Two Types of BGP Sessions

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

IBGP is routed (using IGP)
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.
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!

Routers make independent selections!

192.44.78.0/24

AS 1

AS 2

AS 3

AS 4

192.44.78.0/24
ASPATH = 4 2 1

192.44.78.0/24
ASPATH = 4 3 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
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.

- Set Local Pref = 100 for all routes from AS 1
- Set Local Pref = 50 for all routes from AS 3

**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

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

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

192.0.2.0/24 ASPATH = 2

Customer

AS 2

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.
What is a BGP Wedgie? [RFC 4264]

- 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

Half 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 its upstream provider (AS 3 routes)
And the Routings are...

Intended Routing

AS 1 → AS 2 → AS 3 → AS 4

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

Unintended Routing

AS 1 → AS 4

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

Recovery

Bring down AS 1-2 session

Bring it back up!

• 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!

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?
**Full Wedgie Example**

- AS 1 implements backup links by sending AS 2 and AS 5 a “depref me” communities.
- AS 2 implements its community so that the resulting local pref is below that of its upstream provider’s and its 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

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.