L11: Inter-domain Routing with BGP
Lecture 14
Michaelmas, 2016
Timothy G. Griffin
Computer Lab
Cambridge UK
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

<table>
<thead>
<tr>
<th>Value</th>
<th>Code</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ORIGIN</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>2</td>
<td>AS_PATH</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>3</td>
<td>NEXT_HOP</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>4</td>
<td>MULTI_EXIT_DISC</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>5</td>
<td>LOCAL_PREF</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>6</td>
<td>ATOMIC_AGGREGATE</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>7</td>
<td>AGGREGATOR</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>8</td>
<td>COMMUNITY</td>
<td>[RFC1997]</td>
</tr>
<tr>
<td>9</td>
<td>ORIGINATOR_ID</td>
<td>[RFC2796]</td>
</tr>
<tr>
<td>10</td>
<td>CLUSTER_LIST</td>
<td>[RFC2796]</td>
</tr>
<tr>
<td>11</td>
<td>DPA</td>
<td>[Chen]</td>
</tr>
<tr>
<td>12</td>
<td>ADVERTISER</td>
<td>[RFC1863]</td>
</tr>
<tr>
<td>13</td>
<td>RCID_PATH / CLUSTER_ID</td>
<td>[RFC1863]</td>
</tr>
<tr>
<td>14</td>
<td>MP_REACH_NLRI</td>
<td>[RFC2858]</td>
</tr>
<tr>
<td>15</td>
<td>MP_UNREACH_NLRI</td>
<td>[RFC2858]</td>
</tr>
<tr>
<td>16</td>
<td>EXTENDED COMMUNITIES</td>
<td>[Rosen]</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>reserved for development</td>
<td></td>
</tr>
</tbody>
</table>

From IANA: [http://www.iana.org/assignments/bgp-parameters](http://www.iana.org/assignments/bgp-parameters)

Most important attributes

Not all attributes need to be present in every announcement

### How Can Routes be Classified?

**BGP Communities**

- A community value is 32 bits
- Used for signally within and between ASes
- Very powerful BECAUSE it has no (predefined) meaning

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

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

RFC 1997 (August 1996)

Reserved communities

- no_export = 0xFFFFF01: don’t export out of AS
- no_advertise 0xFFFFFFFF: don’t pass to BGP neighbors
BGP Route Processing

Open ended programming. Constrained only by vendor configuration language

Receive BGP Updates → Apply Import Policies → Best Route Selection → Best Route Table → Apply Export Policies → Transmit BGP Updates

Based on Attribute Values

Best Routes

Apply Policy = filter routes & tweak attributes

Install forwarding Entries for best Routes.

IP Forwarding Table

Route Selection Summary (A lexicographic product)

Highest Local Preference

Enforce relationships

Shortest AS_PATH

Lowest MED

I-BGP < e-BGP

Lowest IGP cost to BGP egress

Lowest router ID

traffic engineering

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)

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 an ugly hack ...

Padding will (usually) force inbound traffic from AS 1 to take primary link.
... But Padding Does Not Always Work

AS 1
provider

AS 3
provider

192.0.2.0/24
ASPATH = 2

customer
AS 2

primary
backup

192.0.2.0/24
ASPATH = 2

ASPATH = 2 2 2 2 2 2 2 2 2 2 2 2 2 2

customer

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

192.0.2.0/24
ASPATH = 2

customer
AS 2

primary
backup

192.0.2.0/24

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

AS 3: 3:70

COMMUNITY = 3:70
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

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 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 been 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.