How far can you go with IX Route Servers only?

Ben Cartwright-Cox
BGP.Tools / Port 179 LTD

a6fc0eaffbba21aa7c46c1394cfcd118

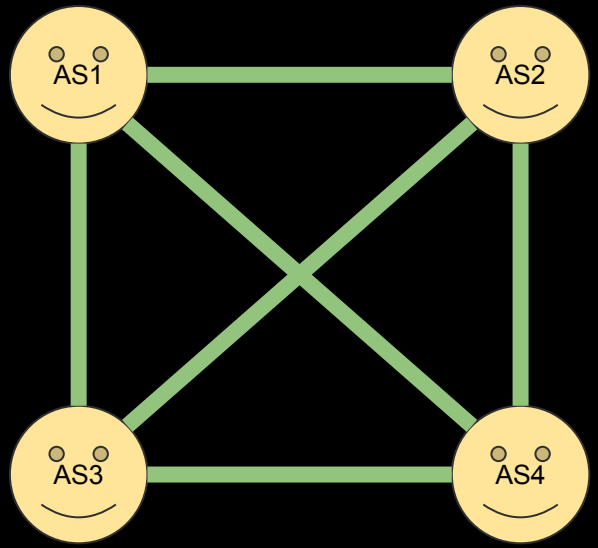
As I am sure you know, IX's (almost always) have services

- AS112
  - Helps soak up some DNS queries that should not have escaped a DNS recursor
- NTP/Time
  - Useful if you need to provide your router with local time
- Route Collectors
  - (sort of?) Useful if you want to double check a members policy
  - Most IX's route collectors (if they have them) are in some state of disrepair
- Route Servers
  - The major service that on the IX LAN!

a6fc0eaffbba21aa7c46c1394cfcd118

IX Route servers

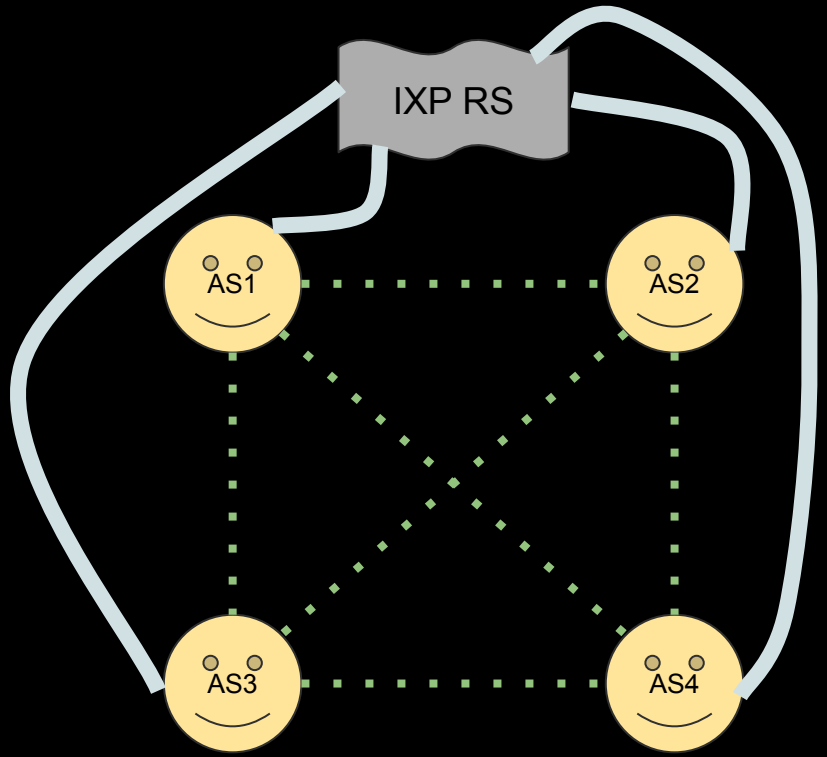
- Attempts to solve the problem of peers needing to create BGP sessions with nearly every member of a exchange they want to exchange traffic with
- This is bad because networks are lazy/busy, and may not setup sessions when asked



a6fc0eaffbba21aa7c46c1394cfcd118

IX Route Servers (RS)

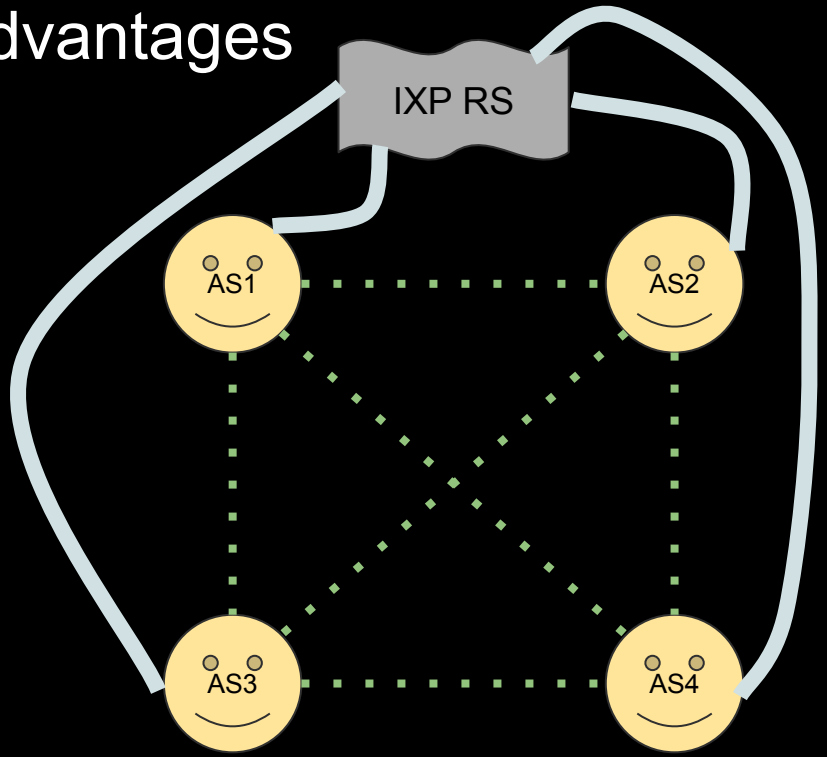
- Instead, everyone peers with the RS, and it distributes the routes sent to it by other members to everyone.
- Since everyone is on the same subnet/layer 2. The BGP next hop is not changed.
- The RS can control terabits of traffic with a 100mbps port.



a6fc0eaffbba21aa7c46c1394cfcd118

IX Route Servers (RS) Bonus Advantages

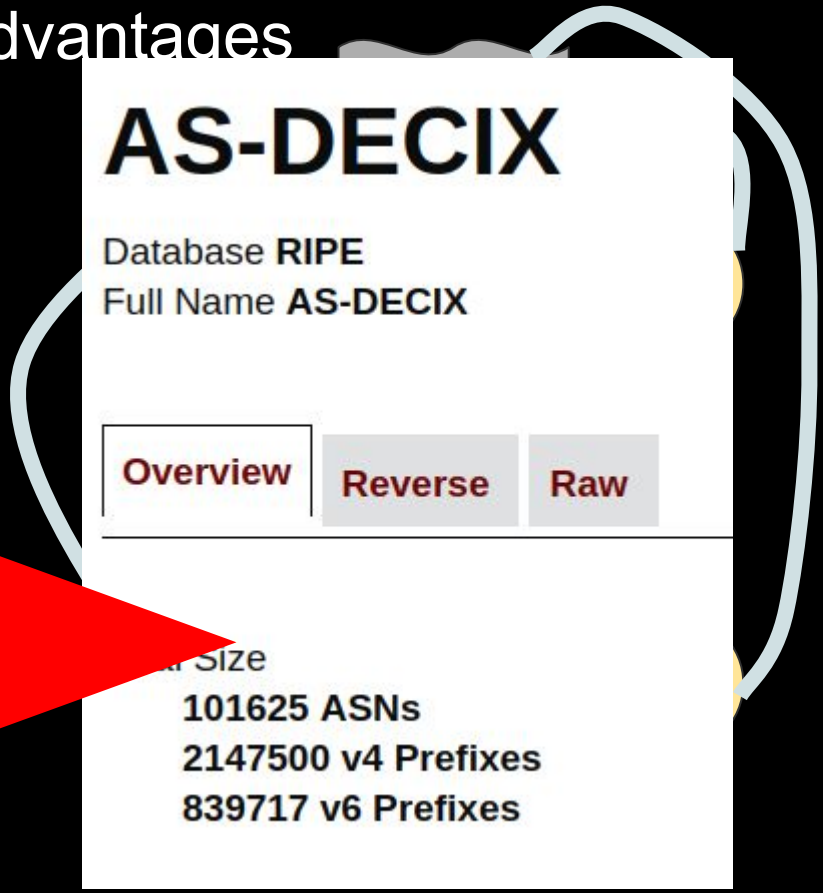
- Your average person (despite what they say publicly) does not have a good/secure BGP peer configuration
- Modern RS are ***far safer*** to peer on than bi-lat peering, due to better IRR/RPKI/Sanity/PeerLock automations
- Yes you may have the magic config, but most of the IX is importing almost anything they are sent



a6fc0eaffbba21aa7c46c1394cfcd118

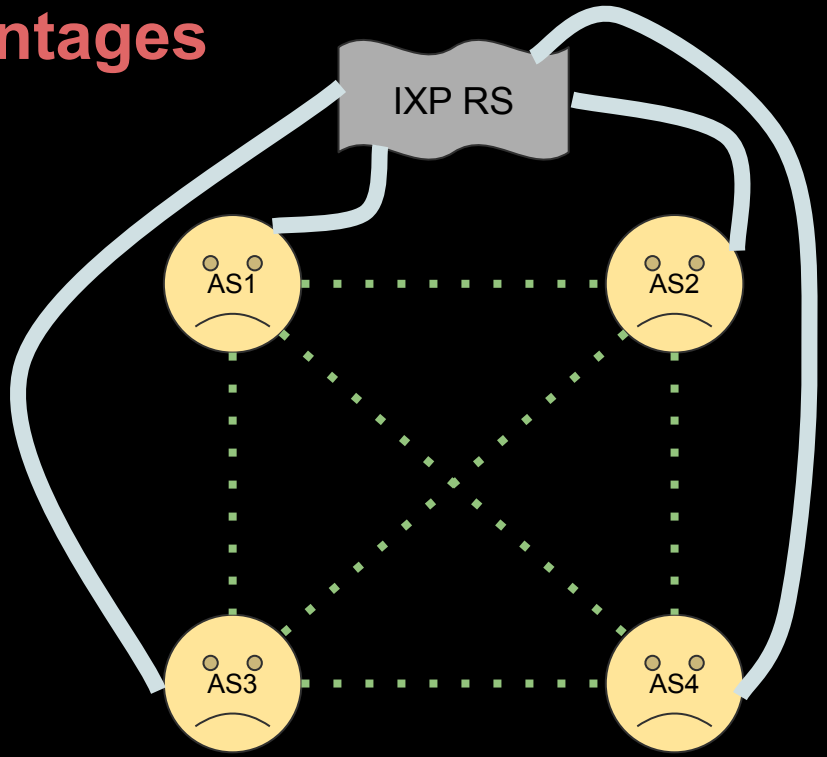
IX Route Servers (RS) Bonus Advantages

- Your average person (despite what they say publicly) does not have a good/secure BGP peer configuration
- Modern RS are ***far safer*** to peer on than bi-lat peering, due to better IRR/RPKI/Sanity/PeerLock automations
- Yes you may have the magic config, but most of the IX is importing almost anything they are sent



IX Route Servers (RS) **Disadvantages**

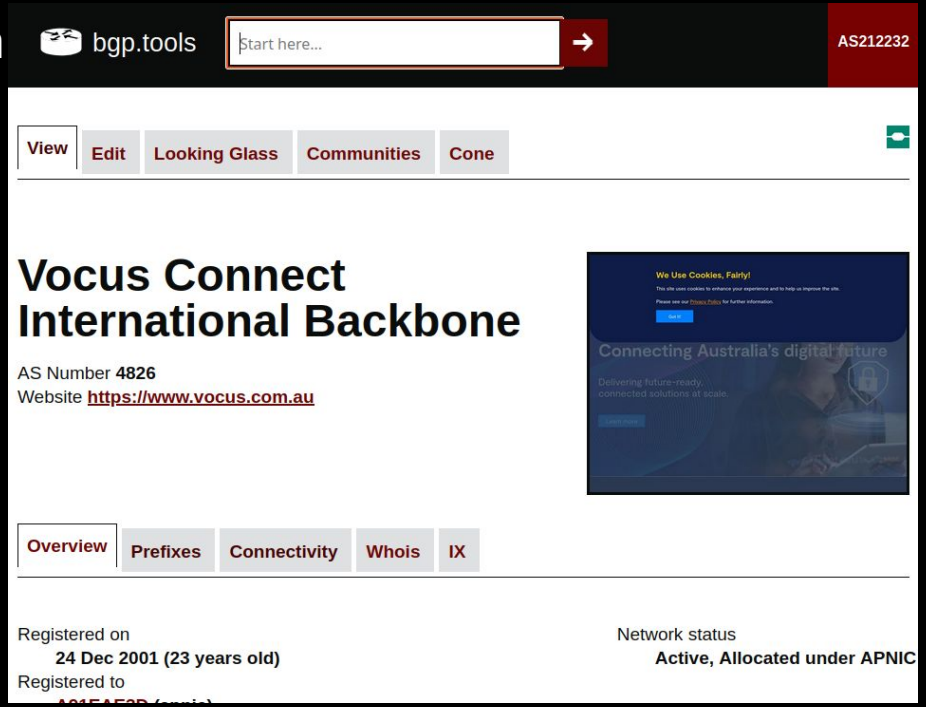
- ***Some*** networks import/export everything to the Route Server
- ***Some*** networks just import from the route server, but don't send their routes to them
    - There are some good reasons to do this, some networks are very sensitive poor routing and Route Servers are considered high risk
- ***Others*** will export/send routes, but not import anything



a6fc0eaffbba21aa7c46c1394cfcd118

Quick tl;dr of what bgp.tools is

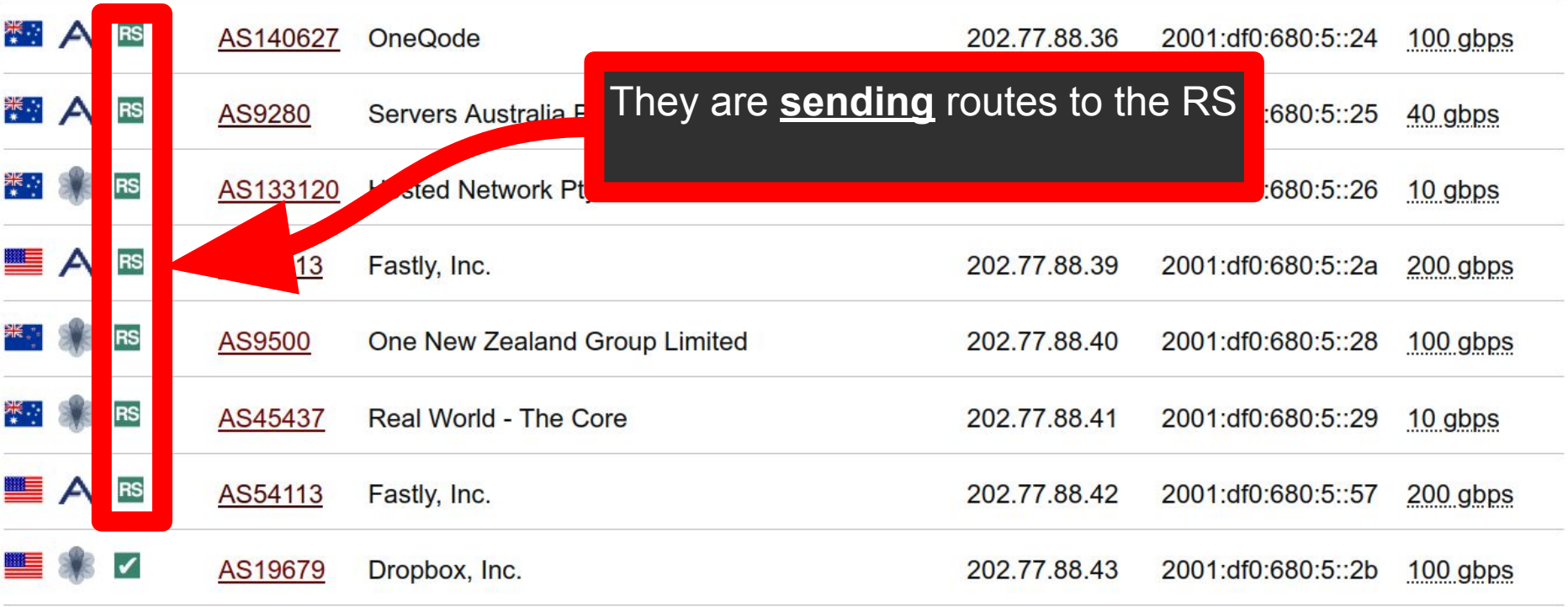
- General internet routing information site
- Runs its own BGP Route Collector that you should feed!
- Has a large IX presence for route collection
- Can also offer rapid BGP (and other BGP related things) network monitoring



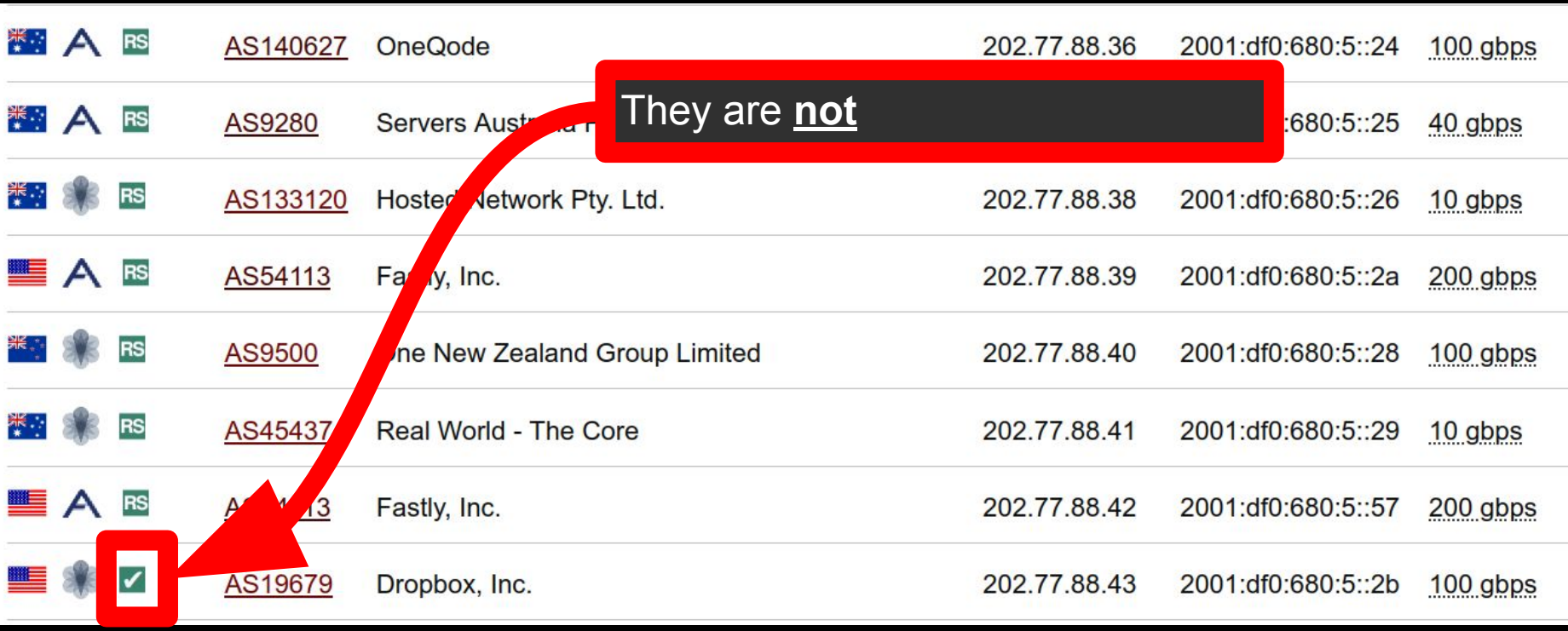
RS	AS140627	OneQode	202.77.88.36	2001:df0:680:5::24	100 gbps
RS	AS9280	Servers Australia Pty Ltd	202.77.88.37	2001:df0:680:5::25	40 gbps
RS	AS133120	Hosted Network Pty. Ltd.	202.77.88.38	2001:df0:680:5::26	10 gbps
RS	AS54113	Fastly, Inc.	202.77.88.39	2001:df0:680:5::2a	200 gbps
RS	AS9500	One New Zealand Group Limited	202.77.88.40	2001:df0:680:5::28	100 gbps
RS	AS45437	Real World - The Core	202.77.88.41	2001:df0:680:5::29	10 gbps
RS	AS54113	Fastly, Inc.	202.77.88.42	2001:df0:680:5::57	200 gbps
✓	AS19679	Dropbox, Inc.	202.77.88.43	2001:df0:680:5::2b	100 gbps

[1] https://bgp.tools/ixp/EdgeIX+-+Sydney

a6fc0eaffbba21aa7c46c1394cfcd118

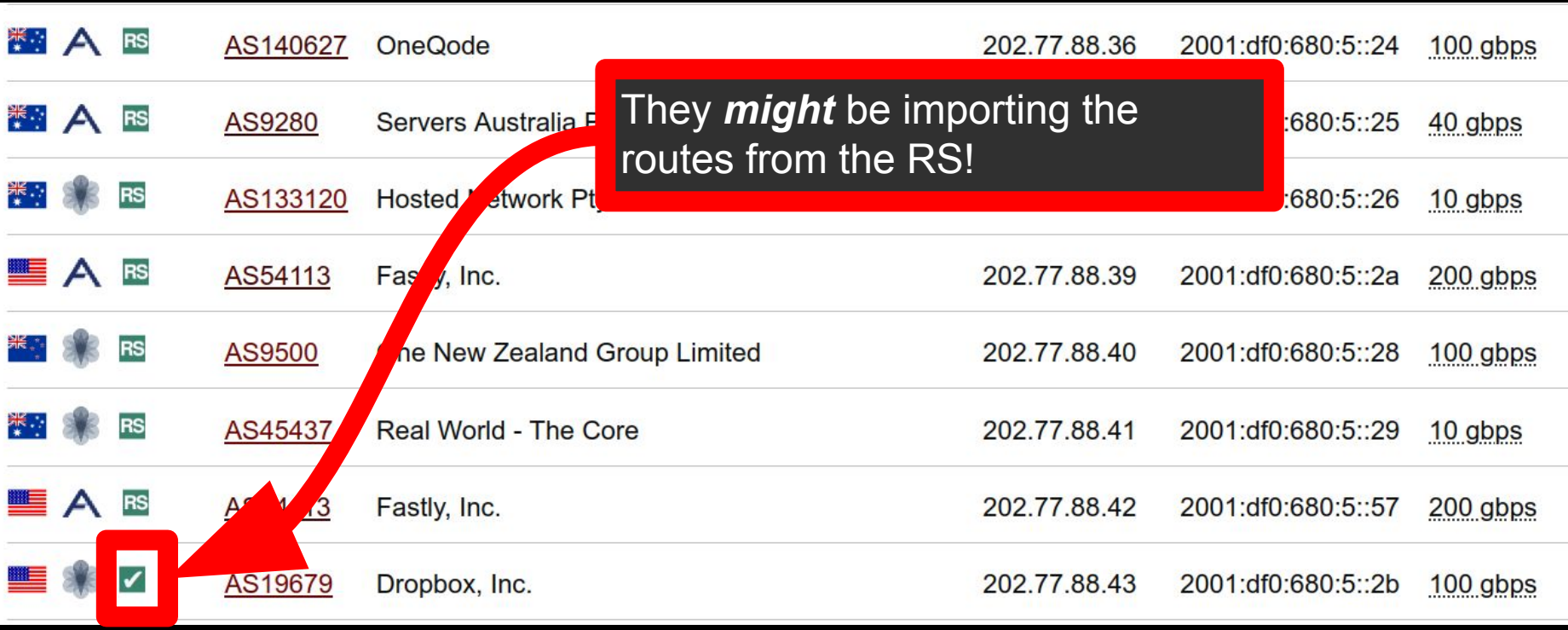


a6fc0eaffbba21aa7c46c1394cfcd118

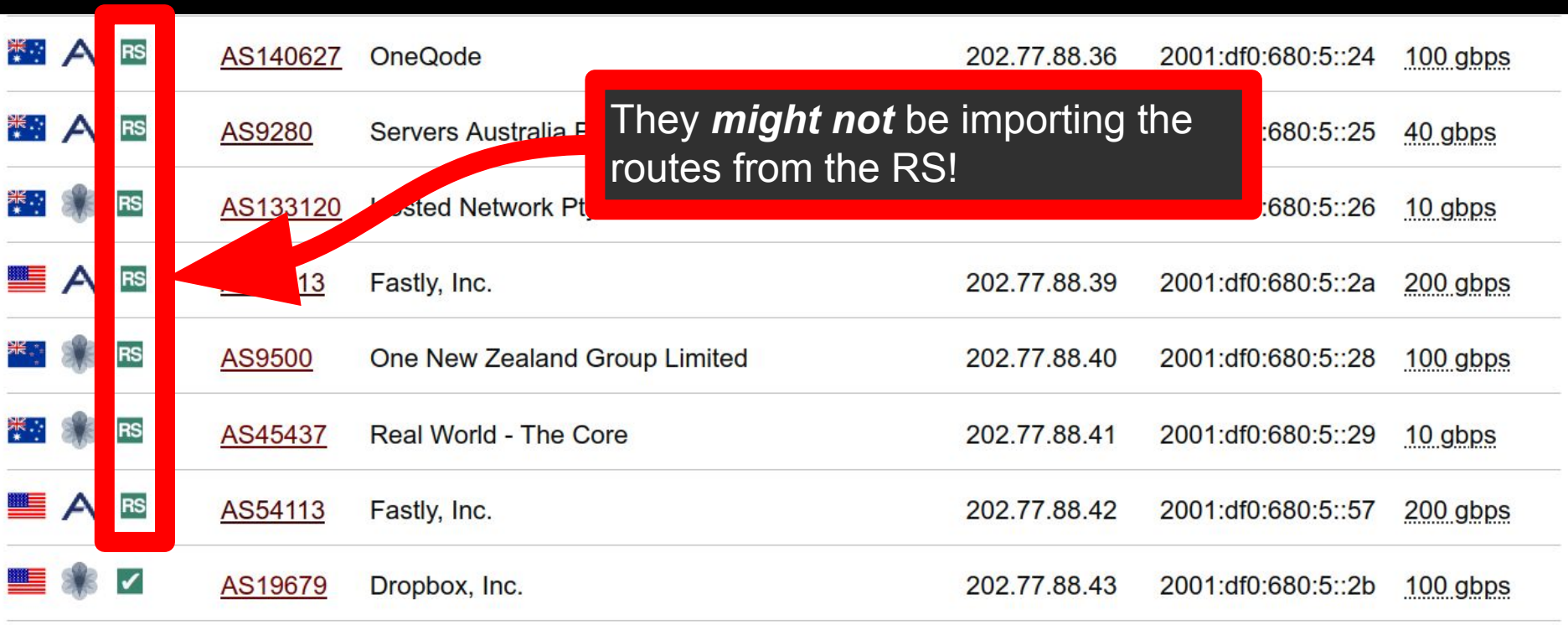


AS140627	OneQode	202.77.88.36	2001:df0:680:5::24	100 gbps
AS9280	Servers Aust[illegible]	[illegible]:680:5::25		40 gbps
AS133120	Hosted [illegible]Network Pty. Ltd.	202.77.88.38	2001:df0:680:5::26	10 gbps
AS54113	Fa[illegible]y, Inc.	202.77.88.39	2001:df0:680:5::2a	200 gbps
AS9500	[illegible]ne New Zealand Group Limited	202.77.88.40	2001:df0:680:5::28	100 gbps
AS45437	Real World - The Core	202.77.88.41	2001:df0:680:5::29	10 gbps
A[illegible]13	Fastly, Inc.	202.77.88.42	2001:df0:680:5::57	200 gbps
AS19679	Dropbox, Inc.	202.77.88.43	2001:df0:680:5::2b	100 gbps

a6fc0eaffbba21aa7c46c1394cfcd118



a6fc0eaffbba21aa7c46c1394cfcd118



a6fc0eaffbba21aa7c46c1394cfcd118

How good is the reach for RS's for outbound heavy nets?

- This is the most simple thing to answer, because all you need to do is get all of the RIB's of all of the IXs RS's you are on (bgp.tools is currently on 113~ IXs), and perform a total unique route count vs the full table

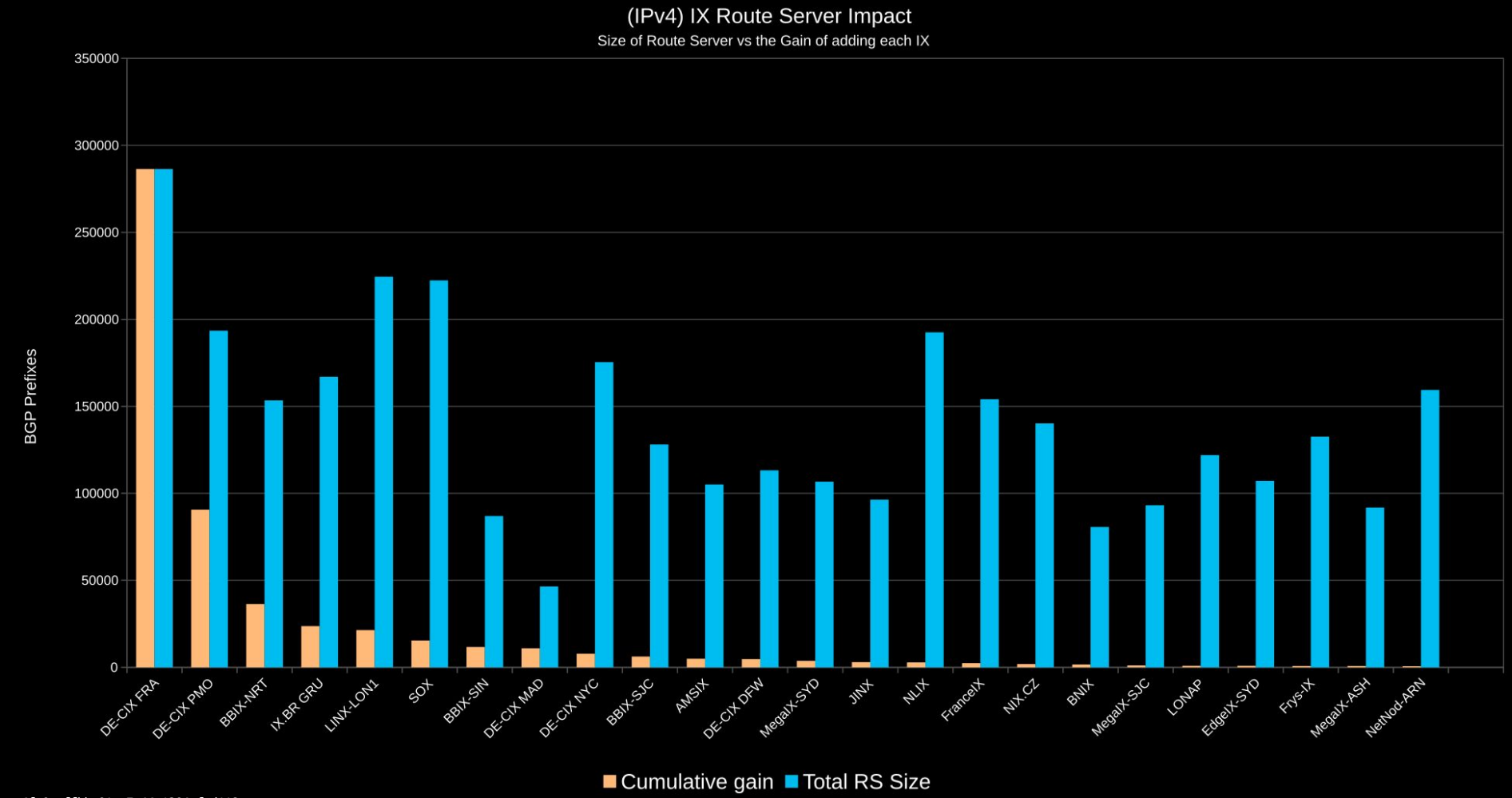
- A full table as of the time of writing is
  - IPv4: 986,502 routes
  - IPv6: 222,168 routes

"Total" Outbound RS Reachability

- IPv4 - 56.6%
  - 557,996 Reachable via RS
  - 986,502 Total Prefixes in DFZ

- IPv6 - 60%
  - 133,401 Reachable via RS
  - 222,168 Total Prefixes in DFZ

a6fc0eaffbba21aa7c46c1394cfcd118



a6fc0eaffbba21aa7c46c1394cfcd118

Obvious (to me) cavates to this

1. This is a survey of 118 exchanges (a lot of exchanges)
    a. You are not likely to go out and build a network like this to offload this traffic, you would just buy more transit instead
    b. Also this 118 exchanges is missing a few notable groups of exchanges, notably, Equinix

2. Just because it's 60% of your prefixes does not mean it is 60% of your traffic
    a. As far as I can tell for eyeball traffic profiles, 50%+ of all traffic is concentrated in just 5 ASNs
        i. Like Meta, Akamai, Google, Netflix, Amazon | (The "Magna" networks?)

3. This calculation does not account for people pushing more specifics into the IX RS for traffic engineering, A technique very popular in markets like Brazil

a6fc0eaffbba21aa7c46c1394cfcd118

Full list of exchanges

AMS-IX
OGIX
INTERIX
NMBINX
SONIX Stockholm
PIT-IX
CINX
QIX Montreal
JINX
DINX
GPC Missouri
STUIX
FD-IX - Indianapolis
EdgeIX - Brisbane
THINX Warsaw
InterLAN-IX
SIX.SK
France-IX AURA
EdgeIX - Sydney
BNIX
IX.br (PTT.br) São Paulo
NIX.SK

YXEIX
SOX Serbia
Frys-IX
EdgeIX - Perth
France-IX Toulouse
LINX LON1
Stuttgart-IX
FIXO
EdgeIX - Auckland
BreizhIX
NIX.CZ
EdgeIX - Adelaide
LONAP
EdgeIX - Melbourne
France-IX Lille
France-IX Marseille
BCIX
GetaFIX Davao
GetaFIX Cebu
IRAQ-IXP
IXP.mk
Lillix

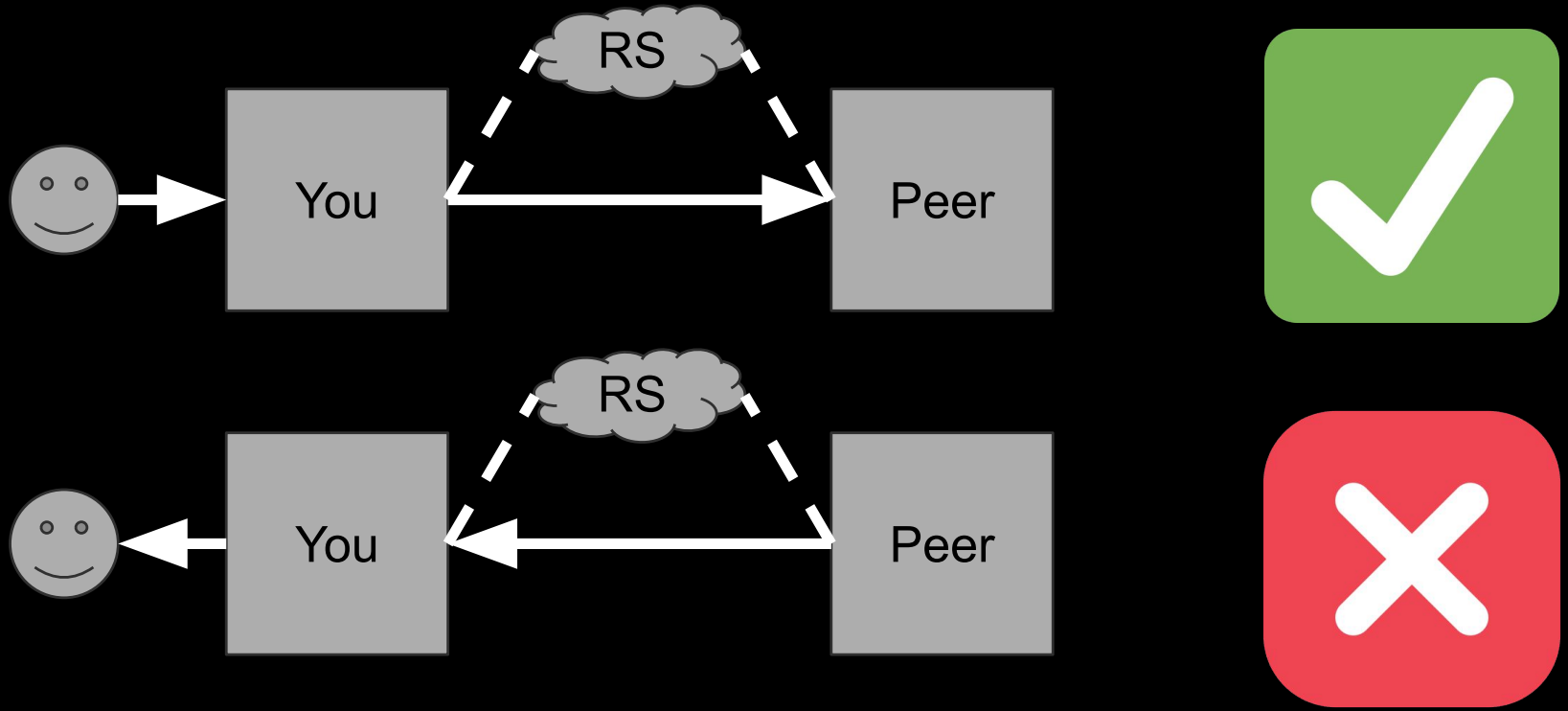
Netnod Helsinki GREEN
Netnod Copenhagen GREEN
MegaIX Chicago
BBIX Amsterdam
BBIX Tokyo
BBIX Singapore
BBIX Dallas
BBIX US-West
BBIX Chicago
BBIX London
MegaIX Auckland
Netnod Helsinki BLUE
France-IX Paris
Netnod Sundsvall
Netnod Gothenburg
MegaIX Perth
MegaIX Singapore
MegaIX Seattle
MegaIX Atlanta
MegaIX Dallas
MegaIX Miami
MegaIX Charlotte
MegaIX Toronto

MegaIX Los Angeles
MegaIX Denver
BIX.BG
DE-CIX Phoenix
DE-CIX Hamburg
MegaIX Las Vegas
MegaIX Ashburn
DE-CIX Palermo
Douala-IX
MegaIX Bay Area
Netnod Copenhagen BLUE
DE-CIX Barcelona
RomandIX
MegaIX Dusseldorf
MegaIX Sofia
NL-ix
MegaIX Melbourne
DE-CIX Madrid
DE-CIX Dusseldorf
DE-CIX Istanbul
Netnod Stockholm BLUE
Netnod Stockholm GREEN

DE-CIX Munich
MegaIX Brisbane
MegaIX New York
DE-CIX New York
DE-CIX Richmond
DE-CIX Chicago
LU-CIX
DE-CIX Dallas
MegaIX Frankfurt
DE-CIX Marseille
B-IX
MegaIX Adelaide
DE-CIX Frankfurt
MegaIX Berlin
MegaIX Hamburg
MegaIX Sydney
DE-CIX Lisbon
ONIX
MSK-IX Moscow
MegaIX Munich
DE-CIX Leipzig

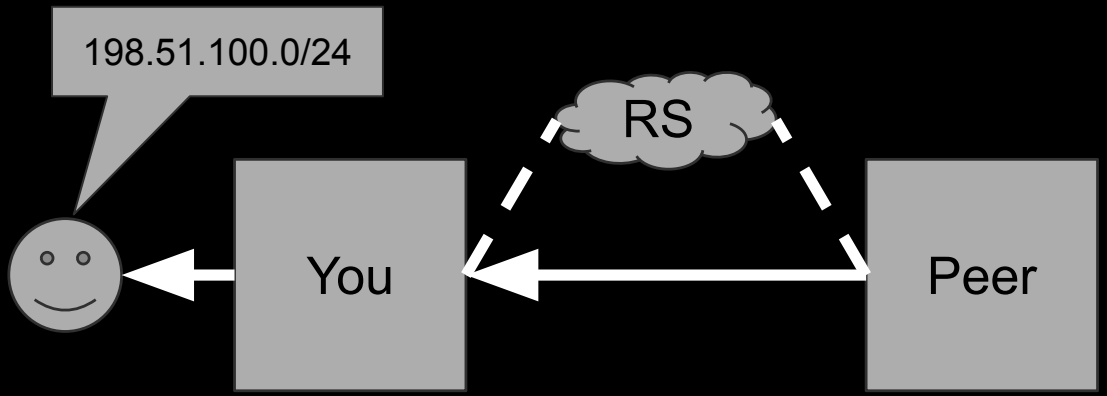
a6fc0eaffbba21aa7c46c1394cfcd118

Traffic directions covered



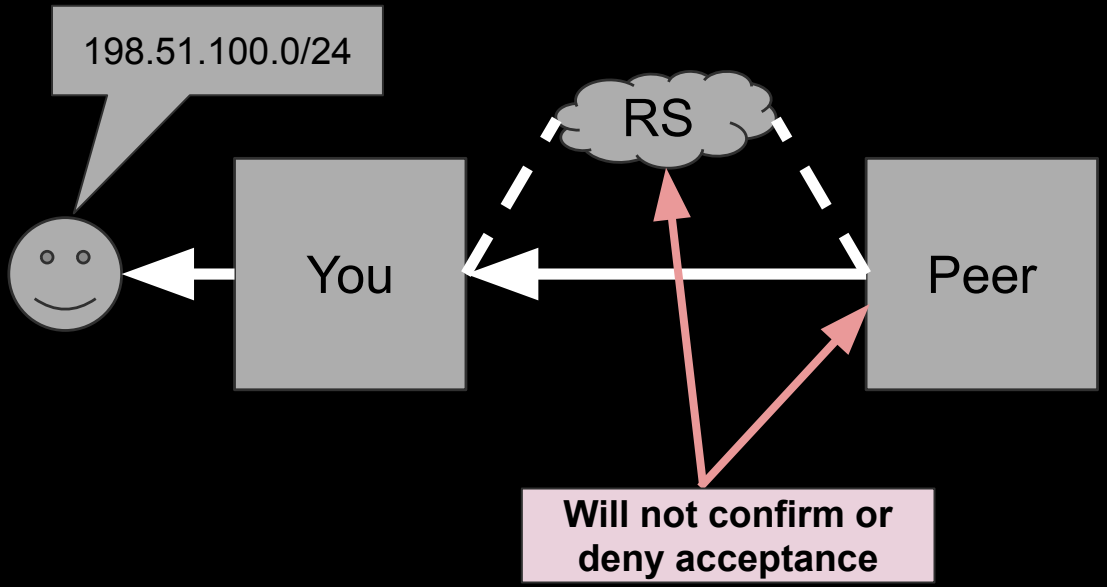
a6fc0eaffbba21aa7c46c1394cfcd118

Inbound will need a different strategy

- 

Inbound will need a different strategy

- BGP (famously) has no feedback signal to a peer for "yes I accepted a prefix"



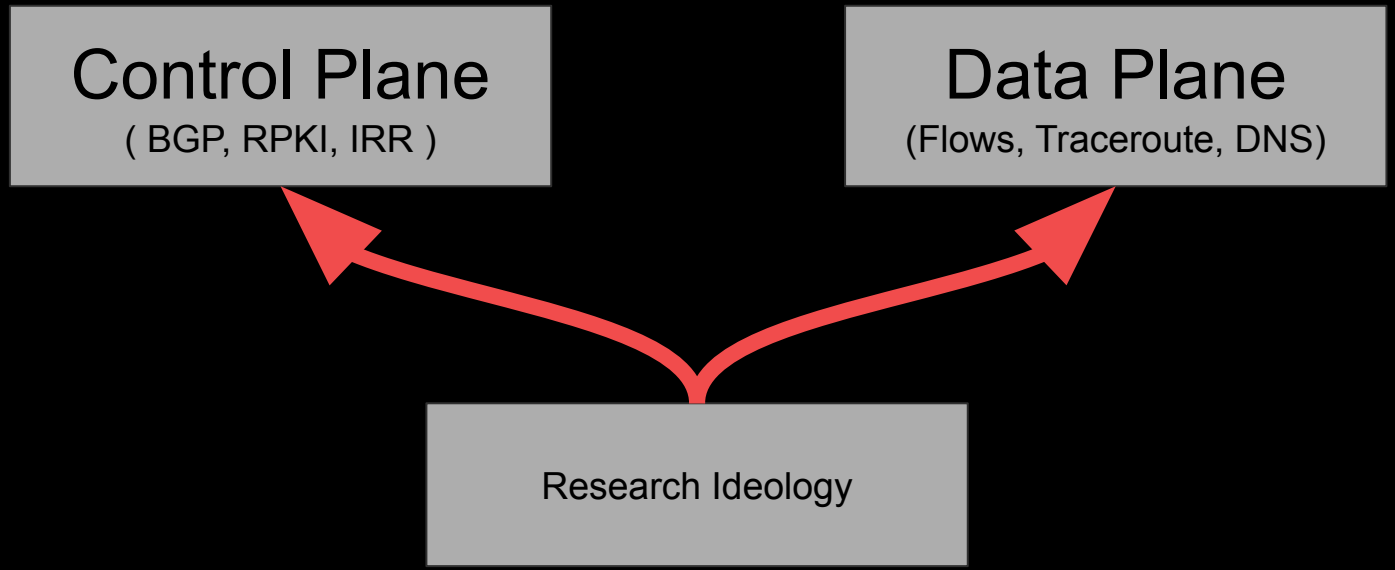
Inbound will need a different strategy

- BGP (famously) has no feedback signal to a peer for "yes I accepted a prefix"

- So we need to find another way to see if they have accepted the prefix
  - One way could be to use the bgp.tools live data set, however even though this is big (3k sessions), that is not big enough to cover the "whole" internet

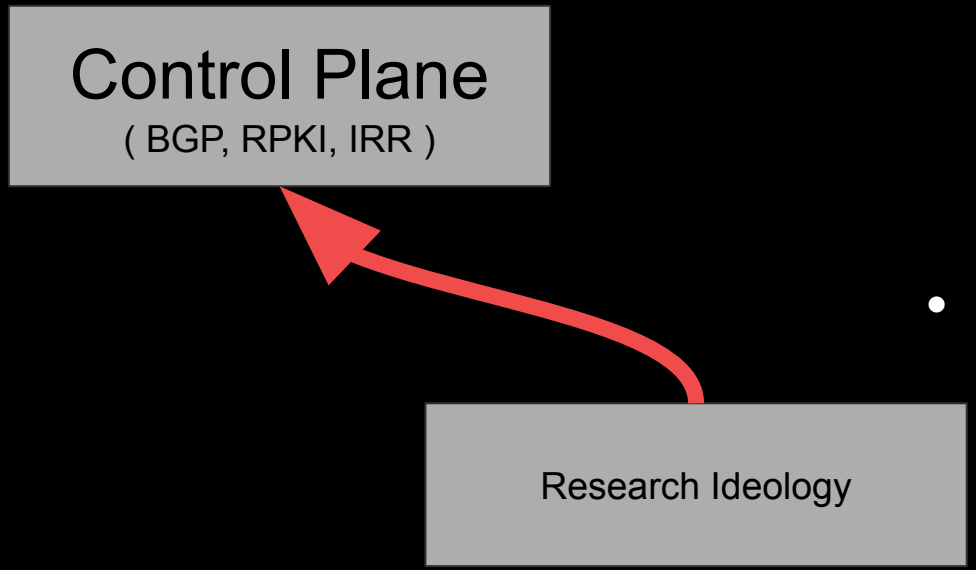
- So we need to figure out a way to do test this on the data plane!

Two ways to look at the internet



a6fc0eaffbba21aa7c46c1394cfcd118

Two ways to look at the internet



Pros:

- Lots of downloadable data
- bgp.tools has 3000~ BGP sessions

Cons:

- Control Plane does not always align with the Data Plane
- You cannot prove that every ISP did something with just 3000~ BGP sessions

a6fc0eaffbba21aa7c46c1394cfcd118

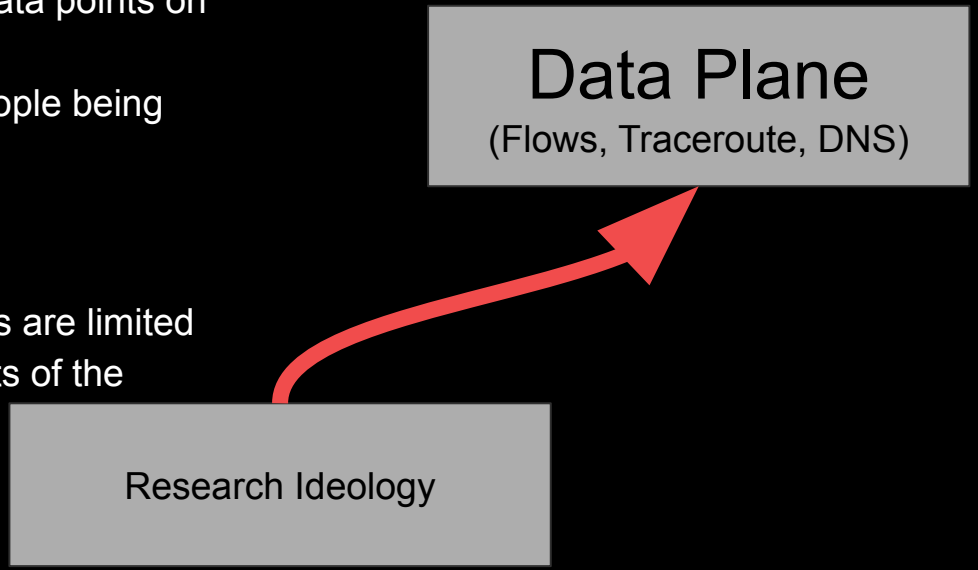
Two ways to look at the internet

Pros:

- Provides a set of more real data points on where traffic goes
- Easier to do without other people being involved

Cons:

- Coverage for testing networks are limited
- Hard to take atomic snapshots of the internet with this
- Firewalls can get in the way



a6fc0eaffbba21aa7c46c1394cfcd118

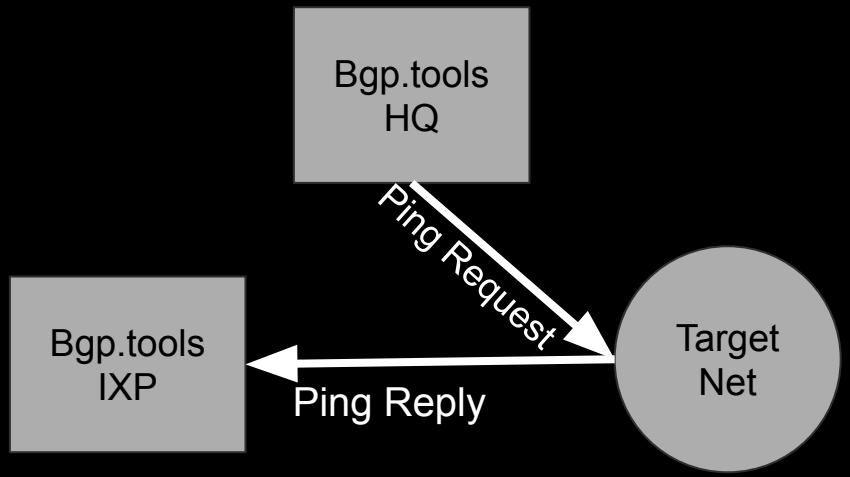
Actual strategy

- Announce a prefix to all route servers
- From somewhere else, send ICMP pings to internet address, with the source IP being that RS only prefix
- If they are accepting a route, it will go back to the IX node, if they don't the reply as nowhere to go

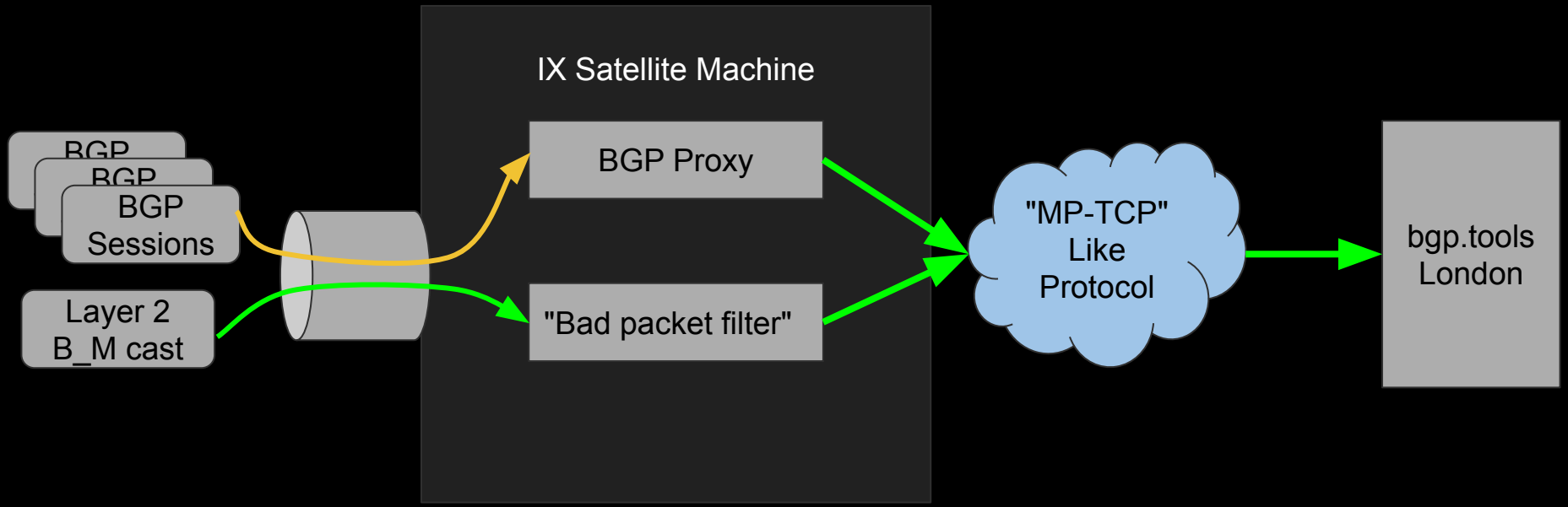
a6fc0eaffbba21aa7c46c1394cfcd118

Actual strategy

- Announce a prefix to all route servers
- From somewhere else, send ICMP pings to internet address, with the source IP being that RS only prefix
- If they are accepting a route, it will go back to the IX node, if they don't the reply as nowhere to go

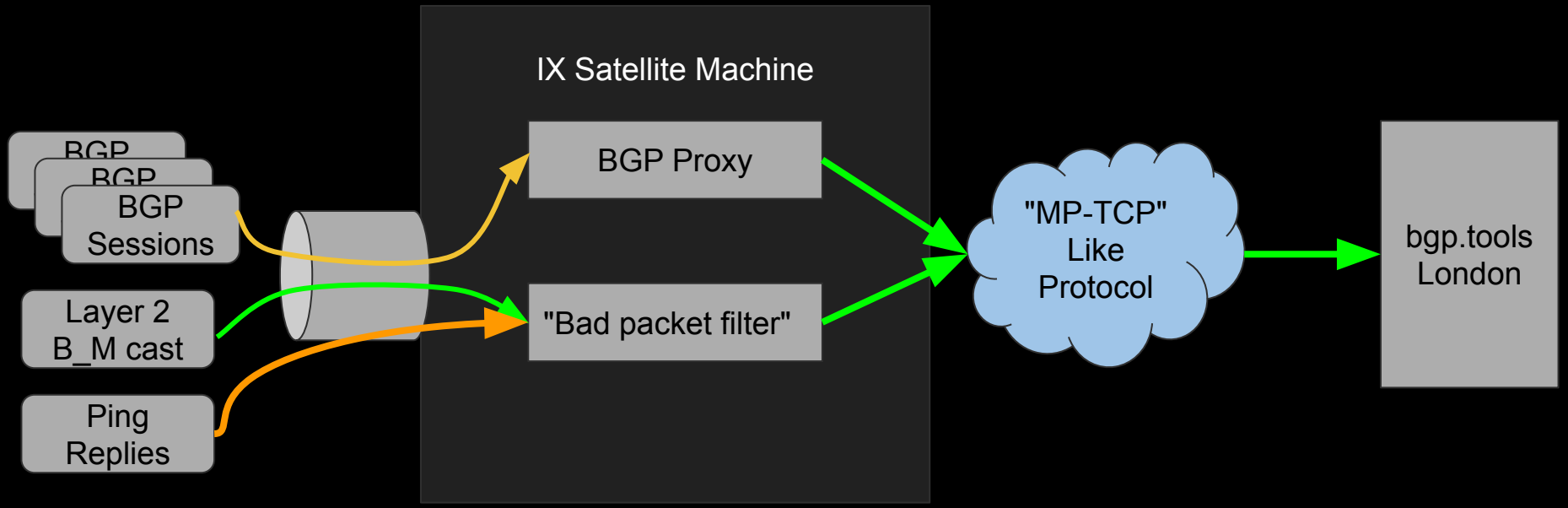


How does that work?



a6fc0eaffbba21aa7c46c1394cfcd118

How does that work?



a6fc0eaffbba21aa7c46c1394cfcd118

Actual strategy

- Announce a prefix to all rout rs
- From somewhere else et address, with the source IP being that RS only
- If they are accepting node, if they don't the reply as nowhere

203.10.63.0/24

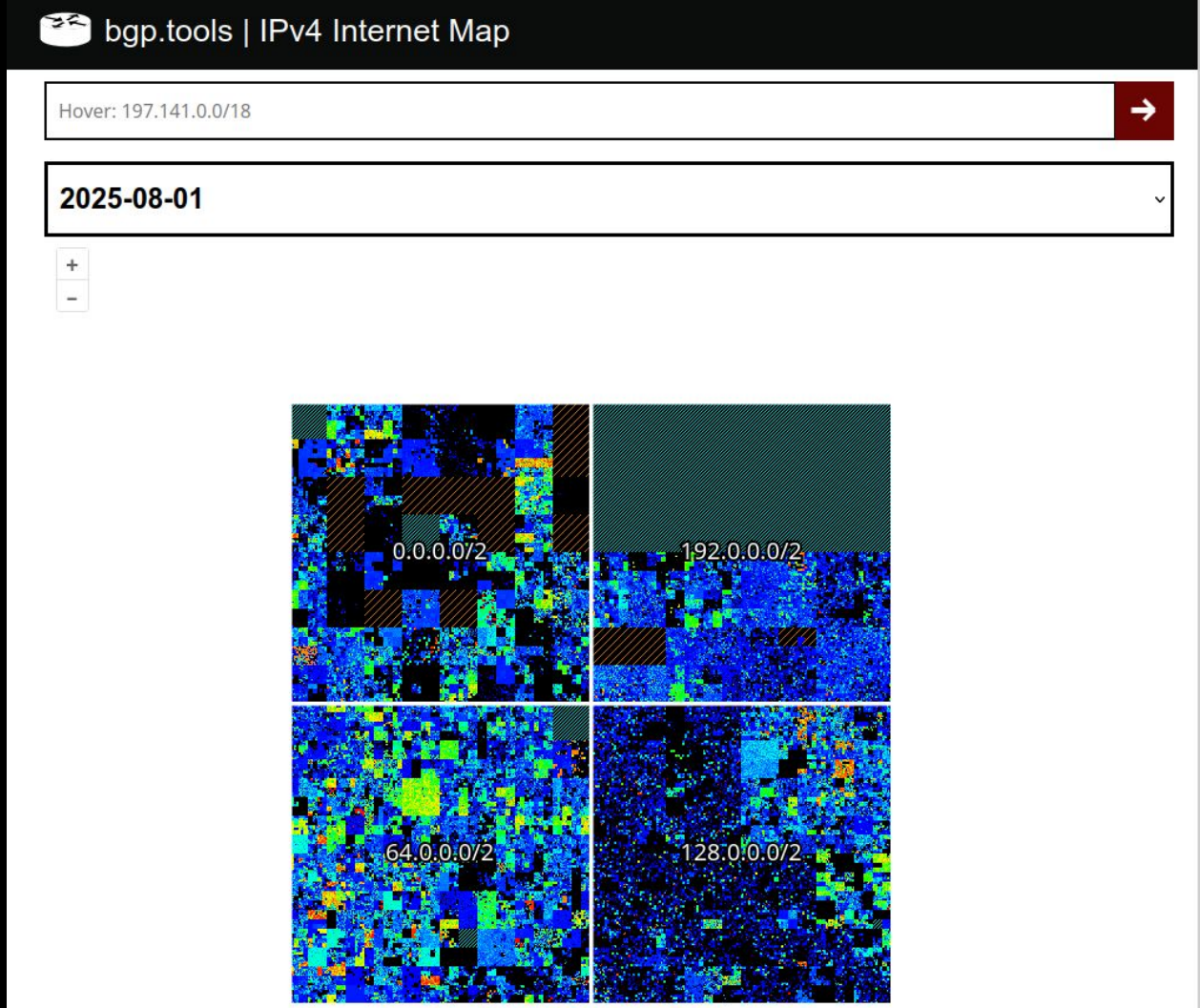
Thanks to Geoff Huston/APNIC!

a6fc0eaffbba21aa7c46c1394cfcd118

Optimisations

- To speed up things, we will only be pinging 1 IP out of each /24 (one that is known to reply)

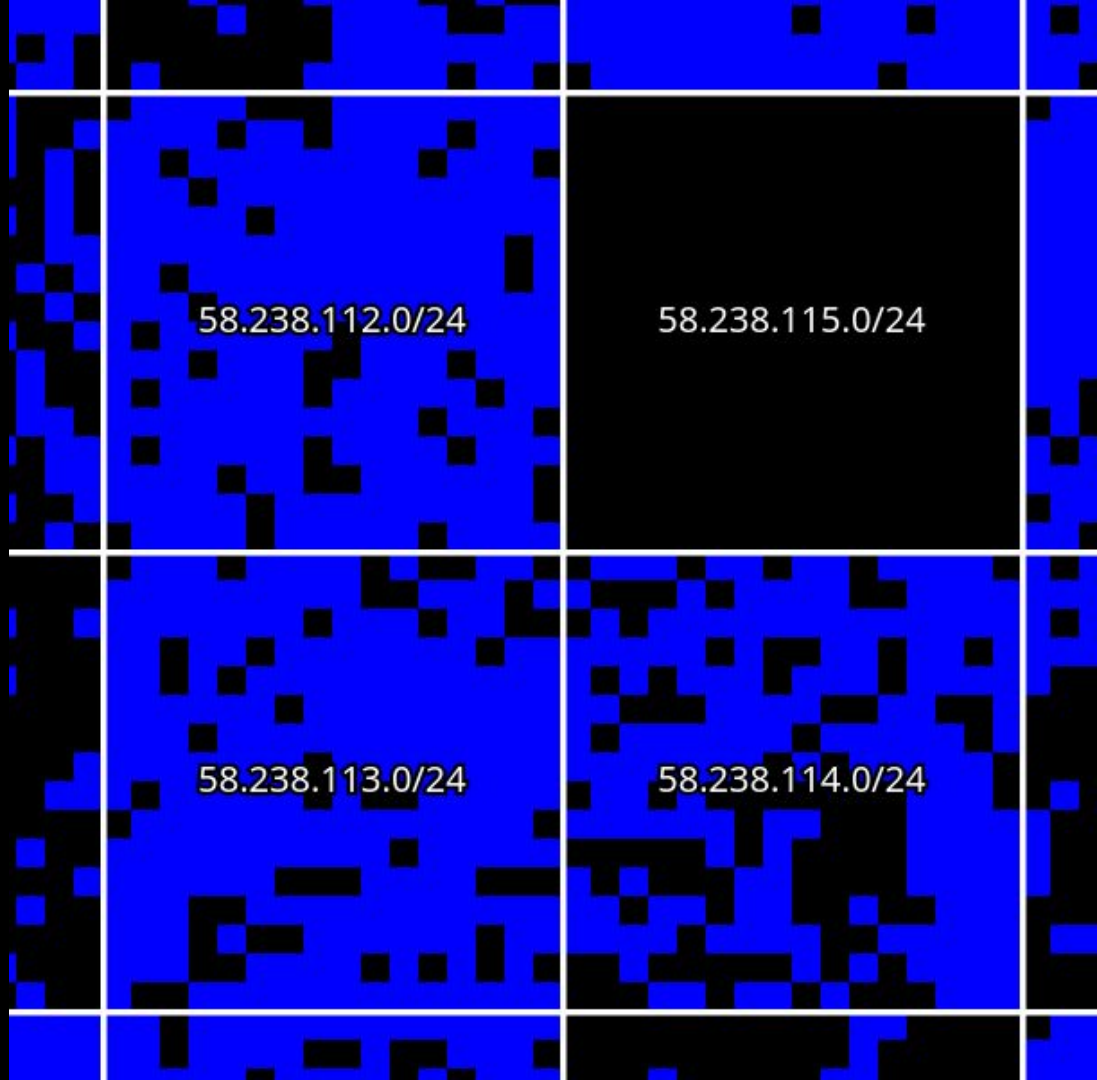
- Bgp.tools already knows what responds to pings because there is map.bgp.tools !



Getting left behind

- Because not everything responds to ICMP, some prefixes are just left behind in this test because there is no easy way to test them

- This is sadly common for CGNAT blocks



If you use all* RS-es, you get

- 980,966 prefixes can be tested
- 15% (144,526) of prefixes can reach you if you are IX RS only
- 85% (836,440) of prefixes can't use route servers paths
- Using 118 internet exchanges

a6fc0eaffbba21aa7c46c1394cfcd118

Most packets go to the following exchanges:

Internet Exchange	% of Prefixes	Prefix Replied
DE-CIX Frankfurt	15.64%	18190
BBIX Singapore	14.47%	16825
IX.br (PTT.br) São Paulo	9.74%	11324
LINX LON1	9.69%	11269
AMS-IX	6.39%	7430
NL-ix	5.34%	6207
BBIX Tokyo	4.57%	5318
BBIX US-West	3.71%	4319
DE-CIX New York	3.37%	3922
BIX.BG	3.24%	3772
DE-CIX Istanbul	2.98%	3461
Everyone Else (long tail)	47.81%	55592

a6fc0eaffbba21aa7c46c1394cfcd118

Localised surveys (EPF Hosts)

- 11.9% {117k} of prefixes can reach you
- Using LINX LON1, DE-CIX FRA, AMS-IX, NETNOD ARN
- Most packets go to the following exchanges:
  - DE-CIX Frankfurt (43%)
  - LINX LON1 (39%)
  - AMS-IX (17%)
  - Netnod Stockholm (5.4%)

a6fc0eaffbba21aa7c46c1394cfcd118

Thoughts

- The difference between the EPF host IXs and all 118~ exchanges is 27473 routes
  - Now those 27,473 routes may matter a lot!
  - But clearly a lot of this is *dominated* by a small number of networks with large downstream customer counts who also import route servers
- Many people export to route servers, few people import from them
  - 557,996 exported prefixes to RS
  - 144,526 prefixes from networks who import from RS
- That's a huge difference!
- Some of this is also limited by some prefixes (maybe the "juicy" CGNAT) ones being untestable

a6fc0eaffbba21aa7c46c1394cfcd118

Bonus Thoughts

- Because this system uses the same pipeline bgp.tools uses to detect "naughty packets" on exchanges, I also know what networks these route server replies came from
- The results surprised me! On the All IXPs run the top source networks were:
  - AS58453 / China Mobile International / LINX LON 1
  - AS18403 / FPT Telecom Company / BBIX Singapore
  - AS9002 / RETN Limited / BBIX Singapore
  - AS6939 / Hurricane Electric LLC / MegaIX Chicago
  - AS7713 / PT Telkom Indonesia Tbk / BBIX US-West
  - Etc, In total 4995 different routers send replies

a6fc0eaffbba21aa7c46c1394cfcd118

Thanks!

Questions? Comments? Stories?
Shy? Email epf@benjojo.co.uk
(or fedi/mastodon @benjojo@benjojo.co.uk)

a6fc0eaffbba21aa7c46c1394cfcd118