Advanced Systems Topics

Lent Term, 2010

Timothy G. Griffin
Computer Lab
Cambridge UK
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?
**LOCAL PREFERENCE**

Higher Local preference values are more preferred

Local preference used ONLY in iBGP

- AS 1
  - local pref = 100
- AS 2
  - local pref = 90
  - local pref = 80
- AS 3
- AS 4
- 13.13.0.0/16
How Can Routes be Classified?

BGP Communities!

A community value is 32 bits

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

Reserved communities
no_export = 0xFFFFFF01: don’t export out of AS
no_advertise 0xFFFFFF02: don’t pass to BGP neighbors

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

RFC 1997 (August 1996)
Tweak 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
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
High bandwidth Provider backbone

Low bandwidth customer backbone

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.

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 ...
... But Padding Does Not Always Work

AS 1
provider
192.0.2.0/24
ASPATH = 2

AS 2
primary
customer
192.0.2.0/24

AS 3
provider
192.0.2.0/24
ASPATH = 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 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.
COMMUNITY Attribute to the Rescue!

AS 1
provider

AS 3
provider

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

192.0.2.0/24
ASPATH = 2
COMMUNITY = 3:70

192.0.2.0/24
ASPATH = 2

primary
backup

AS 2

customer

AS 1
provider

192.0.2.0/24
ASPATH = 2

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 are 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) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null
IGP Tie Breaking Can Export Internal Instability to the Whole Wide World

192.44.78.0/24

AS 1

AS 2

AS 3

AS 4

192.44.78.0/24

ASPATH = 4 2 1

FLAP FLAP

192.44.78.0/24

ASPATH = 4 3 1

FLAP
MEDs Can Export Internal Instability

192.44.78.0/24
MED = 15

192.44.78.0/24
MED = 56 OR 10

Heavy Content Web Farm
Implementation Does Matter!

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

BGP Impact of SQL Worm, 1/25/2003, (plotted by Tim Griffin using Route-Views data)
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”
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 = 15 minutes) -- 6 flaps, one every 10 minutes

penalty
suppress limit
reuse limit

route dampened for nearly 1 hour

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

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

WITHDRAW
UPDATE
**Rate Limiting**

<table>
<thead>
<tr>
<th>MRAI = Minimum Router Advertisement Interval (in seconds)</th>
<th>Your Best Path</th>
<th>MRAI seconds</th>
<th>what you send to your neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>889 736 11</td>
<td>MRAI seconds</td>
<td>Announce 555 67 11</td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>458 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>889 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>555 67 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 555 67 11</td>
<td></td>
<td>Announce 555 67 11</td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>458 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>889 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>458 11</td>
<td></td>
<td>Withdraw</td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>889 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>555 67 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 555 67 11</td>
<td></td>
<td>Announce 7018 555 67 11</td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>458 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>889 736 11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NO ROUTE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7018 7018 555 67 11</td>
<td></td>
<td>Keep quite!</td>
</tr>
</tbody>
</table>
Updates often come in bursts, about every 30 seconds, June 25 2001 (data source = RIPE NCC)

announced + withdrawn prefixes, ISP 1
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.

SSFNet (www.ssfnet.org) simulations,
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”