BGP

Inter-domain routing with the Border Gateway Protocol

Iljitsch van Beijnum
Amsterdam, 13 & 16 March 2007
Routing Between ISPs

Internal routing protocols don't work here: too much information

So: external routing protocols: Exterior Gateway Protocol (EGP)

Another way of looking: not per router, but per network or organization, or rather "autonomous system"
Autonomous Systems
Autonomous Systems
Autonomous Systems
Autonomous Systems

AS2503

AS192

AS29077
The Trees and the Forrest

Tracing the route to www.isoc.nl (212.206.127.42)

1 fa3-0-4-asd8ro2.enertel.nl (195.7.144.85) [AS 12394] 4 msec
2 fa1-0-0-asd1ro6.enertel.nl (195.7.144.145) [AS 12394] 4 msec
3 po0-0-0-asd10ro1.enertel.nl (195.7.154.14) [AS 12394] 4 msec
4 adm-b2-pos2-1.telia.net (213.248.72.133) [AS 1299] 4 msec
5 pos3-2.BR1.AMS3.ALTER.NET (146.188.64.113) [AS 702] 4 msec
6 so-0-2-0.TR1.AMS2.ALTER.NET (146.188.3.213) [AS 702] 4 msec
7 so-5-0-0.XR1.AMS6.ALTER.NET (146.188.8.77) [AS 702] 4 msec
8 so-0-0-0.cr1.hag1.alter.net (212.136.176.110) [AS 702] 4 msec
9 so-4-0-0.cr2.hag1.alter.net (212.136.176.146) [AS 702] 8 msec
10 so-0-0-0.cr2.rtm1.alter.net (212.136.176.121) [AS 702] 8 msec
11 412.atm10-0-0.gw4.rtm1.alter.net (212.136.177.146) [AS 702] 16 msec
12 www.isoc.nl (212.206.127.42) [AS 702] 4 msec
The Trees and the Forrest

Tracing the route to www.isoc.nl (212.206.127.42)

1 fa3-0-4-asd8ro2.enertel.nl (195.7.144.85) [AS 12394] 4 msec
2 fa1-0-0-asd1ro6.enertel.nl (195.7.144.145) [AS 12394] 4 msec
3 po0-0-0-asd10ro1.enertel.nl (195.7.154.14) [AS 12394] 4 msec
4 adm-b2-pos2-1.telia.net (213.248.72.133) [AS 1299] 4 msec
5 pos3-2.BR1.AMS3.ALTER.NET (146.188.64.113) [AS 702] 4 msec
6 so-0-2-0.TR1.AMS2.ALTER.NET (146.188.3.213) [AS 702] 4 msec
7 so-5-0-0.XR1.AMS6.ALTER.NET (146.188.8.77) [AS 702] 4 msec
8 so-0-0-0.cr1.hag1.alter.net (212.136.176.110) [AS 702] 4 msec
9 so-4-0-0.cr2.hag1.alter.net (212.136.176.146) [AS 702] 8 msec
10 so-0-0-0.cr2.rtm1.alter.net (212.136.176.121) [AS 702] 8 msec
11 412.atm10-0-0.gw4.rtm1.alter.net (212.136.177.146) [AS 702] 16 msec
12 www.isoc.nl (212.206.127.42) [AS 702] 4 msec
IDR History

- Gateway-to-Gateway protocol (GGP)
- Exterior Gateway Protocol (EGP)
- BGP 1, 2 and 3
- IDRP and IDPR
- BGP4, RFC 1771, is used exclusively today
BGP's Functions

- Pass on which IP addresses are used where
- Enforce "routing policy"
- Avoid routing loops
- Avoid broken links
- Additionally, when possible: choose shortest path
Which Addresses Go Where

No geographic relevance to IP address distribution

Geographically close != network close

So: for each prefix we need to know where on the network it is used
Who Uses BGP?

Most ISPs, in order to tell other ISPs which address blocks they use

End users?

Most don't, ISP handles routing for their addresses

End users with more than one ISP (who are said to be “multihomed”) do, since they can't depend on a single ISP to handle this
How BGP Works

"Border routers" have connections with border routers from neighboring ASes
(And with all BGP routers within the local AS)
Communication over TCP port 179
Sessions are created manually
How BGP Works (2)

- As soon as the connection is established each router sends a (more or less) complete copy of the "global routing table" to the neighbor.

- Path through neighboring router better? Then use it yourself.

- When this is done, only updates when something changes.
BGP Header
OPEN Header

```
+-----------------+-----------------+-----------------+-----------------+
|          0      |          1      |          2      |          3      |
| 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
|     Version      |                  |                  |                  |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
| My Autonomous System |                  |                  |                  |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
|     Hold Time    |                  |                  |                  |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
| BGP Identifier  |                  |                  |                  |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
| Opt Parm Len    |                  |                  |                  |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
| Optional Parameters |                |                  |                  |
|+-----------------+-----------------+-----------------+-----------------+|
|                  |                  |                  |                  |
```
### UPDATE Header

<table>
<thead>
<tr>
<th>Unfeasible Routes Length (2 octets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Withdrawn Routes (variable)</td>
</tr>
<tr>
<td>Total Path Attribute Length (2 octets)</td>
</tr>
<tr>
<td>Path Attributes (variable)</td>
</tr>
<tr>
<td>Network Layer Reachability Information (variable)</td>
</tr>
</tbody>
</table>
### NOTIFICATION Hdr

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Error code</th>
<th>Error subcode</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Error code</th>
<th>Error subcode</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Error code</th>
<th>Error subcode</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KEEPALIVE Header

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                                                               |
|                           Marker                             |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|                                                               |
|          Length               |          --> 4 <--          |
|+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
BGP Path Attributes

Information routers exchange consists of a range of IP addresses and "path attributes"

Address information (prefix): Network Layer Reachability Information (NLRI)

Some common path attributes are:

- AS path, next hop, origin, Multi Exit Discriminator (or metric) and communities
AS Path

- Mandatory

- First of all: detect and avoid routing loops

- Also important to enforce transit/peering filters

- And for filters that make sure customers don't accidently act as transit ISPs

- (Shortest path is preferred when comparing different routes for the same destination)
Loops

AS1

192.0.2.0/24

AS2

AS4

AS3
Loops

AS1

$192.0.2.0/24$

AS2 BGP table:
$192.0.2.0 1$

AS2

AS4

AS3
Loops

AS1

AS2

AS3

AS4

192.0.2.0/24

AS2 BGP table:
192.0.2.0 1

AS3 BGP table:
192.0.2.0 2 1
Loops

AS1 192.0.2.0/24  AS2 BGP table:
   192.0.2.0  1

AS2  AS3  AS4

AS3 BGP table:
   192.0.2.0  2  1

AS4 BGP table:
   192.0.2.0  2  1
   3  2  1
Loops

AS1 ← 192.0.2.0/24

AS2 BGP table:
192.0.2.0 1

AS3 BGP table:
192.0.2.0 2 1

AS4 BGP table:
192.0.2.0 2 1
3 2 1
Loops

AS1 192.0.2.0/24

AS2 BGP table:
192.0.2.0 1

AS3 BGP table:
192.0.2.0 2 1

AS4 BGP table:
192.0.2.0 2 1

AS2

AS3

AS4
Local Preference

- Only exchanged within the local AS, but mandatory
- Route with the highest local preference is used
- Only when local pref is equal other attributes such as AS path are evaluated
Multi Exit Discriminator

- Optional

- Similar to "metric" in internal routing protocols

- Originally meant to differentiate between multiple similar routes over the same neighboring AS, but useful elsewhere too

- Isn't communicated to external ASes unless specifically set
Communities

- 32 bit value used to convey user-defined information
- Not in the original BGP specification! (So obviously optional)
- Some "well-known" communities
- Most take the shape of AS:nn (such as 701:120) where meaning depends on source AS
Multiprotocol BGP

Multiprotocol Extensions for BGP-4: RFC 2858

Extension to BGP4 that makes it possible to distribute routing information for additional "address families"

Announce capabilities to peer in open message

Put information for new protocol in two new path attributes
MP_REACH_NLRI

Address Family Identifier (2 octets)
Subsequent Address Family Identifier (1 octet)
Length of Next Hop Network Address (1 octet)
Network Address of Next Hop (variable)
Number of SNPAs (1 octet)
Length of first SNPA (1 octet)
First SNPA (variable)

..............
MP_REACH_NLRI (2)

- Length of second SNPA (1 octet)
- Second SNPA (variable)
- ...
- Length of Last SNPA (1 octet)
- Last SNPA (variable)
- Network Layer Reachability Information (variable)
MP_UNREACH_NLRI

+---------------------------------------------------------+
| Address Family Identifier (2 octets)                     |
+---------------------------------------------------------+
| Subsequent Address Family Identifier (1 octet)           |
+---------------------------------------------------------+
| Withdrawn Routes (variable)                             |
+---------------------------------------------------------+

NLRI encoding:

+---------------------------+
|   Length (1 octet)        |
+---------------------------+
|   Prefix (variable)       |
+---------------------------+
IPv6

- RFC 2545: Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing
- Almost completely identical to IPv4
- Ok, the addresses are longer...
- Next hop: global, + optional link local
- Operation of BGP with IPv6 still quite different: free transit, lots of tunnels
BGP Security

- BGP is sensitive to attacks at the TCP, IP and link (usually ethernet) layers.
- But in most cases not simple to exploit! MD5/password offers protection, if hard to manage.
- Future: Secure BGP (S-BGP) and/or secure origin BGP (soBGP).
BGP TCP MD5 Option

**RFC 2385:**

```
+---------+---------+-------------------+
| Kind=19 |Length=18|   MD5 digest...   |
+---------+---------+-------------------+
```

The MD5 digest is always 16 bytes in length, and the option would appear in every segment of a connection.
S-BGP

Secure BGP, created by BBN

Proof-of-concept implementation available

Map prefix to AS

Include destination AS in BGP update

Sign everything

Carry information in path attributes
soBGP

- Secure Origin BGP, by people from Cisco
- Authenticate relationship prefix and AS
- Extensible
- Authentication data in new BGP message
S-BGP vs soBGP

- S-BGP much heavier: 500k signature checks on startup
- S-BGP needs private keys in routers
- soBGP could offload crypto to server
Problems

- All that crypto!
- Memory
- Where is a list of which prefix goes with which AS?
- Can be, but isn't done today without S-/soBGP (with huge filters)
- Chicken/egg
Policy

- Only send routes when allowed: don't provide services unless paid for
- Additionally: configured preference
- And: check whether what the neighbor sends is correct
Balancing Traffic

- With more than one external connection BGP automatically balances the traffic.
- But not always in the most desirable way.
- Traffic engineering chapter from my book is online at the O'Reilly website, the URL is http://www.oreilly.com/catalog/bgp/
BGP Path Selection

1. Prefer the path with the largest LOCAL_PREF
2. Prefer the path with the shortest AS_PATH
3. Prefer the path with the lowest multi-exit discriminator (MED)
4. Prefer external (eBGP) over internal (iBGP) paths
5. Prefer the path with the lowest IGP metric to the BGP next hop
6. Prefer the route coming from the BGP router with the lowest router ID
7. Prefer the path coming from the lowest neighbor address
Too Effective?

- When two ISPs both peer with mostly the same networks, paths will tend to be the same length.

- In this case, even small changes radically change traffic patterns.
iBGP Scaling

- iBGP Full mesh requirement (ouch!)

- Fix this using:
  - Route reflectors: client talks to reflector, reflectors and non-clients in full mesh
  - Confederations: split in sub-ASes, full mesh inside, semi-eBGP to others
Full Mesh
Route Reflection
Confederations
Loopback

\[ \text{Rtr1} \quad a \quad \text{Rtr2} \]

\[ \text{b} \quad \text{Rtr3} \quad \text{y} \quad \text{Rtr4} \]

\[ \text{x} \]
Loopback
Loopback
Peering

- Private peering (direct link), or
- Exchange such as AMS-IX, LINX, DE-CIX
- Big layer 2 network, everyone connects a router
- Direct BGP sessions with others to exchange routing info
Policy for Transit

- Transit is provided to paying customers
- X provides transit to Y
- X routes packets from/to far away on behalf of Y
- So X announces to everyone that Y is reachable through them
- And X announces to Y that the whole world is reachable through them
Policy for Peering

- Exchange traffic without money changing hands (usually)
- X peers with Z
- X announces to Z that X’s customers are reachable through X, but NOT the rest of the world
- Z does the same, so all traffic has X or a customer of X as its source and Z or a customer of Z as its destination
Transit vs Peering

AS 200

AS 300

AS 400

AS 100

X
Transit vs Peering

- AS 200
- AS 300
- AS 400

AS 100

Y

X
Transit vs Peering

AS 200

AS 300

AS 400

AS 100

Y

X
Transit vs Peering

AS 200

AS 300

AS 400

AS 100

Y

X

Z
Transit vs Peering

AS 200

AS 300

AS 400

AS 100

Y

X

Z
Peering/Transit Example

**AS1: Large ISP**
- Default 4
- 164.0.0.0 2 4
- 165.0.0.0 2 5
- 166.0.0.0 3 6

**AS2: ISP 2**
- Default 1
- 164.0.0.0 4
- 164.0.0.0 1 4
- 165.0.0.0 5
- 166.0.0.0 3 6
- 166.0.0.0 1 3 6

**AS3: ISP 3**
- Default 1
- 164.0.0.0 4
- 164.0.0.0 1 4
- 165.0.0.0 2 5
- 165.0.0.0 1 2 5
- 166.0.0.0 6

**AS4: Multihomed**
- Default 2 1
- 164.0.0.0
- 165.0.0.0 2 5
- 165.0.0.0 1 2 5
- 166.0.0.0 2 3 6
- 166.0.0.0 1 3 6

**AS5: Single homed**
- Default 2 1
- 164.0.0.0 2 4
- 165.0.0.0
- 166.0.0.0 2 3 6

**AS6: Single homed**
- Default 3 1
- 164.0.0.0 3 2 4
- 165.0.0.0 3 2 5
- 166.0.0.0
Multilateral Peering

- Everyone peers with the route server
- Route server propagates all routes
- Next hop address isn't changed
- So traffic is exchanged directly
- See http://www.openpeering.nl/
More Interconnects

- Must decide where to jump from network A to network B
- Early exit / hot potato: get rid of traffic as soon as possible (happens automatically)
- Late exit / cold potato: keep it in your network as long as possible (rare)
Early Exit
Early Exit (2)
Routing Table Pathology

- 214000 routes in global routing table
- Often poor aggregation
- So: filter on prefix length:
  - < /24: forget it
  - RIR block: never any problems
  - In between: sometimes filtered
That's it!

- http://www.bgpexpert.com/
- http://www.oreilly.com/catalog/bgp/
- RFC 1771 (BGP version 4)
- RFC 4271 (revised BGP4 standard document)
- RFC 2385 (BGP TCP MD5 Option)