete
- Complex policies (weights, communities setting preferences, and so on) increase chances of routing anomalies.
  - … yet this is the current trend!

Load Balancing Example


AS 3
peer
peer
AS 4
provider
provider
customer
customer
AS 2
AS 5
primary link for prefix P2
backup link for prefix P1
primary link for prefix P1
backup link for prefix P2
AS 1

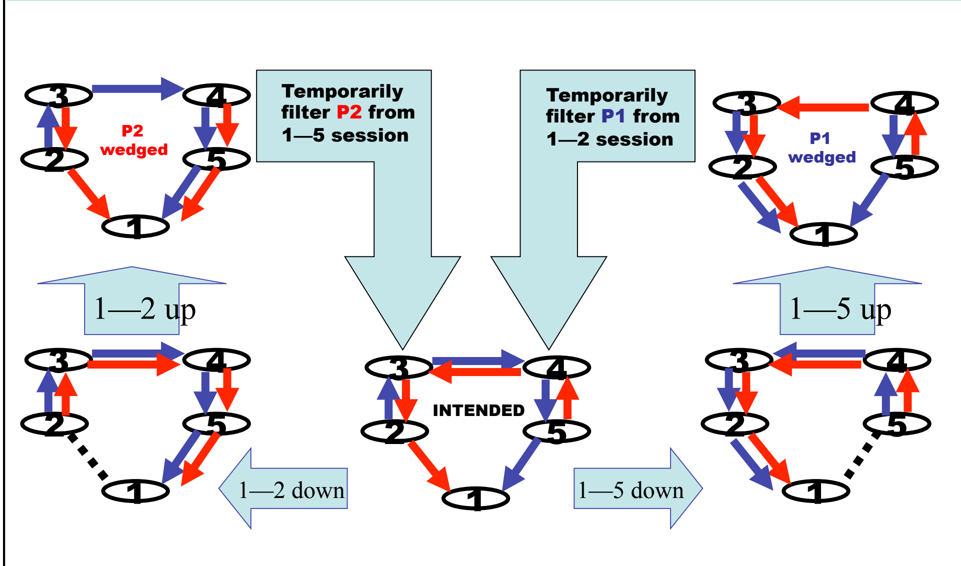
Simple session reset my not work!!

Can’t un-wedge with session resets!



Note that when bringing **all up** we could actually land the system in any one of the 4 stable states --- depends on message order….

Recovery

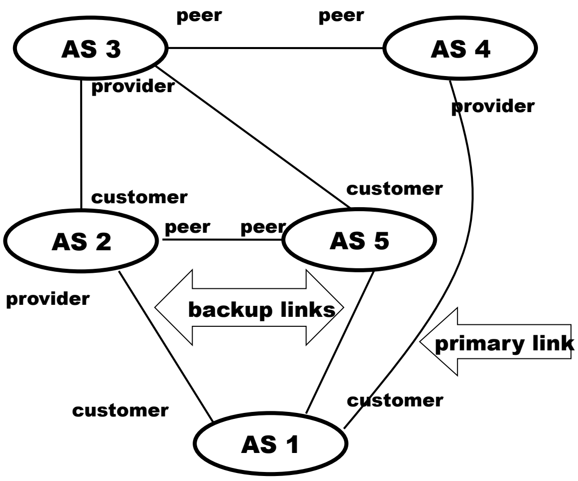


Temporarily filter P2 from 1—5 session

Temporarily filter P1 from 1—2 session

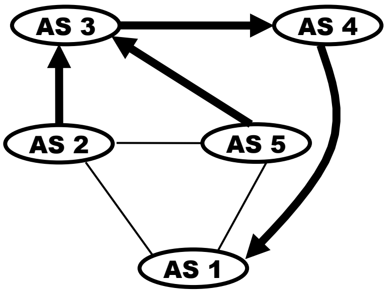
Who among us could figure this one out?
When 1—2 is in New York and 1—5 is in Tokyo?

Advanced Example

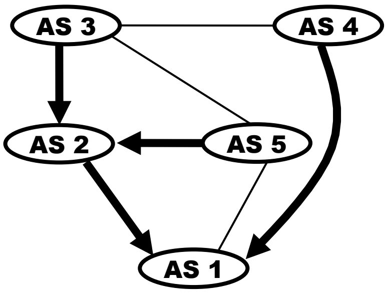


- AS 1 implements backup links by sending AS 2 and AS 3 a “depref me” communities.
- AS 2 implements its community so that the resulting local pref is below that of its upstream providers and it’s peers (AS 3 and AS 5 routes)
- AS 5 implements its community so that the resulting local pref is below its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

And the Routings are…

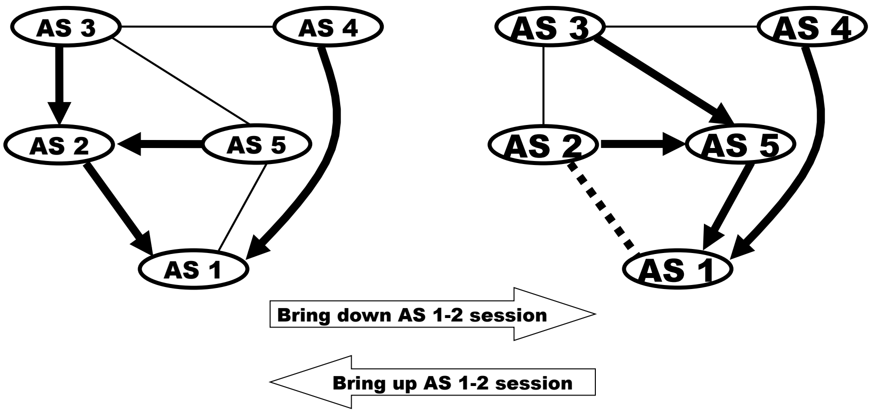


Intended Routing



Unintended Routing

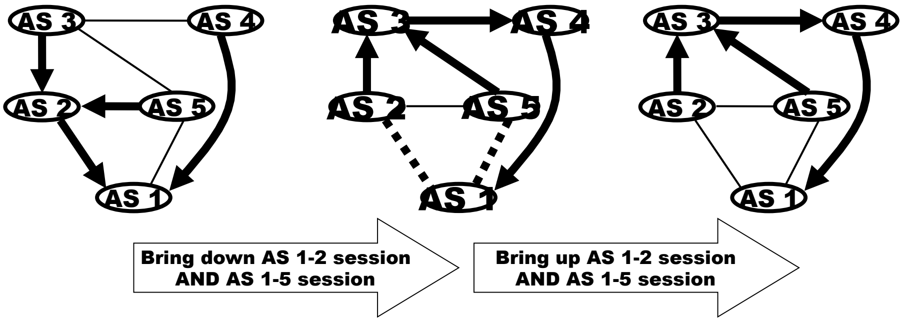
Resetting 1—2 does not help!!



Bring down AS 1-2 session

Bring up AS 1-2 session

Recovery



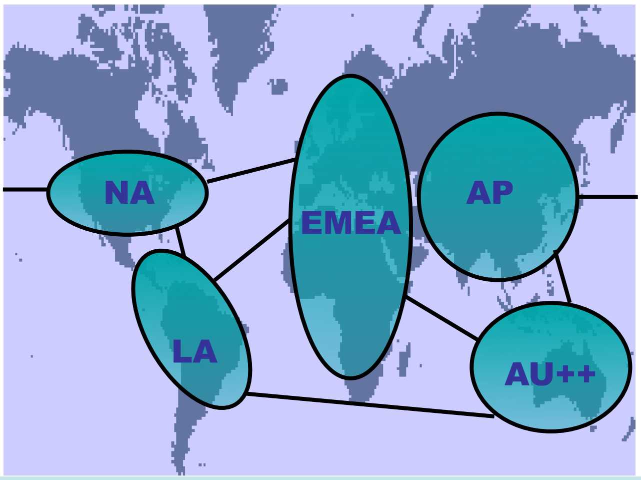
Bring down AS 1-2 session
AND AS 1-5 session

Bring up AS 1-2 session
AND AS 1-5 session

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

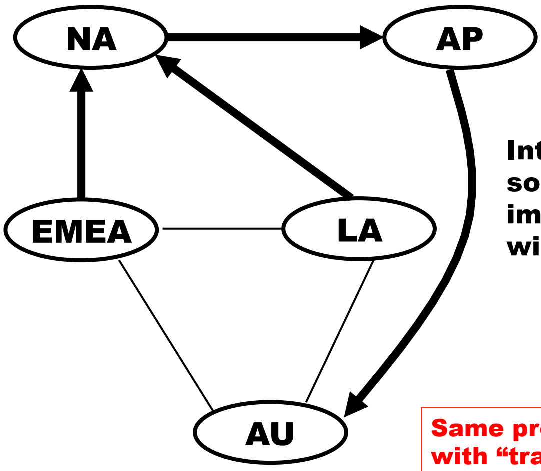
Try to convince AS 5 and AS 1 that their session has be reset (or filtered) even though it is not associated with an active route!

That Can’t happen in MY network!!



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

Does this look familiar?



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

Same problems can arise with “traffic engineering” across regional networks.