INNOG Routing 101 - Oct 2020

Anurag Bhatia, Hurricane Electric

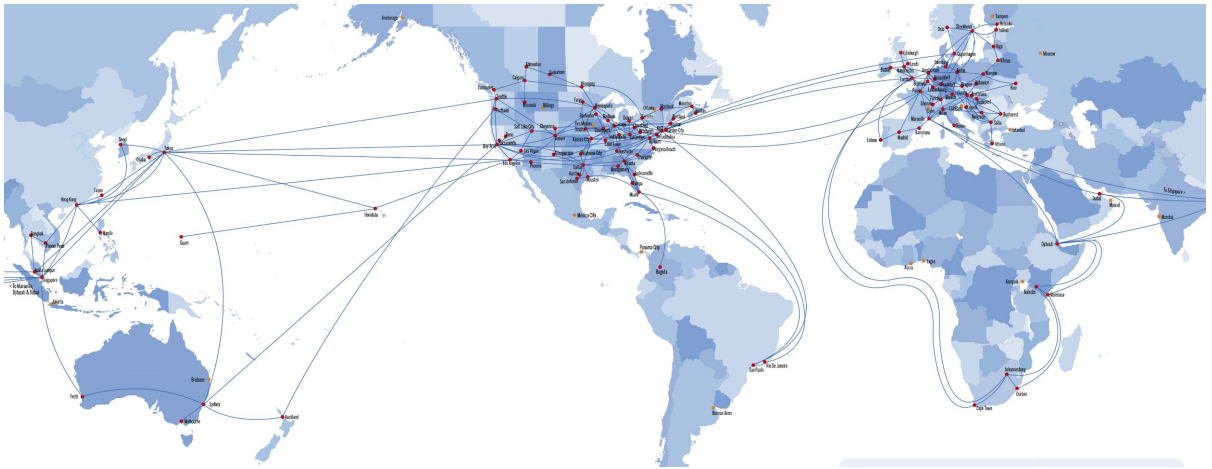
About me...

Working at Global backbone operator - Hurricane Electric and spend lot of time in looking at traceroutes, global routing, interesting patterns, issues etc

Besides routing have lot of interest in DNS, root DNS, IXPs, Network automation, tooling and virtualisation.

About Hurricane Electric

- Operating a Global IP backbone spanning across 45 countries, 240+ cities and multiple 100G rings across Atlantic & Pacifc.
- Connected with 8492 networks in 29000+ BGP sessions across 242 cities.
- Operating at multi-terabit scale with 100s of Terabit of edge capacity.
- Doing filtering based on IRR and RPKI :-)



Overview

1. Fundamentals of interconnection
2. Transit free networks
3. Hot potato Vs cold potato routing
4. BGP path selection process
5. Inbound and outbound traffic optimisation
6. Route hijack Vs route leak
7. Tools for analysis and troubleshooting

Starting with fundamentals of interconnection

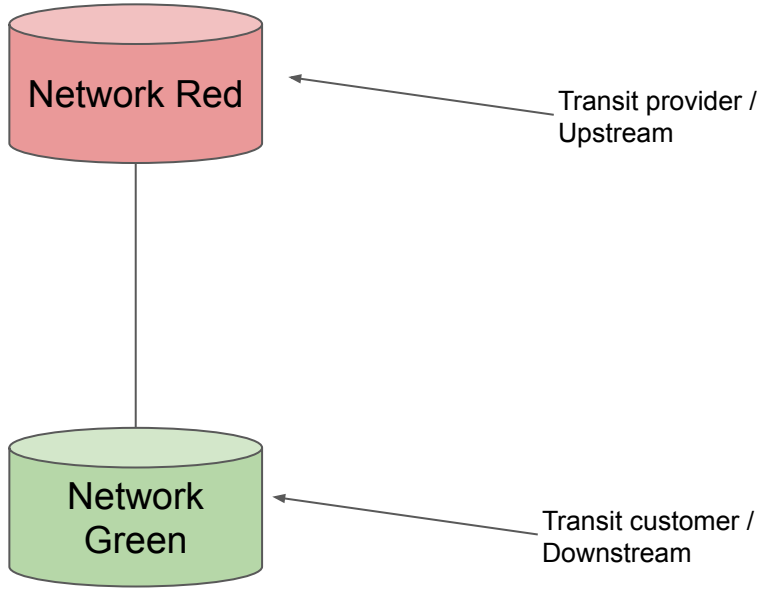
Peering Vs Transit

Peering

1. Peering partners exchange routes with each other.
2. Not to be confused with “peer” in BGP. A BGP peer can be a “peer” or “transit”.
3. Peering can be pair or free but it’s largely settlement free.
4. None of them act as transit for each other
5. Can be done over dedicated private links - Private Network Interconnect (PNI) or IXPs

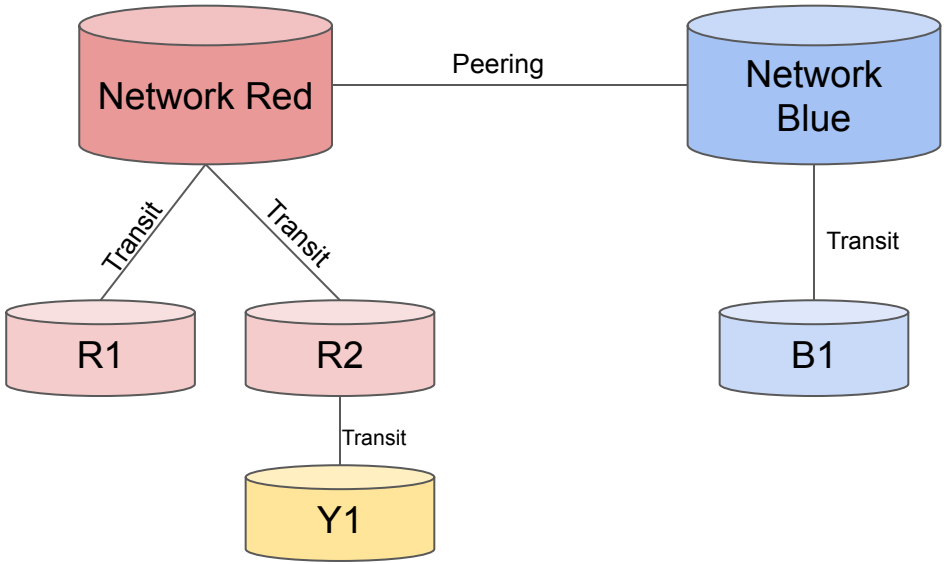
Transit

1. Transit provider gives a “guarantee” of global reachability
2. Transit is usually paid relation where transit customer pays to transit provider
3. A network can maintain a mix of transits and peering



Peering & Transit topology

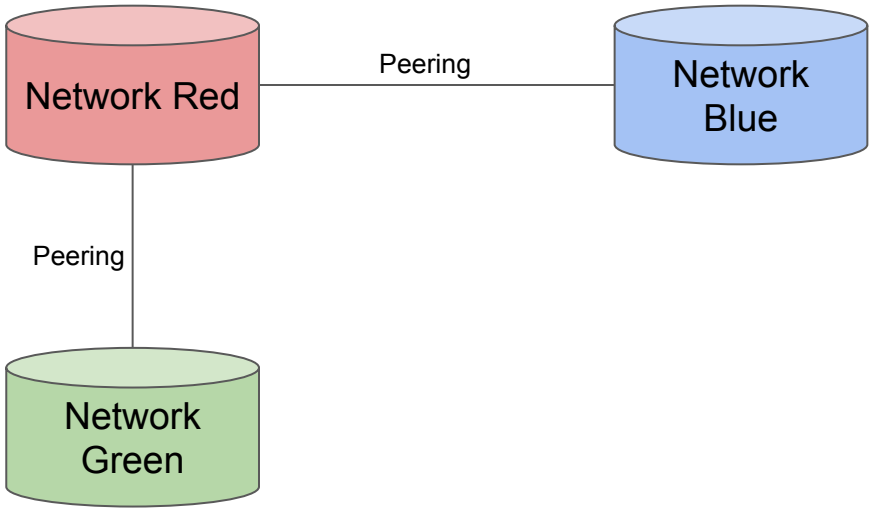
1. Network Red has “peering” relationship with Network Blue.
2. Network red would announce all it’s own + downstream routes including R1, R2 & Y1 to network blue.
3. Network blue would announce its own and downstream B1 routes to network red.



Peering ≠ Transit

Peering & Transit topology

1. Network green's routes will reach network red.
2. Network blue's routes will reach network red as well.
3. Network red's routes will reach both green and blue.
4. Network green's routes will **NOT** reach network blue via network red.



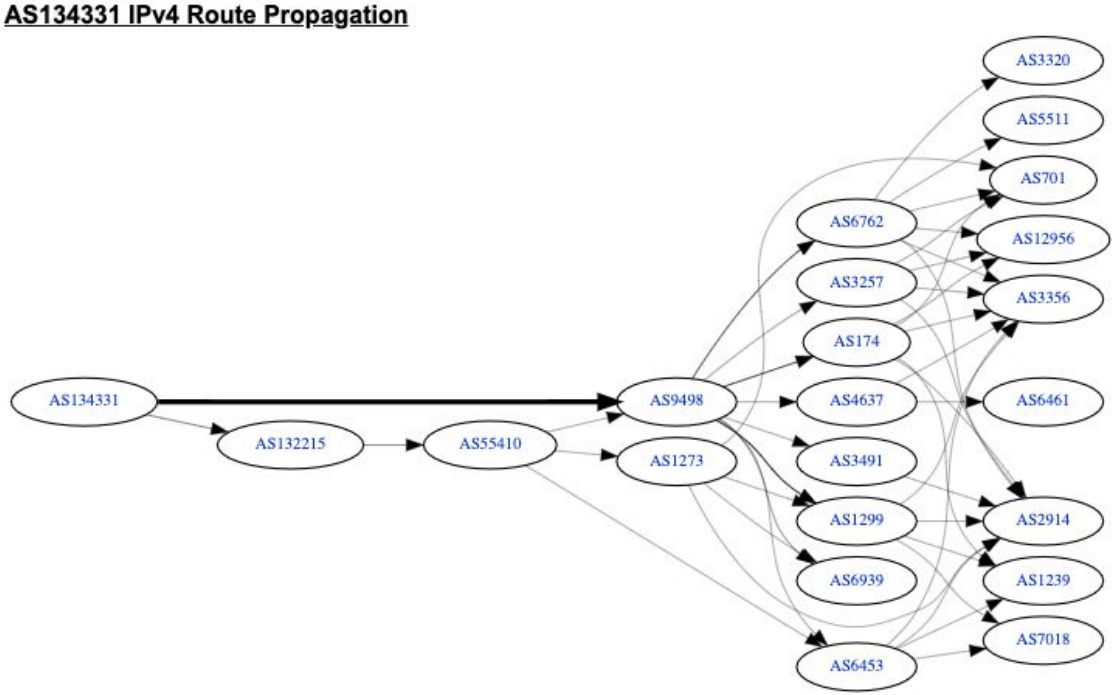
How a transit provider can guarantee reachability to any network in the world?

Transit free / Tier 1 network

Transit free / Tier 1 network

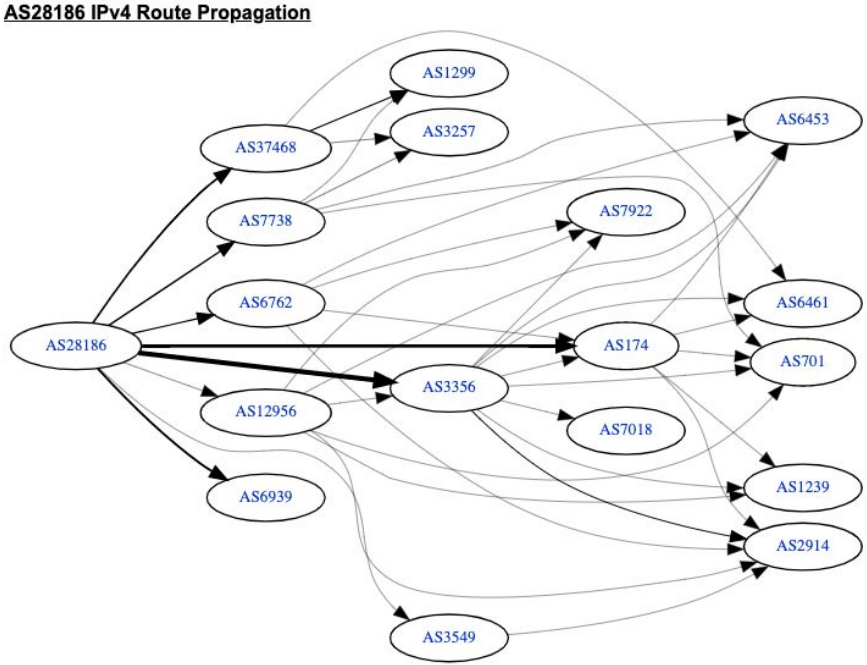
1. Advanced warning: Quite abused/misused term in marketing
2. (Strictly technically speaking) A transit free network is “transit free” and does not takes transit from anyone.
3. Transit free network can reach all networks in the world either through it downstream customers or it’s peers.
4. Approximately 15 networks globally are considered to tier 1. Wikipedia has the list here. Some of examples: AT&T AS7018, Verizon AS701, GTT AS3257, Centurylink / Level 3 AS3356, Tata Communications AS6453 etc.
5. All transit free networks peer with all other transit free networks. Remember: No transit by definition!

How Smartlink AS134331 connects to the world?



Source: https://bgp.he.net/AS134331#_graph4

What about a network from Brazil?



https://bgp.he.net/AS28186#_graph4

Typical global connectivity design

- Peer with networks with whom you have high traffic via PNIs and at IXPs.
- Buy IP transit locally in the country or nearby at large hubs like Singapore, Marseille (France), London (UK), Frankfurt (Germany), New York (US) etc.
- Quite similar to airplane traffic routes: India -> Europe & India -> Singapore quite direct. India -> Myanmar is often indirect.

Some misc points about Transit free networks

1. Important part as they stitch all networks in the world together and gives us a guaranteed reachability. Everyone cannot connect to everyone!

2. Not that important based on traffic as very large part of traffic is from content networks (like Google, Facebook, Akamai, Cloudflare, Netflix etc) towards eyeball networks (where end user - eyeballs sit) like Airtel, Jio, ACT, Ishaan, Blazenet, Spectra etc. Large part of traffic never hits transit free networks.

3. Somewhere around top 30 ASNs in the world result in as much as 90% traffic. That's 30 ASNs out of 65000+ ASNs visible in the world. <- Level of concentration of content.

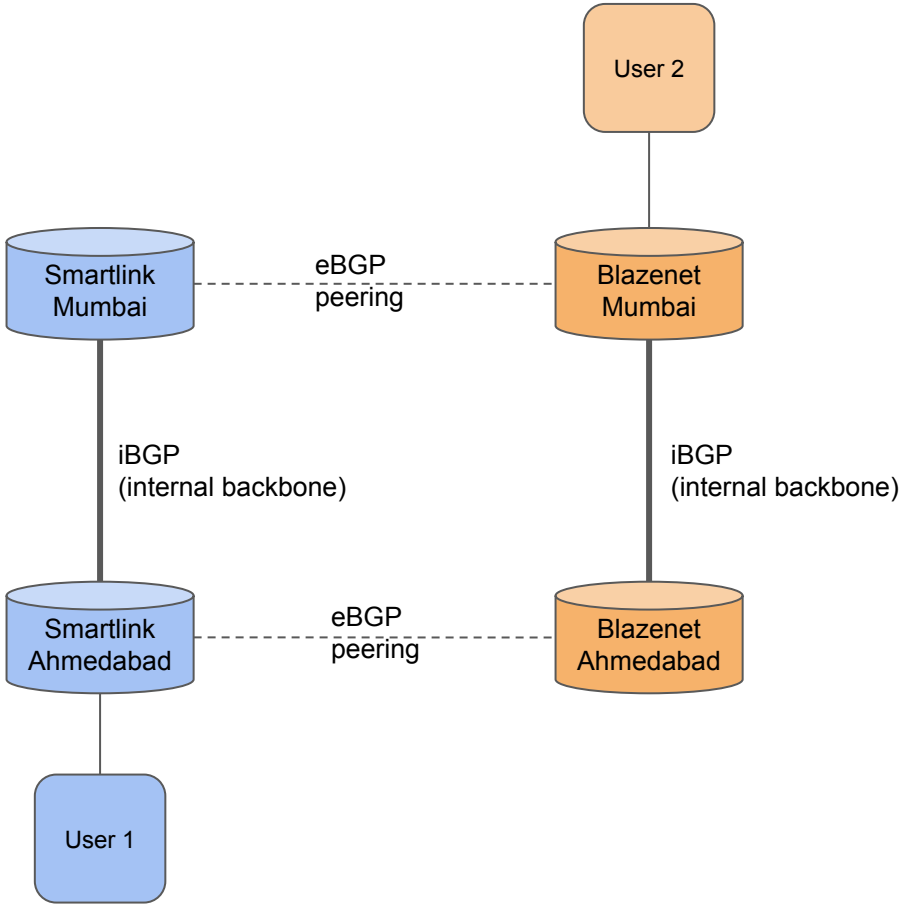
4. You will often find better connectivity with networks who follow open peering policy in comparison to transit free networks.

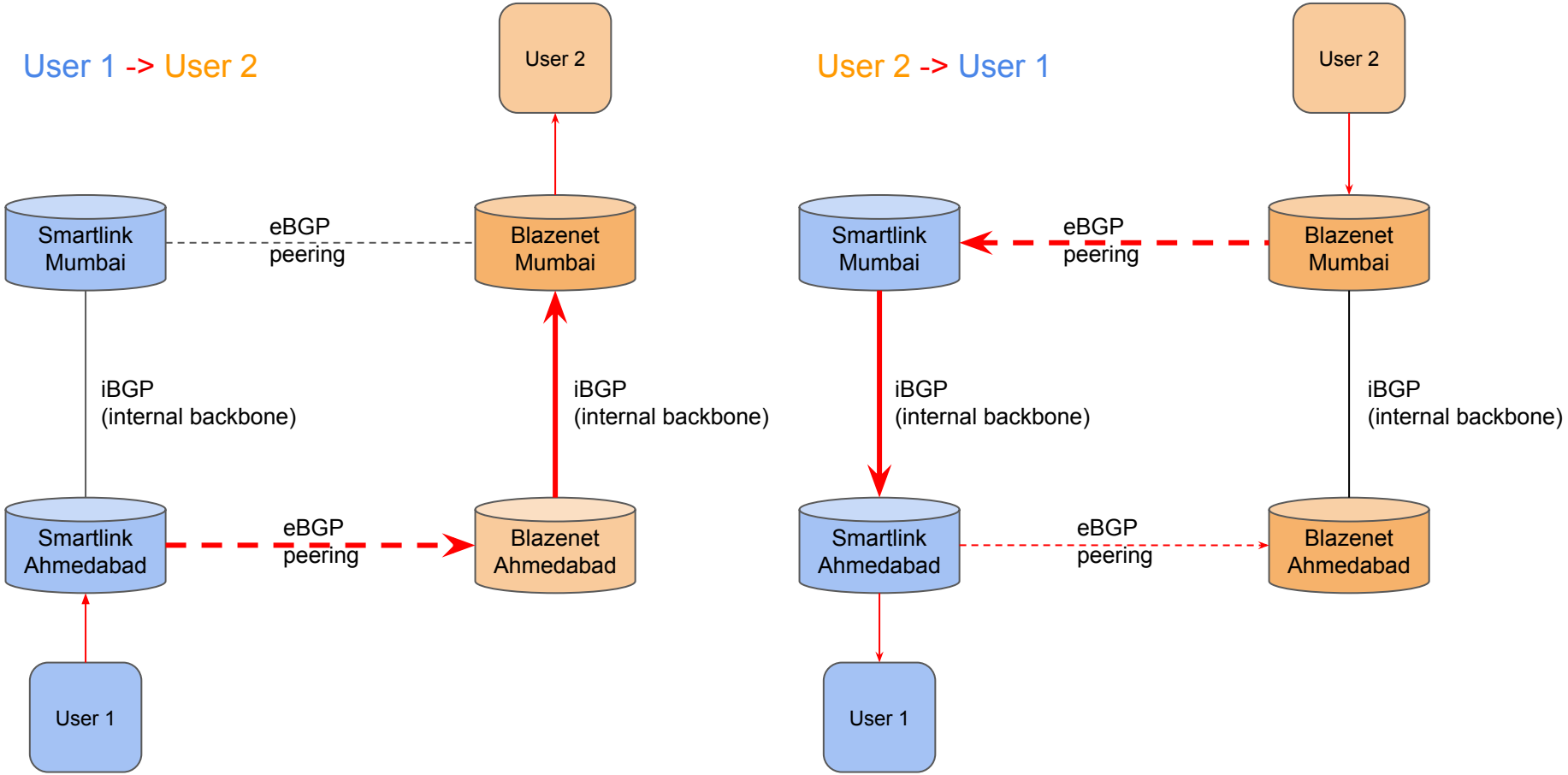
5. To become a tier1 / transit free network, you must be peered to all other tier 1 transit free networks!

Hot potato Vs Cold potato routing

Hot potato

- Potato is “hot” and hand off as quickly as possible - nearest exit.
- Most of networks (excluding large content networks) by default follow it
- Results in asymmetric routing at longer distances




User 1 -> User 2
User 2
Smartlink Mumbai
eBGP peering
Blazenet Mumbai
iBGP (internal backbone)
iBGP (internal backbone)
Smartlink Ahmedabad
eBGP peering
Blazenet Ahmedabad
User 1
User 2 -> User 1
User 2
Smartlink Mumbai
eBGP peering
Blazenet Mumbai
iBGP (internal backbone)
iBGP (internal backbone)
Smartlink Ahmedabad
eBGP peering
Blazenet Ahmedabad
User 1

Cold potato

- Potato is “cold” and hence carry it far as possible
- Often followed by content networks who want to carry traffic on their own backbone instead of “nearest exit” for optimisations
- Can result in case if routing announcements are not consistent.

Real world example of hot potato

CenturyLink

LONDON ENGLAND Traceroute results for:
65.19.151.10 (65.19.151.10)

Tracing route to 65.19.151.10
1 ae-1-3111.edge4.London1.Level3.net (4.69.141.230) 0ms 0ms 0ms
2 10ge2-1.core1.lon2.he.net (216.66.88.237) 0ms 0ms 24ms
3 100ge4-1.core1.nyc4.he.net (72.52.92.166) 67ms 68ms 68ms
4 100ge8-1.core1.sjc2.he.net (184.105.81.218) 132ms 134ms 133ms
5 100ge8-2.core3.fmt1.he.net (72.52.92.57) 191ms 170ms 132ms

HURRICANE ELECTRIC INTERNET SERVICES

Looking Glass

Welcome to Hurricane Electric's Network Looking Glass. The information provided by and the support of this service are on a best effort basis. These are some of our routers at core locations within our network. We also operate a public route server accessible via telnet at route-server.he.net.

Show options

core3.fmt1.he.net> traceroute 4.69.141.230 source 216.218.252.128 numeric

Target	4.69.141.230
Hop Start	1
Hop End	30

Hop	Packet 1	Packet 2	Packet 3	Hostname
1	23 ms	1 ms	<1 ms	**100ge6-1.core1.sjc2.he.net** (72.52.92.58)
2	2 ms	<1 ms	1 ms	**100ge0-36.core2.sjc2.he.net** (184.104.192.214)
3	2 ms	1 ms	1 ms	**level3-as3356.port-channel9.core2.sjc2.he.net** (65.19.191.118)
4	1 ms	1 ms	1 ms	**ae-35-3505.ebr3.SanJose1.Level3.net** (4.69.219.62)
5	88 ms	67 ms	88 ms	**ae-10-10.ebr4.NewYork6.Level3.net** (4.69.160.77)
6	73 ms	71 ms	68 ms	**ae-1-11.ebr3.NewYork6.Level3.net** (4.69.209.37)
7	150 ms	139 ms	134 ms	**ae-41-41.ebr1.London2.Level3.net** (4.69.140.9)
8	133 ms	133 ms	136 ms	**ae-40-40.ebr1.London1.Level3.net** (4.69.140.13)
9	132 ms	133 ms	195 ms	**ae-1-3111.edge4.London1.Level3.net** (4.69.141.230)

Entry cached for another 59 seconds. 2020-10-29 19:55:44 UTC

BGP best path selection

BGP best path selection

- Comes into picture only when comparing route learnt from BGP only and NOT BGP Vs other protocol.
- Longest prefix match / most specific prefix always wins and hence BGP path selection rules apply when multiple routes exists with the same mask.
- Route1: 10.0.0.0/22 and Route 2: 10.0.0.0/24 - Route 2 will always be prefered over Route 1.

Easy to remember sentence:

We love oranges as oranges means energy in our roots

BGP path selection algorithm

"We love oranges as oranges means energy in our roots"

1. We - Weight - Cisco specific, higher weight wins
2. Love - Local preference, higher localpref wins
3. Oranges - Locally Originated wins (over externally originated)
4. AS (path) - Shorter AS_PATH wins
5. Oranges - Origin code preference i>e>? - Not common these days
6. Means - MED, low MED wins (used in multiple sessions across same set of ASNs)
7. Energy - eBGP wins over iBGP (enables hot potato)
8. IN - Lowest IGP metric to next-hop
9. Our - Oldest route
10. Roots - Lowest router-id

Typical ISPs preference for sending traffic

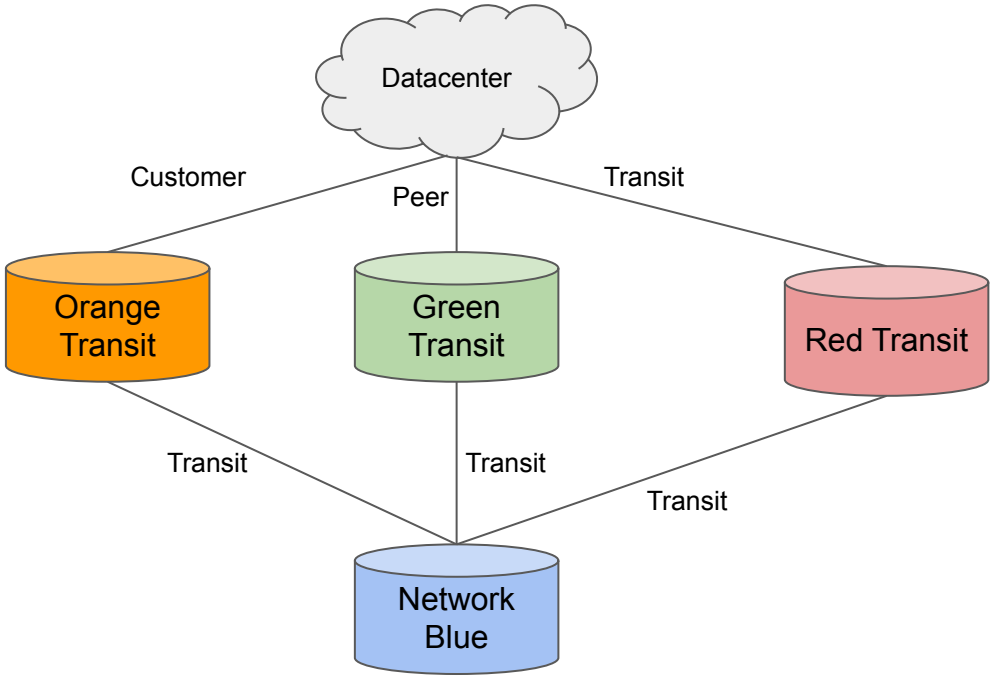
1. Send traffic from customer path if possible
2. Send traffic from peering path if possible
3. Send traffic traffic from transit (least prefered)

Typical ISPs preference for sending traffic

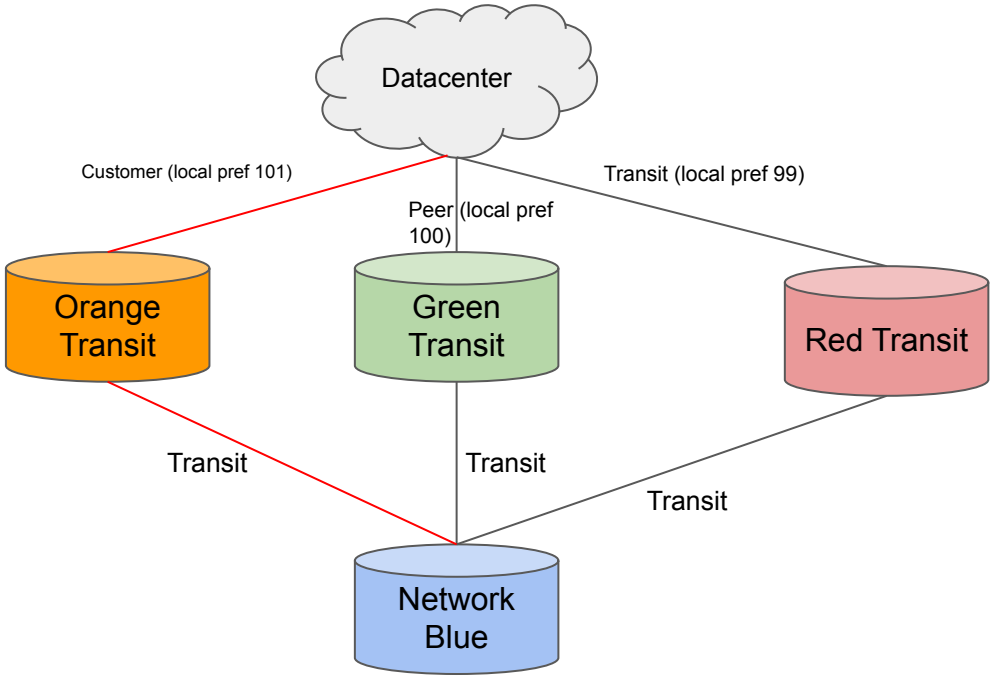
1. Send traffic from customer path if possible <- Highest local preference
2. Send traffic from peering path if possible <- Medium local preference
3. Send traffic traffic from transit (least prefered) <- Lowest local preference

Remember local preference wins over AS_PATH length

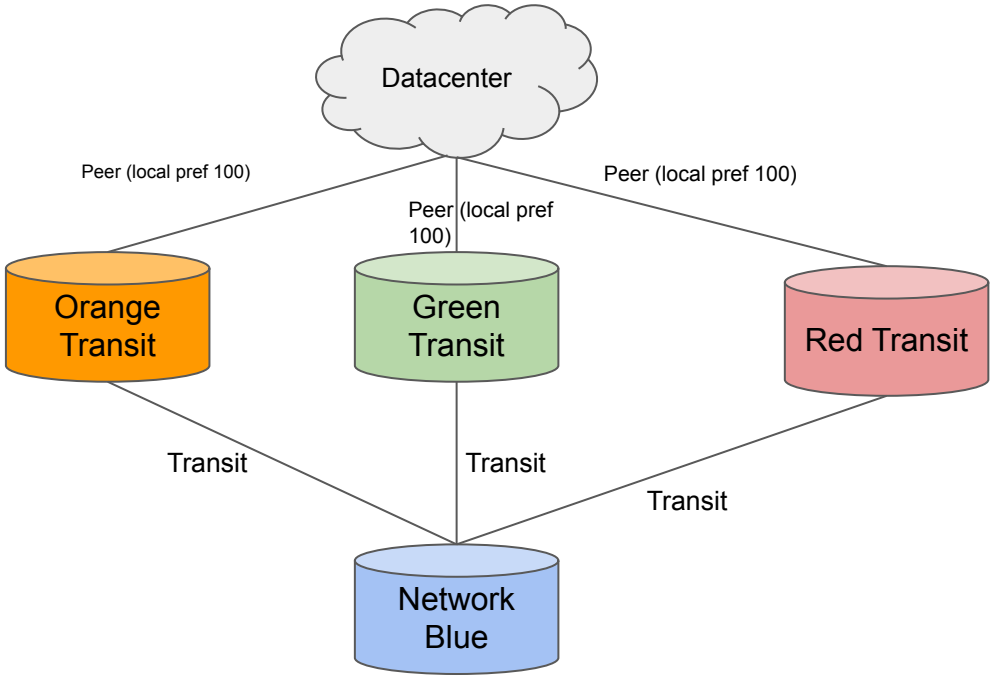
Why my prepends don’t always work?



Why my prepends don’t always work?



Case when prepend can work



Multihoming & traffic balancing

Remember routes flow direction is opposite to traffic flow

Routing information flow - opposite of traffic flow

1. Routes announces to peer bring inbound traffic

2. Routes learnt from peer impacts outbound traffic

Outbound traffic control

1. Set high localpref on customer routes
2. Set second highest local pref on PNI (direct peering over private ports)
3. Set third highest local pref on peering routes via IXP
4. Set lowest local pref on transit routes

Inbound traffic control

1. Announce aggregate pools to all transits like e.g 10.0.0.0/22.

2. Announce a more specific pool to individual transit like 10.0.0.0/23 to transit 1 + 10.0.2.0/23 to transit 2 to load balance inbound traffic

3. Announce most specific to all peers like e.g 10.0.0.0/24 + 10.0.1.0/24 + 10.0.2.0/24 + 10.0.3.0/24

4. /24 is just for example. Idea is to have three levels of aggregation. Do not de-aggregate your pools all the way to /24. If you have a /21, announce /21 to all transits, /22 to selective transit to balance traffic and /23 to peers.

But what if I simply announce same set of prefixes to all?

BGP AS_PATH selection reminder...

1. Localpref wins over AS_PATH and you may get in cases where despite of your preference you do not get max traffic.

2. If local pref is same, AS_PATH length wins and that may not give best results

Route hijack

What happens when a route in hijacked

1. A network starts “originating” a prefix which does not belongs to it

2. Can happen due to a mistake/misconfiguration or intentional

3. Origin AS changes in this case

4. In Indian networks it’s often caused by people re-writing AS_PATH while prepending in the Mikrotik config.
   set-bgp-prepend (1, 2, 3...) Vs set-bgp-prepend-path

Route leak

What happens when a route in leaked

1. A network starts “leaking” routes which does not belongs to it’s customers to its peers or upstream

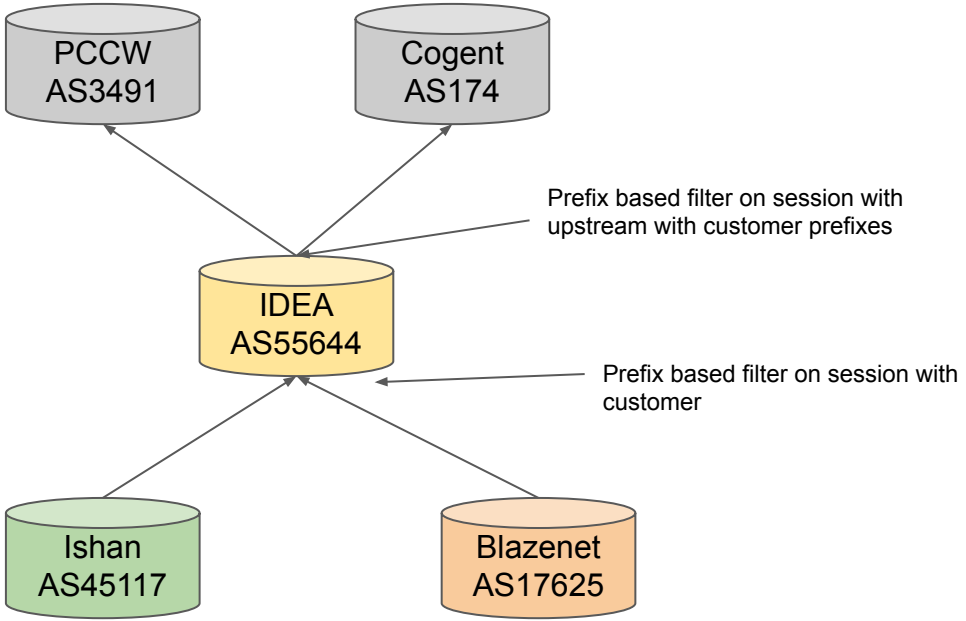
2. Can happen due to a mistake/misconfiguration or intentional

3. Origin AS remains same in this case

4. Often caused because people use BGP prefix list without AS_PATH matching on their peers and upstreams. Either AS_PATH match should be used OR simply use BGP communities internally

A real world route leak case in India which has a visible impact...

Example case of route leak cases due misconfiguration



Problem in this approach...

1. Vodafone/IDEA AS55644 is just matching customer prefixes and announcing to upstream. Since downstreams like Ishan & Blazenet are multi-homed, IDEA may learn those routes indirectly via Tata AS4755 or Airtel AS9498 and may still “leak”

2. Turns out at some stage both Ishan and Blazenet stop purchasing IDEA transit and IDEA left the older config with prefix filters with upstream.

3. This results in IDEA leaking 271 routes to its upstreams like PCCW Global AS3491 and Cogent AS174.

A look at AS_PATHs of leaked routes

Routes belong to Ishan:

```
27.54.160.0/24|328145 3491 55644 4755 45117
27.54.160.0/24|12779 174 55644 55644 55644 55644 55644 55644 55644 55644 55644 4755 45117
27.54.160.0/24|1299 3491 55644 4755 45117
27.54.160.0/24|6894 174 55644 55644 55644 55644 55644 55644 55644 55644 55644 4755 45117
27.54.161.0/24|328145 3491 55644 4755 45117
27.54.161.0/24|12779 174 55644 55644 55644 55644 55644 55644 55644 55644 55644 4755 45117
27.54.161.0/24|1299 3491 55644 4755 45117
27.54.161.0/24|6894 174 55644 55644 55644
```

and more!!

A look at AS_PATHs of leaked routes

Routes belong to Blazenet:

```
27.109.13.0/24|328145 3491 55644 4755 45820 17625
27.109.13.0/24|12779 174 55644 55644 55644 55644 55644 55644 55644 55644 55644 4755 45820
17625
27.109.13.0/24|1299 3491 55644 4755 45820 17625
27.109.13.0/24|6894 174 55644 55644 55644 55644 55644 55644 55644 55644 55644 4755 45820
17625
120.72.94.0/24|328145 3491 55644 4755 45820 17625
120.72.94.0/24|1299 3491 55644 4755 45820 17625
202.131.112.0/24|328145 3491 55644 4755 17625
202.131.112.0/24|1299 3491 55644 4755 17625
```

Does it has any real world impact?

Yes, bad routing!

Trace from Munich Germany to Blaznet in Gujarat, India

```
anurag@devops01 ~/temp> mtr -wr 202.131.112.1
Start: 2020-10-30T04:32:10+0530
HOST: devops01.muc.anuragbhatia.com          Loss%   Snt   Last   Avg  Best  Wrst StDev
  1.|-- 10.20.70.1                             0.0%    10    0.1   0.1   0.1   0.2   0.0
  2.|-- gw.giga-dns.com                        30.0%    10    0.5   0.6   0.5   0.8   0.1
  3.|-- gw02.giga-hosting.biz                   0.0%    10    0.6   2.9   0.6  22.1   6.8
  4.|-- mcn-b3-link.telia.net                   0.0%    10    0.8   1.0   0.8   1.4   0.2
  5.|-- ffm-bb2-link.telia.net                  0.0%    10   11.6  12.9  11.6  20.7   2.8
  6.|-- ffm-b5-link.telia.net                   0.0%    10   11.5  11.7  11.5  12.3   0.3
  7.|-- ffm-b4-link.telia.net                   0.0%    10   11.3  11.6  11.3  12.5   0.4
  8.|-- pccw-ic-319976-ffm-b4.c.telia.net       0.0%    10   12.1  12.0  11.6  13.3   0.5
  9.|-- HundredGE0-3-0-0.br03.sin03.pccwbtn.net 0.0%    10  246.6 246.8 246.6 247.6   0.3
 10.|-- HundredGE0-3-0-0.br03.sin03.pccwbtn.net 0.0%    10  246.5 246.7 246.4 247.2   0.3
 11.|-- 63-218-249-146.static.pccwglobal.net    0.0%    10  293.8 292.6 292.4 293.8   0.4
 12.|-- 223.196.22.229                          0.0%    10  286.3 286.4 286.2 286.6   0.1
 13.|-- 223.196.6.26                           0.0%    10  282.1 282.1 282.0 282.4   0.1
 14.|-- 223.196.24.33                          0.0%    10  284.8 284.9 284.8 285.2   0.1
 15.|-- 14.141.20.29.static-vsnl.net.in        0.0%    10  187.7 186.0 185.6 187.7   0.6
 16.|-- ???                                   100.0    10    0.0   0.0   0.0   0.0   0.0
 17.|-- 59.165.224.218.man-static.vsnl.net.in  0.0%    10  193.4 193.4 193.3 193.5   0.1
 18.|-- 202.131.98.146                         0.0%    10  192.7 193.8 192.7 203.1   3.3
 19.|-- 202.131.101.206                        0.0%    10  192.4 192.5 192.4 192.6   0.1
 20.|-- 202.131.113.202                        0.0%    10  209.4 198.8 191.2 228.5  12.3
 21.|-- 202.131.107.26                         0.0%    10  190.9 190.9 190.7 191.0   0.1
 22.|-- 202.131.112.1                          0.0%    10  191.0 191.0 190.7 192.3   0.5
anurag@devops01 ~/temp>
```

Telia (Germany) > PCCW (Germany) > PCCW (Singapore) > IDEA (India) > Tata Comm AS4755 (India) > Tata Comm (Gujarat) > Blazenet (Gujarat)

Better way to filter downstreams

- Use BGP communities. These are simply tags which can be applied on the routes and then these tags can be matched & actions can be performed.
- Consider tagging routes with type of network (customer, peer or upstream) + location.
- Prefix list based filters only on the session with downstream. Beyond that simply tag matching across all peers / upstream. Prefix list should be used only once.

Some of tools to troubleshoot routing issues

1. bgp.he.net :-)
2. rt-bgp.he.net <- new & shows realtime data
3. Oregon Route Views - telnet route-views.oregon-ix.net
4. RIPE Atlas - atlas.ripe.net (web UI as well as command line tools)
5. Various looking glasses
6. BGPalerter - https://github.com/nttgin/BGPalerter / Slides / Video

Live quick demo of tools...

Questions?

anurag@he.net
https://he.net