Border Gateway Protocol (BGP)

(Bruce Maggs and Nick Feamster)
BGP Primer

Autonomous System Number

AT&T
7018
12/8

Sprint
1239
144.223/16

CMU
9
128.2/16

bmm.pc.cs.cmu.edu
128.2.205.42

AS Path

Block of IP addresses
BGP Details

• AS that owns a prefix “originates” an advertisement with only its AS number on path

• AS advertises only its primary path to a prefix (the path it actually uses) to its neighbors

• Primary path for an IP address must be chosen from received advertisements with most specific (longest) prefix containing address, e.g., for 128.2.205.42, 128.2.205/24 is preferred over 128.2/16

• Advertisement contains entire AS path to prevent cycles

• Router withdraws the advertisement if the path is no longer available
Problems with BGP

• Not secure – susceptible to route “hijacking”

• Routing policy determined primarily by economics, not performance

• Slow to converge (and not guaranteed)

• During convergence, endpoints can be disconnected even when valid routes exist
What Causes Transient Disconnection?

All of Hari’s providers use him to get to MIT

BGP Rule:
An AS advertises only its current forwarding path

Nobody offers Hari an alternate path
What Causes Transient Disconnection?

Hari knows no path to MIT

Hari drops Peter’s and AT&T’s packets in addition to his own

LOSS!
What Causes Transient Disconnection?

Hari withdraws path

AT&T and Peter move to alternate paths
What Causes Transient Disconnection?

Hari withdraws path

AT&T and Peter move to alternate paths
AT&T announces the Sprint path to Hari
→ Traffic flows

Transient Packet Loss
Two Flavors of BGP

- **External BGP (eBGP):** exchanging routes between ASes
- **Internal BGP (iBGP):** disseminating routes to external destinations among the routers within an AS

**Question:** What’s the difference between IGP and iBGP?
Example BGP Routing Table

The full routing table

> show ip bgp

<table>
<thead>
<tr>
<th>Network</th>
<th>Next Hop</th>
<th>Metric</th>
<th>LocPrf</th>
<th>Weight</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>*&gt;i3.0.0.0</td>
<td>4.79.2.1</td>
<td>0</td>
<td>110</td>
<td>0</td>
<td>3356 701 703 80 i</td>
</tr>
<tr>
<td>*&gt;i4.0.0.0</td>
<td>4.79.2.1</td>
<td>0</td>
<td>110</td>
<td>0</td>
<td>3356 i</td>
</tr>
<tr>
<td>*&gt;i4.21.254.0/23</td>
<td>208.30.223.5</td>
<td>49</td>
<td>110</td>
<td>0</td>
<td>1239 1299 10355 10355 i</td>
</tr>
<tr>
<td>* i4.23.84.0/22</td>
<td>208.30.223.5</td>
<td>112</td>
<td>110</td>
<td>0</td>
<td>1239 6461 20171 i</td>
</tr>
</tbody>
</table>

Specific entry. Can do longest prefix lookup:

> show ip bgp 130.207.7.237
BGP routing table entry for 130.207.0.0/16
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  10578 11537 10490 2637
Prefix: 192.5.89.89 from 18.168.0.27 (66.250.252.45)
  Origin IGP, metric 0, localpref 150, valid, internal, best
  Community: 10578:700 11537:950
  Last update: Sat Jan 14 04:45:09 2006
Route Attributes and Route Selection

BGP routes have the following attributes, on which the route selection process is based:

- **Local preference**: numerical value assigned by routing policy. Higher values are more preferred.
- **AS path length**: number of AS-level hops in the path
- **Multiple exit discriminator (“MED”)**: allows one AS to specify that one exit point is more preferred than another. Lower values are more preferred.
- **Shortest IGP path cost to next hop**: implements “hot potato” routing
- **Router ID tiebreak**: arbitrary tiebreak, since only a single “best” route can be selected
Other BGP Attributes

- **Next-hop**: IP address to send packets en route to destination. *Question*: How to ensure that the next-hop IP address is reachable? Either import external address (e.g., 4.79.2.1) into internal routing tables, or use next-hop-self neighbor command to advertise (in iBGP) own address, as shown.

- **Community value**: Semantically meaningless. Used for passing around “signals” and labelling routes.
Local Preference

- Control over *outbound* traffic
- *Not* transitive across ASes
- Coarse hammer to implement route preference
- Useful for preferring routes from one AS over another (e.g., primary-backup semantics)
AS Path Length

- Among routes with highest local preference, select route with shortest AS path length
- Shortest AS path != shortest path, for any interpretation of “shortest path”
AS Path Length Hack: Prepending

- Attempt to control inbound traffic
- Make AS path length look artificially longer
- How well does this work in practice vs. e.g., hacks on longest-prefix match?
Multiple Exit Discriminator (MED)

- Mechanism for AS to control how traffic enters, given multiple possible entry points.

![Diagram of MED concept with cities and traffic flows]

- San Francisco (MED: 20)
- New York (MED: 10)
- Los Angeles
Hot-Potato Routing

- Prefer route with shorter IGP path cost to next-hop
- *Idea:* traffic leaves AS as quickly as possible

**Common practice:** Set IGP weights in accordance with propagation delay (e.g., miles, etc.)
Internet Business Model (Simplified)

- **Customer/Provider:** One AS pays another for reachability to some set of destinations

- **“Settlement-free” Peering:** Bartering. Two ASes exchange routes with one another.

Preferences implemented with local preference manipulation
Filtering and Rankings

**Filtering:** route advertisement

**Ranking:** route selection

### Table

<table>
<thead>
<tr>
<th>Type of neighboring AS</th>
<th>Ranking</th>
<th>Filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer</td>
<td>Most preferred</td>
<td>Advertise to all other ASes</td>
</tr>
<tr>
<td>Peer</td>
<td>Less preferred than routes through customer, more preferred than routes through provider</td>
<td>Advertise to customer ASes</td>
</tr>
<tr>
<td>Provider</td>
<td>Least preferred</td>
<td></td>
</tr>
</tbody>
</table>
Who owns a prefix?

- Organizations are granted prefixes of addresses, e.g., 128.2/16, by regional Internet registries ARIN, RIPE NCC, APNIC, AFRINIC, LACNIC

Source: http://www.apnic.net/about-APNIC/organization/history-of-apnic/history-of-the-regional-internet-registries

- Organizations also separately register AS numbers, but no linkage between AS numbers and prefixes.
Route Hijacking

• Any network can advertise that it knows a path to any prefix!
• No way to check if the path is legitimate.
• Highly specific advertisements (e.g., 128.2.205/24) will attract traffic.
• To mitigate risk, network operators manually create filters to limit what sorts of advertisements they will trust from their peers.
Why Hijack Routes?

- Steal some IP addresses temporarily, send SPAM until the addresses are blacklisted.
- Create a sinkhole to divert traffic away from a Web site, making it unavailable.
- Eavesdrop on traffic but ultimately pass it along.
The AS 7007 Incident

• On April 25, 1997, AS 7007 (MAI Network Services) leaked its entire routing table with all prefixes broken down (probably due to a bug) to /24 with original AS paths stripped off to AS 1790 Sprint.
• After MAI turned off their router, Sprint kept advertising the routes!
• See http://www.merit.edu/mail.archives/nanog/1997-04/msg00444.html
The Business Game and Depeering

• Cooperative competition (brinksmanship)
• Much more desirable to have your peer’s customers
  – Much nicer to get paid for transit
• Peering “tiffs” are relatively common

31 Jul 2005: Level 3 Notifies Cogent of intent to disconnect.
16 Aug 2005: Cogent begins massive sales effort and mentions a 15 Sept. expected depeering date.
31 Aug 2005: Level 3 Notifies Cogent again of intent to disconnect (according to Level 3)
5 Oct 2005 9:50 UTC: Level 3 disconnects Cogent. Mass hysteria ensues up to, and including policymakers in Washington, D.C.
7 Oct 2005: Level 3 reconnects Cogent

During the “outage”, Level 3 and Cogent’s singly homed customers could not reach each other. (~ 4% of the Internet’s prefixes were isolated from each other)
Depeering Continued

Resolution...

**Level 3 and Cogent Reach Agreement on Equitable Peering Terms**
Friday October 28, 7:00 am ET

BROOMFIELD, Colo. and WASHINGTON, Oct. 28 /PRNewswire-FirstCall/ -- Level 3 Communications (Nasdaq: LVLT - News) and Cogent Communications (Amex: COI - News) today announced that the companies have agreed on terms to continue to exchange Internet traffic under a modified version of their original peering agreement. The modified peering arrangement allows for the continued exchange of traffic between the two companies' networks, and includes commitments from each party with respect to the characteristics and volume of traffic to be exchanged. Under the terms of the agreement, the companies have agreed to the settlement-free exchange of traffic subject to specific payments if certain obligations are not met.

...but not before an attempt to steal customers!

As of 5:30 am EDT, October 5th, Level(3) terminated peering with Cogent without cause (as permitted under its peering agreement with Cogent) even though both Cogent and Level(3) remained in full compliance with the previously existing interconnection agreement. Cogent has left the peering circuits open in the hope that Level(3) will change its mind and allow traffic to be exchanged between our networks. **We are extending a special offering to single homed Level 3 customers.**

Cogent will offer any Level 3 customer, who is single homed to the Level 3 network on the date of this notice, one year of full Internet transit free of charge at the same bandwidth currently being supplied by Level 3. Cogent will provide this connectivity in over 1,000 locations throughout North America and Europe.
Policy Interactions

Varadhan, Govindan, & Estrin, “Persistent Route Oscillations in Interdomain Routing”, 1996
Customers, Providers, and Peers

GloboNet

RegioNet

LocoNet

CountryNet

Provider

Customer

MajorNet

MinorNet

Peer

Peer
Valley Free Paths

- GloboNet
- MajorNet
- RegioNet
- CountryNet
- MinorNet
- LocoNet

Zero or more customer-to-provider links
Zero or one peer-to-peer link
Zero or more provider-to-customer links
Theorem: If there are no customer-provider cycles, and every AS prefers routes learned from its customers, and all advertised routes are valley free, then BGP is guaranteed to converge.