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

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

Mr. BGP says that path 4 1 is better than path 3 2 1.

Duh!
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?
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 = 0xFFFFFFFF01: don’t export out of AS
- no_advertise 0xFFFFFFFF02: 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

---

**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

High bandwidth
Provider backbone

2865

17

Heavy
Content
Web Farm

Low bandwidth
customer backbone

15

56

San Diego

NYC

SFF

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

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 AS PATH Padding. Yes, this is a Glorious Hack ...
But Padding Does Not Always Work

AS 1
provider

192.0.2.0/24
ASPATH = 2

primary

customer
192.0.2.0/24

AS 2

AS 3
provider

192.0.2.0/24
ASPATH = 2 2 2 2 2 2 2 2 2 2 2 2 2 2

backup

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 1
provider

192.0.2.0/24
ASPATH = 2

primary

customer
192.0.2.0/24

AS 2

AS 3
provider

192.0.2.0/24
ASPATH = 2

backup

192.0.2.0/24
COMMUNITY = 3:70

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
BGP Dynamics

- How many updates are flying around the Internet?
- How long Does it take Routes to Change?

The goals of
(1) fast convergence
(2) minimal updates
(3) path redundancy
are at odds.

Pick any two!!

Hourly Average of Per-Second Updated and Withdrawn Prefix Rate

http://bgpupdates.potaroo.net

Jan 26 2009
Hourly Peak of Per-Second Updated and Withdrawn Prefix Rate

Results will vary depending on location...

Q: Why All the Updates?

- The Internet is large, so isn’t there always something going on somewhere? (That is, BGP is just doing a good job of keeping things connected!)
- Is BGP exploring many alternate paths during convergence?
- Are IGP instabilities being exported to the interdomain world?
- Have bad tradeoffs been made in router software implementation?
- Are BGP sessions being reset due to congestion?
- Weird policy interactions like MED oscillation?
- Gnomes, sprites, and fairies
- ....

A: NO ONE REALLY KNOWS ...
BGP does a very good job hiding information!
Routing Change: Path Exploration

- Initial situation
  - Destination 0 is alive
  - All ASes use direct path

- When destination dies
  - All ASes lose direct path
  - All switch to longer paths
  - Eventually withdrawn

- E.g., AS 2
  - \((2,0) \Rightarrow (2,1,0)\)
  - \((2,1,0) \Rightarrow (2,3,0)\)
  - \((2,3,0) \Rightarrow (2,1,3,0)\)
  - \((2,1,3,0) \Rightarrow \text{null}\)

IGP Tie Breaking Can Export Internal Instability to the Whole Wide World
MEDs Can Export Internal Instability

Implementation Does Matter!

Thanks to Abha Ahuja and Craig Labovitz for this plot.
Conjestion can take down BGP sessions! The SQL Slammer worm

Two BGP Mechanisms for Squashing Updates

- Rate limiting on sending updates
  - Send batch of updates every MinRouteAdvertisementInterval (MRAI) seconds (+/- random fuzz)
  - Default value is 30 seconds
  - A router can change its mind about best routes many times within this interval without telling neighbors

- Route Flap Dampening
  - Punish routes for “misbehaving”
Route Flap Dampening (RFC 2439)

Routes are given a penalty for changing. If penalty exceeds suppress limit, the route is dampened. When the route is not changing, its penalty decays exponentially. If the penalty goes below reuse limit, then it is announced again.

- Can dramatically reduce the number of BGP updates
- Requires additional router resources

How it works

Route Flap Dampening (half-life = 25 minutes) — 6 flaps, one every 10 minutes

penalty for each flap = 1000
Problems with Flap Damping:
punishes small updates for
“well connected” destinations

The prefix is not “misbehaving” -- it is BGP!!

Rate Limiting

MRAI = Minimum Router Advertisement Interval (in seconds)

<table>
<thead>
<tr>
<th>MRAI seconds</th>
<th>Your Best Path</th>
<th>what you send to your neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>889 736 11</td>
<td>Announce 555 67 11</td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>458 11</td>
<td>Withdraw</td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>889 736 11</td>
<td>Announce 7018 555 67 11</td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>458 11</td>
<td>Keep quite!</td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
</tr>
</tbody>
</table>

With quite!
30 Second Bursts

Updates often come in bursts, about every 30 seconds, June 23 2001 (data source = RIPE NCC)

Rate limiting in action (IBGP vs EBGP)

AS 7018 IBGP (MRAI = 4 Seconds)

AS 7018 EBGP (MRAI = 30 Seconds)

10 minute bins
Why is Rate Limiting Needed?

Rate limiting dampens some of the oscillation inherent in a vectoring protocol.


Two Main Factors in Delayed Convergence

- BGP can explore many alternate paths before giving up or arriving at a new path
  - No global knowledge in vectoring protocols
- Rate limiting timer slows everything down

Current interval (30 seconds) was picked “out of the blue sky”