IXPs
Internet Exchange Points (IXPs) + Public Peering

Carrier Hotels provide shared interconnect (switching fabric) between ISPs

Allow ISPs to BGP peer with a large number of organizations through a single link

Peerings that use shared fabric ("public interconnect") known as a "Public Peering"

You still have to negotiate the BGP peering with others on the exchange

In the U.S., private peering more common. In Europe, public peering more common.
Many ISPs Will be Your Friend at an IXP

Most content providers will peer with you over public exchanges

So will cloud providers

Little downside for them not to if they're already on the exchange

Most ISPs (even Tier 1s) will sell you transit at an IXP (via private peering)
Multilateral Peering Exchanges

Typically, you'd establish peering relationships with others at an exchange.

This can become cumbersome if there are a large number of players (could be 100s or 1000s of organizations at a single IXP!)

Route server allows those with open policies to all BGP peer with only a single entity to both advertise its routes and collect routes from others on the exchange.

Known as **Multilateral Peering**
Packet Clearing House — Where are U.S. IXPs?
Packet Clearing House — Europe
Packet Clearing House — Asia

1111 IXPs shown - Number of IXPs by Country

Tokyo, Japan
- JPNAP Tokyo
- Japan Internet Exchange
- Network Service Provider IXP (Distributed IX in Edo)
- BBIX Tokyo
- Network Service Provider IXP-6
- Equinix IBX Tokyo
- JPIX Nihonbashii
- Media Exchange Co.
- Network Service Provider IXP
- Pihana Tokyo
- JPNAP Tokyo II - Ikebukuro
- Asia Smart IX [BBIX Asia]
- Tokyo Lambda Exchange

Source: pch.net
Equinix — Largest Commercial IXP Provider
Equinix Internet Exchange San Jose (Bay Area)

Facilities:
- Equinix SV1/SV5/SV10 - San Jose
- Equinix SV2 - Santa Clara
- Equinix SV3 - San Jose
- Equinix SV4 - Sunnyvale
- Equinix SV8 - Palo Alto

~210 (publicly listed) participants on San Jose Exchange. ~101 on Palo Alto Exchange.

~500 organizations listed between those 5 hotels (many more than on public exchange)
## European Example: AMS-IX

### AMS-IX Points of Presence

<table>
<thead>
<tr>
<th>Data Centre</th>
<th>Phone</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 NorthC Amsterdam (AMS01)</td>
<td>+31 (0)20 486 9773</td>
<td>Kabelweg 48a, Amsterdam</td>
</tr>
<tr>
<td>2 Digital Realty AMS17</td>
<td>-</td>
<td>Science Park 120, Amsterdam</td>
</tr>
<tr>
<td>3 Digital Realty AMS04</td>
<td>+31 (0)20 480 4415</td>
<td>H.J.E. Wenckebachweg 127, Amsterdam</td>
</tr>
<tr>
<td>4 Equinix AMI/2</td>
<td>+31 (0)53 434 0570</td>
<td>Luttenbergweg 4, Amsterdam</td>
</tr>
<tr>
<td>5 Equinix AM3</td>
<td>+31 (0)20 808 0015</td>
<td>Science Park 610, Amsterdam</td>
</tr>
<tr>
<td>6 Equinix AM5</td>
<td>+31 (0)20 592 8263</td>
<td>Schepenbergenweg 42, Amsterdam</td>
</tr>
<tr>
<td>7 Equinix AM6</td>
<td>+31 (0)53 436 2666</td>
<td>Duivendrechtse kade 80A, Amsterdam</td>
</tr>
<tr>
<td>8 Equinix AM7</td>
<td>+31 (0)53 434 0570</td>
<td>Kuiperbergweg 13, Amsterdam</td>
</tr>
<tr>
<td>9 EuNetworks</td>
<td>+31 (0)20 354 8098</td>
<td>Paul van Vlissingenstraat 16, Amsterdam</td>
</tr>
<tr>
<td>10 Iron Mountain</td>
<td>+31 (0)20 316 5170</td>
<td>J.W. Lucasweg 35, Haarlem</td>
</tr>
<tr>
<td>11 Global Switch</td>
<td>+31 (0)20 666 6300</td>
<td>Henk Sneevlietweg 2-6, Amsterdam</td>
</tr>
<tr>
<td>12 Interxion</td>
<td>+31 (0)20 880 7700</td>
<td>Tupolevlaan 101, Schiphol-Rijk</td>
</tr>
<tr>
<td>13 Interxion</td>
<td>+31 (0)20 560 6600</td>
<td>Science Park 121, Amsterdam</td>
</tr>
<tr>
<td>14 Nikhef</td>
<td>+31 (0)20 592 2037</td>
<td>Science Park 105, Amsterdam</td>
</tr>
</tbody>
</table>

### Key Statistics

<table>
<thead>
<tr>
<th>Members</th>
<th>882[1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ports</td>
<td>1,438  [1]</td>
</tr>
<tr>
<td>Peers</td>
<td>1,316  [1]</td>
</tr>
<tr>
<td>Peak in</td>
<td>9.022 Tbps[2]</td>
</tr>
<tr>
<td>Peak out</td>
<td>10.287 Tbps[2]</td>
</tr>
</tbody>
</table>
European Example: AMS-IX

Most details about AMS-IX are public at https://www.ams-ix.net/
PeeringDB is a free database of networks, IXPs, and facilities.

https://peeringdb.com/
Tier 1 Network
Backbone
Lumen
Hurricane Electric
The 8 U.S. Interconnection Regions

- Seattle
- San Francisco (Bay Area)
- Los Angeles
- New York Area
- Washington, D.C.
- Atlanta
- Chicago
- Dallas
What happens if Tier-1s De-Peer?

"Cogent has decided not to exchange traffic directly with TeliaSonera's AS 1299 or indirectly with AS 1299 through a third-party provider," Telia told its customers. "As a result, Cogent has partitioned the Internet and disrupted the flow of traffic between Cogent and TeliaSonera customers."

“Cogent has been controversial in the ISP market for low bandwidth pricing and its public disputes over peering with AOL (2003), Level 3 Communications (2005), France Telecom (2006), Limelight Networks (2007), Telia Carrier (March 2008), and Sprint Nextel (October 2008).”

![Classification of Disputes Collected](source: Anatomy of the Internet Peering Disputes)

*a Cute Epic Axis* · 3yr ago
Packet Whisperer
Whenever there is a peering dispute between Cogent and anyone else, assume first that Cogent is in the wrong and willing to use underhanded tactics and cronyism to resolve it. Only deviate from this when comprehensive proof shows otherwise. They no longer get the benefit of the doubt due to their past bullshit.
Flattening
Three Changes in the World:

• An increasing fraction of Internet traffic originates from a few CPs or CDNs (e.g., Google, YouTube, Akamai, Cloudflare). This shift is due to the large penetration of video streaming.

• The major CDNs and CPs have expanded to almost every region of the developed world, so that they can be co-located with many ASes at Internet Exchange Points (IXPs).

• IXPs have increased rapidly in number, making it easy and cheap for an AS to establish peering links with other ASes co-located at the same IXP.
Tier 1 Peerings Mattered Less
How Comcast became a toll-collecting, nuke-wielding hydra

Comcast wants cash to deliver cached Netflix traffic to its subscribers. Has ...

Down : Up > 5:1

Settlement Free Peering

Commercial Agreement
(L3 is Both ISP and CDN)
A Small Transit Provider Case Study

AS19653 – Small Transit Provider in Climax, Michigan
Founded in 1911 as an Independent Telephone Company.
Started as a CLEC in 1996.
Independent ILEC-CLEC-ISP.  CLLI = CLMXMIXI

2011 – Joined NANOG
Telephone Company (ILEC-CLEC)
Tier 3 ISP
100% transit (two OC-12s)

2017 – (after 18 NANOGs)
Packet Optical Service Provider
Tier 2 ISP
88% Peering
12% transit
More than 100G in upstream ports

Source: The Internet is Flat: Revisited
AS19653 Evolution of Transit to Peering

<table>
<thead>
<tr>
<th></th>
<th>Peering</th>
<th>Transit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2011</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>2012</td>
<td>36%</td>
<td>64%</td>
</tr>
<tr>
<td>2013</td>
<td>68%</td>
<td>32%</td>
</tr>
<tr>
<td>2014</td>
<td>74%</td>
<td>26%</td>
</tr>
<tr>
<td>2015</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>2016</td>
<td>86%</td>
<td>14%</td>
</tr>
<tr>
<td>2017</td>
<td>88%</td>
<td>12%</td>
</tr>
</tbody>
</table>

A Small Transit Provider Case Study

SOURCE: AS19653
AS19653 From Transit to IXP Peering to CDN

A Small Transit Provider Case Study

SOURCE: AS19653
Percentage of Total Traffic AS19653 Transit/Peering/CDN

A Small Transit Provider Case Study

SOURCE: AS19653
Modern Routing Practices
CS249i: The Modern Internet
Project 1 Notes
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)

```json
{
    "sub_type": "rib_ipv4_unicast",
    "sequence_number": 0,
    "prefix": "103.127.54.0/24",
    "entries": [
        {
            "peer_index": 1,
            "originated_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 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.

```
| Matching Routes | Status Codes | Network     | Next Hop     | Learned      | Metric | LocPrf | Weight | Path | Origin | RO#
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMEx</td>
<td>A - Aggregate</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>
</tr>
<tr>
<td>ME</td>
<td>B - Best</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>
</tr>
<tr>
<td>ME</td>
<td>C - Not Install Best</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>
</tr>
<tr>
<td>ME</td>
<td>D - Confederation</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>
</tr>
<tr>
<td>I</td>
<td>E - Damped</td>
<td>8.8.8.0/24</td>
<td>206.83.10.13</td>
<td>216.218.252.230 (6939)</td>
<td>15</td>
<td>100</td>
<td>0</td>
<td>15169</td>
<td>IGP</td>
<td>✓</td>
</tr>
<tr>
<td>I</td>
<td>F - History</td>
<td>8.8.8.0/24</td>
<td>198.32.118.39</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>
</tr>
<tr>
<td>I</td>
<td>G - Local</td>
<td>8.8.8.0/24</td>
<td>198.32.161.20</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>
</tr>
<tr>
<td>I</td>
<td>H - IBGP</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>
</tr>
<tr>
<td>I</td>
<td>I - Multipath</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>
</tr>
<tr>
<td>I</td>
<td>J - Not Installed Multipath</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>
</tr>
<tr>
<td>I</td>
<td>K - State</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>
</tr>
<tr>
<td>I</td>
<td>L - Best-External</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>
</tr>
<tr>
<td>I</td>
<td>M - Filtered</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>
</tr>
<tr>
<td>I</td>
<td>N - Suppressed</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>
</tr>
<tr>
<td>I</td>
<td>O - Best-Internal</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>
</tr>
<tr>
<td>I</td>
<td>P - S - State</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>
</tr>
</tbody>
</table>
```
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](http://archive.routeviews.org/europe) (quagga bgpd, from route-views2.oregon-ix.net)
  - [MRT format RIBs and UPDATES](http://archive.routeviews.org/europe) (quagga bgpd, from route-views3 as of Aug 13, 2013)
  - [v6 MRT format RIBs and UPDATES](http://archive.routeviews.org/europe) (quagga bgpd, from route-views6.oregon-ix.net)
  - [MRT format RIBs and UPDATES from AMS-IX Collector](http://archive.routeviews.org/europe) (FRR bgpd, from route-views.amsix.routeviews.org)
  - [MRT format RIBs and UPDATES from Chicago](http://archive.routeviews.org/europe) (FRR bgpd, from route-views.chicago.routeviews.org)
  - [MRT format RIBs and UPDATES from NICe CL Collector](http://archive.routeviews.org/europe) (FRR bgpd, from route-views.chile.routeviews.org)
  - [MRT format RIBs and UPDATES from Equinix Ashburn](http://archive.routeviews.org/europe) (quagga bgpd, from route-views.eqix.routeviews.org)
  - [MRT format RIBs and UPDATES from FL-IX](http://archive.routeviews.org/europe) (FRR bgpd, from route-views.flix.routeviews.org)
  - [MRT format RIBs and UPDATES from GOREX](http://archive.routeviews.org/europe) (FRR bgpd, from route-views.gorex.routeviews.org)
  - [MRT format RIBs and UPDATES from ISC (PAIX)](http://archive.routeviews.org/europe) (quagga bgpd, from route-views.isc.routeviews.org)
  - [MRT format RIBs and UPDATES from KIXP](http://archive.routeviews.org/europe) (quagga bgpd, from route-views.kixp.routeviews.org)
  - [MRT format RIBs and UPDATES from JINX](http://archive.routeviews.org/europe) (quagga bgpd, from route-views.jinx.routeviews.org)
  - [MRT format RIBs and UPDATES from LINX](http://archive.routeviews.org/europe) (quagga bgpd, from route-views.linx.routeviews.org)
  - [MRT format RIBs and UPDATES from NAPAfrica](http://archive.routeviews.org/europe) (FRR bgpd, from route-views.napafrica.routeviews.org)
  - [MRT format RIBs and UPDATES from NWAX](http://archive.routeviews.org/europe) (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/
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.
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

Last Year

- Valid: 31.69%
- Invalid: 0.61%
- Not-Found: 67.70%

Total: 945,584

Today

- Valid: 40.95%
- Invalid: 1.13%
- Not-Found: 57.92%

Total: 1,021,163

RPKI-ROV Analysis of Unique Prefix-Origin Pairs (IPv4)

Valid: 418,151
Not-Found: 591,466
Invalid: 11,546
RPKI Deployment History

RPKI-ROV History of Unique Prefix-Origin Pairs (IPv4)

NIST RPKI Monitor: RPKI-ROV Analysis  Protocol: IPv4  RIR: All
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.
Remote Peering
Remote peering model

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

Your Network

Nearest Reseller PoP

Reseller Network

Data center

Cross-connect

France-IX Convergence hub

Reseller 10G port(s) on France-IX peering platform

France-IX Community

Single physical port delivered at your router (with potentially several VLANs)
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>