Modern Routing Practices

CS249i: The Modern Internet
A Few Notes on BGP

BGP — routers share *most efficient* (shortest path) with their neighbors

- You don't see *all* routes. Rather routes to all routed prefixes.

Your Routing Information Base (RIB) is organized by routed prefix

- A bunch of details, only a few really matter to you:
  - (Routed) Prefix
  - AS Path
  - Next Hop

Traffic is sent to the route that matches the most specific prefix
rib_ipv4_unicast Entry (in JSON)

{
  "sub_type": "rib_ipv4_unicast",
  "sequence_number": 0,
  "prefix": "103.127.54.0/24",
  "entries": [
    {
      "peer_index": 1,
      "orginated_time": 1632281047,
      "path_identifier": 0,
      "path_attributes":
        {
          "type": 2,
          "as_paths": [
            {
              "segment_type": 2,
              "num": 8,
              "asns": [65400, 65105, 32, 46749, 46749, 6939, 137367, 17995]
            }
          ]
        }
    },
    {
      "type": 3,
      "nexthop": "171.67.69.32"
    },
    {
      "type": 1,
      "value": 0
    },
    {
      "type": 4,
      "metric": 0
    },
    {
      "type": 8,
      "communities": [32, 454801053]
    }
  ]
},
"route_family": 65537}
Traceroute + Router Interfaces

traceroute to google.com (142.250.72.174), 30 hops max

1  _gateway (171.67.69.32)  0.388 ms  0.369 ms  0.360 ms
2   *   *   *
3  10.214.4.249 (10.214.4.249)  1.043 ms
4  dc-sf-rtr-vl12.SUNet (171.66.0.207)  1.082 ms
5  dc-sfo-agg4--stanford-100g.cenic.net (137.164.23.178)  1.943 ms
6  dc-svl-agg8--sfo-agg4-100gbe.cenic.net (137.164.11.92)  2.532 ms
7  dc-svl-agg10--svl-agg8-300g.cenic.net (137.164.11.80)  1.860 ms
8  74.125.147.146 (74.125.147.146)  2.982 ms
9  108.170.242.254 (108.170.242.254)  3.95 ms
10 142.250.234.60 (142.250.234.60)  4.26 ms
11 142.250.211.208 (142.250.211.208)  10.564 ms
Looking Glass Servers

Useful to know BGP state at different routers — ISPs will often let you interrogate their public routing infrastructure — known as **Looking Glass** service.

```plaintext
core1.ash1.he.net> show ip bgp routes detail 8.8.8.8

<table>
<thead>
<tr>
<th>Status Codes</th>
<th>Status</th>
<th>Network</th>
<th>Next Hop</th>
<th>Learned</th>
<th>Metric</th>
<th>LocPrf</th>
<th>Weight</th>
<th>Path</th>
<th>Origin</th>
<th>RO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BME</td>
<td>8.8.8.0/24</td>
<td>206.53.170.23</td>
<td>206.53.170.1 (64216)</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>8.8.8.0/24</td>
<td>206.53.170.23</td>
<td>206.53.170.2 (64216)</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>8.8.8.0/24</td>
<td>206.126.236.21</td>
<td>206.126.236.21 (15169)</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.126.237.242</td>
<td>206.126.237.242 (15169)</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.32.118.39</td>
<td>216.218.252.171 (6939)</td>
<td>79</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.32.161.20</td>
<td>216.218.252.99 (6939)</td>
<td>84</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>198.32.132.41</td>
<td>216.218.252.150 (6939)</td>
<td>120</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>198.32.132.41</td>
<td>216.218.252.254 (6939)</td>
<td>120</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.53.203.14</td>
<td>216.218.252.147 (6939)</td>
<td>165</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>208.115.136.21</td>
<td>216.218.252.226 (6939)</td>
<td>165</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.41.110.73</td>
<td>216.218.252.168 (6939)</td>
<td>170</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>198.179.18.72</td>
<td>216.218.252.28 (6939)</td>
<td>199</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.108.255.141</td>
<td>216.218.252.185 (6939)</td>
<td>245</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>198.32.242.133</td>
<td>216.218.252.177 (6939)</td>
<td>270</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>8.8.8.0/24</td>
<td>206.53.174.7</td>
<td>216.218.252.167 (6939)</td>
<td>310</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
```
University of Oregon Route Views

University of Oregon collects router's RIBs from globally distributed set of IXPs and routers

Publishes these on a regular basis at http://archive.routeviews.org/

- Data Archives
  - MRT format RIBs and UPDATES (quagga bgpd, from route-views2.oregon-ix.net)
  - MRT format RIBs and UPDATES (quagga bgpd, from route-views3 as of Aug 13, 2013)
  - MRT format RIBs and UPDATES (quagga bgpd, from route-views4.routeviews.org)
  - v6 MRT format RIBs and UPDATES (quagga bgpd, from route-views6.oregon-ix.net)
  - MRT format RIBs and UPDATES from AMS-IX Collector (FRR bgpd, from route-views.amsix.routeviews.org)
  - MRT format RIBs and UPDATES from Chicago (FRR bgpd, from route-views.chicago.routeviews.org)
  - MRT format RIBs and UPDATES from NIC.cl Collector (FRR bgpd, from route-views.chile.routeviews.org)
  - MRT format RIBs and UPDATES from Equinix Ashburn (quagga bgpd, from route-views.eqix.routeviews.org)
  - MRT format RIBs and UPDATES from FL-IX (FRR bgpd, from route-views.flix.routeviews.org)
  - MRT format RIBs and UPDATES from GOREX (FRR bgpd, from route-views.gorex.routeviews.org)
  - MRT format RIBs and UPDATES from ISC (PAIX) (quagga bgpd, from route-views.isc.routeviews.org)
  - MRT format RIBs and UPDATES from KIXP (quagga bgpd, from route-views.kixp.routeviews.org)
  - MRT format RIBs and UPDATES from JINX (quagga bgpd, from route-views.jinx.routeviews.org)
  - MRT format RIBs and UPDATES from LINX (quagga bgpd, from route-views.linx.routeviews.org)
  - MRT format RIBs and UPDATES from NAPAfrica (FRR bgpd, from route-views.napafrica.routeviews.org)
  - MRT format RIBs and UPDATES from NWAX (quagga bgpd, from route-views.nwax.routeviews.org)
CAIDA ASRank — Inferring AS Relationships

CAIDA collects all routes from RouteViews. Attempt to infer relationships.


---

**ASRank** is CAIDA’s ranking of *Autonomous Systems (AS)* (which approximately map to Internet Service Providers) and organizations (Orgs) (which are a collection of one or more ASes). This ranking is derived from topological data collected by CAIDA’s *Archipelago Measurement Infrastructure* and *Border Gateway Protocol (BGP)* routing data collected by the *Route Views Project* and *RIPE NCC*.

ASes and Orgs are ranked by their customer cone size, which is the number of their direct and indirect customers. Note: We do not have data to rank ASes (ISPs) by traffic, revenue, users, or any other non-topological metric.

[https://asrank.caida.org/](https://asrank.caida.org/)
BGP Security
BGP Hijacking

BGP has no built in security! Any AS can advertise any prefix. Others will choose the shortest path — regardless of whether it's the correct path.
Real World Cases

In April 2018, a Russian provider announced IP prefixes that contained Route53 Amazon DNS servers.

They hijacked Amazon DNS queries so that DNS queries for myetherwallet.com went to attacker-controlled servers, which returned the wrong IP address, and directed HTTP requests to an imposter website.

The hackers were thus able to steal approximately $152,000 in cryptocurrency.

Would HTTPS have helped in this situation?
ISP-Provided Protections

For an end customer, an ISP should only accept that end customer’s IP address block. Any other prefix advertised from that customer should be dropped.

Easy for customers, but difficult for understanding what to filter from other ISPs
RPKI
Resource Public Key Infrastructure (RPKI)

PKI that communicates who owns IP prefixes and the AS number that can originate — in an object known as a Route Origin Authorization (ROA).

RPKI uses X.509 certificates with extensions for IPs and ASNs (RFC 3779)

Each RIR (Internet Registry) posts their public keys — act as the trust anchors
RPKI Deployment History

% Unique Prefix-Origin pairs

- Green: Valid
- Yellow: Not-Found
- Red: Invalid

Timeline from 2014 to 2021
MPLS
MPLS — Multiprotocol Label Switching

Routing technique where path through network is determined at ingress.

A short (Layer 2.5) label is tacked onto the front of the packet. Routers use tag to very quickly forward to the next router. Egress strips label.

Effectively L2 Routing. Avoids expensive L3 IP longest prefix match at each hop.

Complex to configure — but can have significant impact on router performance

Tier 1s often use MPLS on their backbone
Remote peering model

Your Peering VLAN from 100M to 2G or your dedicated nx10GE port

Your Network → Nearest Reseller PoP → Reseller Network → Cross-connect → Data center → Reseller 10G port(s) on France-IX peering platform → France-IX Community

- Single physical port delivered at your router (with potentially several VLANs)
- Nearest Reseller PoP
- Reseller Network
- Data center
- Cross-connect
- France-IX Convergence hub
- France-IX Community
Remote Peering

How does it work?

Remote Peering Provider is already installed at the IXPs.

Waves provisioned, instant turn up.

Neutral RPP no business clash

Peering Focus Speeds IXP deploy Little paperwork One Contract

Remote Peering No Router CapEx No Colocation Fees No Deployment/Install Fees Paperwork Reduction for IXP Near instant turn up
BGP Communities

"BGP Communities" — BGP attribute that is parsed and passed to BGP peers

- Effectively tags that are attached to routes

- Communities are transitive! Passed along multiple routers.

Communities allows an AS to tell its neighbors additional information about the routes it's advertising

Both standardized and non-standard communities exist
No Advertise
No Export
Other Standardized Communities

**NO_EXPORT_SUBCONFED:** Do not advertise outside of your BGP confederation

**NOPEER:** Other routers don't have to propagate the prefix

**BLACKHOLE:** Drop all traffic for this prefix (used to protect against DDoS)
Some NTT Communities

Customers wanting to alter their route announcements to selected peers

NTT BGP customers may choose to prepend to selected peers with the following communities, where nnn is the peer’s ASN:

<table>
<thead>
<tr>
<th>Community</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>65400:nnn</td>
<td>do not advertise to peer nnn in North America</td>
</tr>
<tr>
<td>65401:nnn</td>
<td>prepends o/b to peer nnn 1x in North America</td>
</tr>
<tr>
<td>65402:nnn</td>
<td>prepends o/b to peer nnn 2x in North America</td>
</tr>
<tr>
<td>65403:nnn</td>
<td>prepends o/b to peer nnn 3x in North America</td>
</tr>
<tr>
<td>65410:nnn</td>
<td>announce to peer nnn in North America, disregards 2914:429 and 65500:nnn</td>
</tr>
<tr>
<td>65420:nnn</td>
<td>do not advertise to peer nnn in Europe</td>
</tr>
<tr>
<td>65421:nnn</td>
<td>prepends o/b to peer nnn 1x in Europe</td>
</tr>
</tbody>
</table>