Getting the Most From Your INEX Membership

EOLAS Workshop - December 7th 2023

Barry O'Donovan
barry.odonovan@inex.ie

INTERCONNECTING NETWORKS AND PEOPLE FOR OVER 25 YEARS

Who am I?

- “Internet infrastructure specialist”
- INEX Operations team for ~16 years (contract)
- IXP Manager lead developer and project manager
- Island Bridge Networks - we work on the ISP side too

- More @ https://www.barryodonovan.com/



INEX

- Peering point for the island of Ireland, member owned association, cost recovery, founded 1996
- ~115 members (inc. >95% of eyeballs)
- > 1Tbps of IP data exchanged at peek
- ~7 Tbps of connected edge capacity
- Dual infrastructure, 7 PoPs, own dark fibre
- Opened INEX Cork in 2016
- IXP Manager / Salt / Napalm automation
- Recommended for ISO27001 certification

What we are going to talk about today:

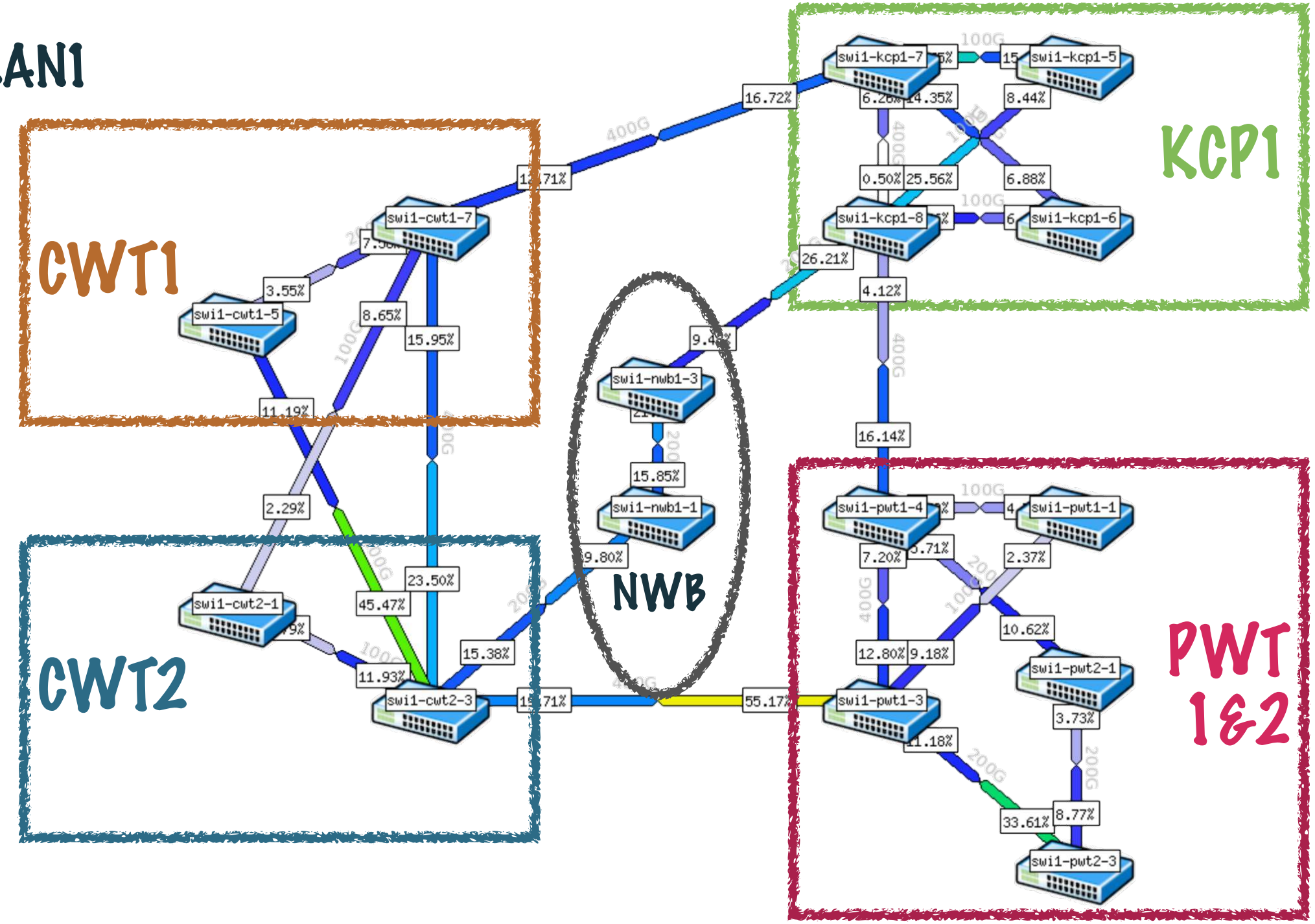
- An “internet primer” - where do IXPs fit in
- INEX topology and technology
- Route servers and route collectors
- High level overview of BGP prefix propagation
- Traffic engineering options
- Who you need bilateral BGP sessions with ***
- Essential internet services available over INEX
- Tools available at INEX via IXP Manager
- Other tools of note


INEX Topology

Topology Details

- Switch config fully automated for Arista and Mellanox including:
  - Member port configs
  - Core port configs
  - VXLAN + BGP
  - Base configs
- INEX has own dark fibre between PoPs and:
  - Spec and install appropriate MUXes
  - Use Coriant's Groove G30 platform *(now Infinera)*
  - Sometimes use coloured SFPs (mgmt only now), BiDis on campus cross connects.
- Fully automated monitoring and alerting systems.
- Route servers, route collectors, AS112, IXP Manager, mgmt network, …

INEX LAN1


CWT1
CWT2
NWB
KCP1
PWT 1&2
swi1-cwt1-7
swi1-cwt1-5
swi1-cwt2-1
swi1-cwt2-3
swi1-nwb1-3
swi1-nwb1-1
swi1-kcp1-7
swi1-kcp1-5
swi1-kcp1-8
swi1-kcp1-6
swi1-pwt1-4
swi1-pwt1-1
swi1-pwt1-3
swi1-pwt2-1
swi1-pwt2-3
16.72%
3.55%
8.65%
15.95%
11.19%
2.29%
23.50%
45.47%
15.38%
11.93%
9.80%
15.85%
55.17%
0.50%
25.56%
6.88%
8.44%
26.21%
4.12%
16.14%
7.20%
2.37%
10.62%
12.80%
9.18%
3.73%
33.61%
8.77%
400G
200G
100G

INEX LAN1



Arista DCS-7280SR-48C6

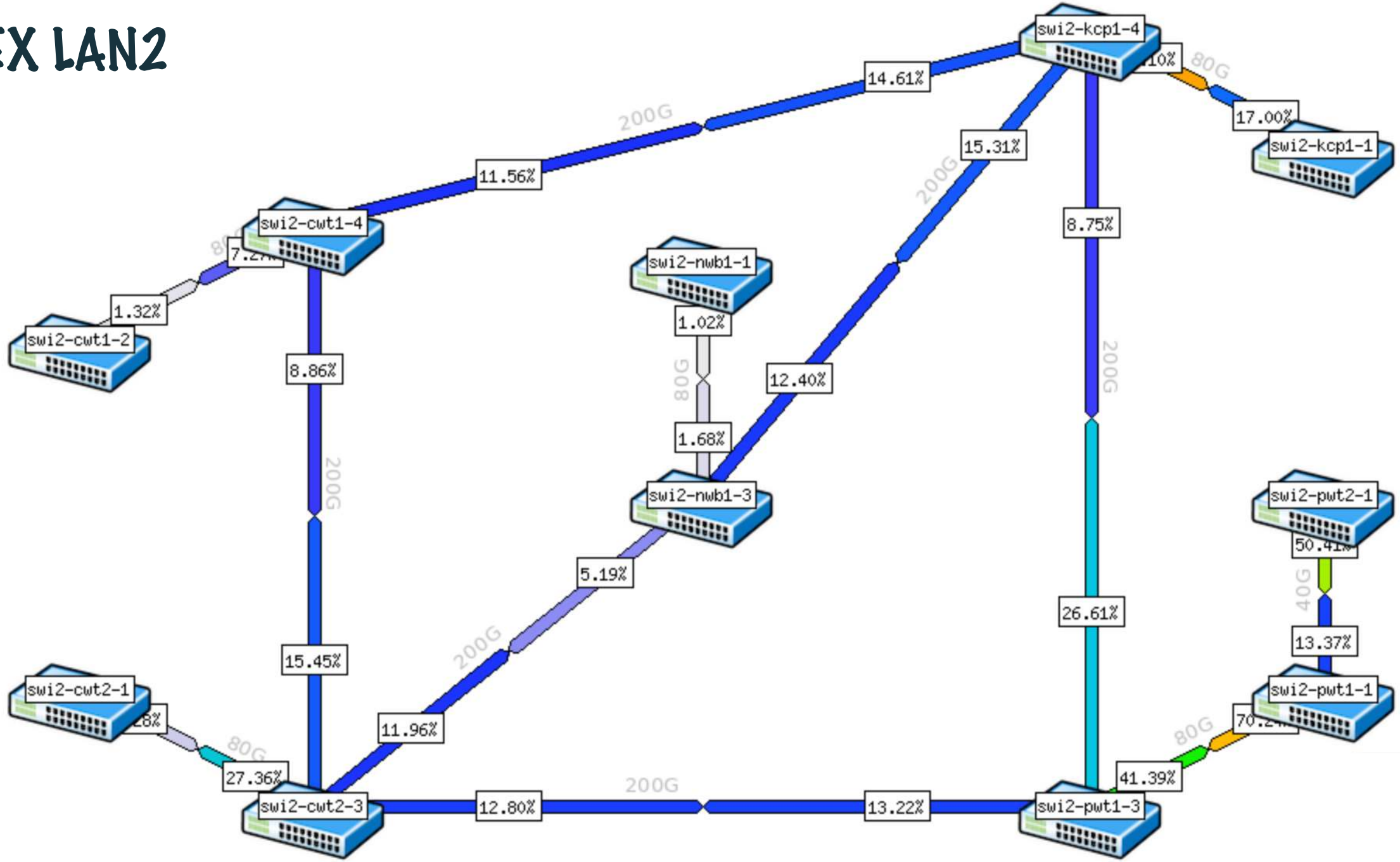


Arista DCS-7060CX2-32S



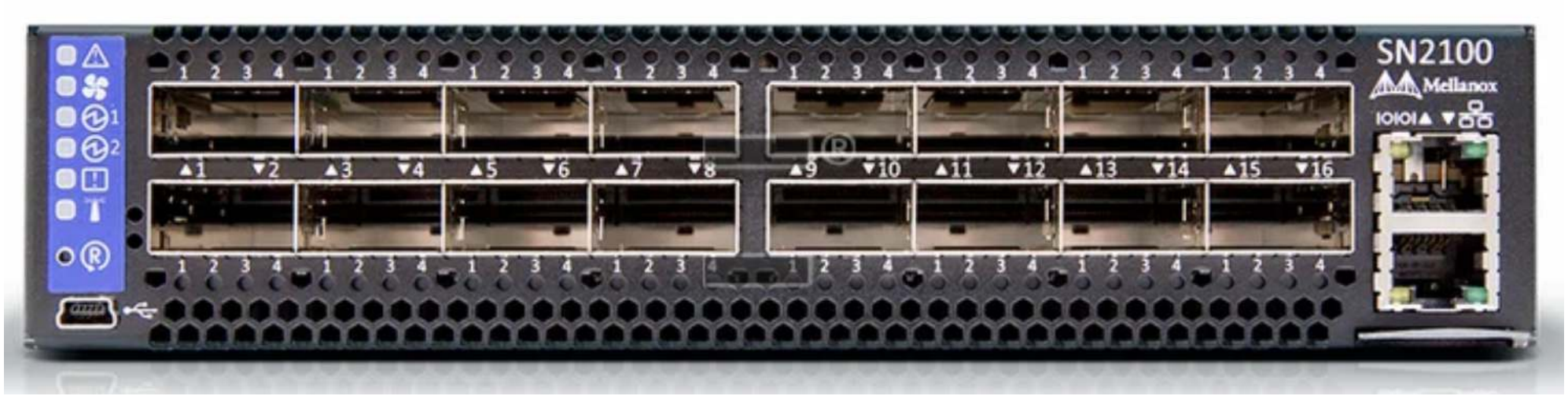
Arista DCS-7280CR3

INEX LAN2

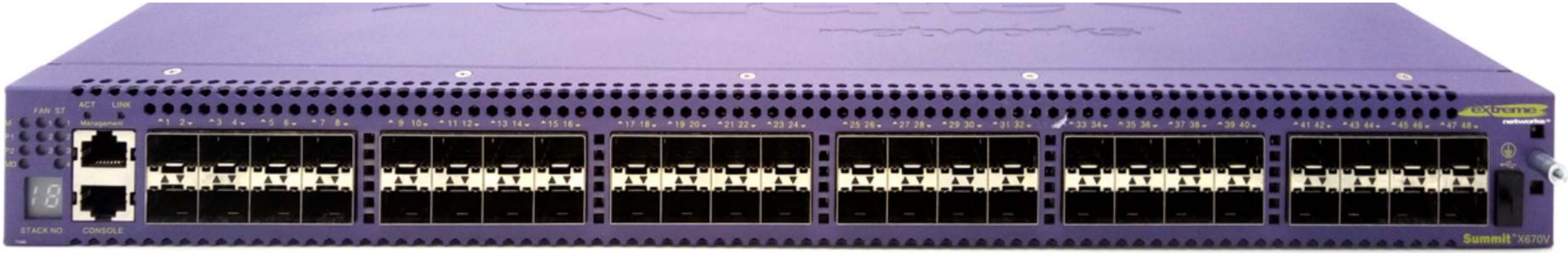


swi2-kcp1-4
14.61%
200G
11.56%
swi2-cwt1-4
1.32%
swi2-cwt1-2
8.86%
200G
15.45%
swi2-cwt2-1
27.36%
swi2-cwt2-3
12.80%
200G
13.22%
swi2-pwt1-3
41.39%
80G
swi2-pwt1-1
13.37%
40G
swi2-pwt2-1
26.61%
200G
8.75%
swi2-kcp1-1
17.00%
80G
15.31%
200G
swi2-nwb1-1
1.02%
80G
1.68%
12.40%
swi2-nwb1-3
5.19%
200G
11.96%

INEX LAN2

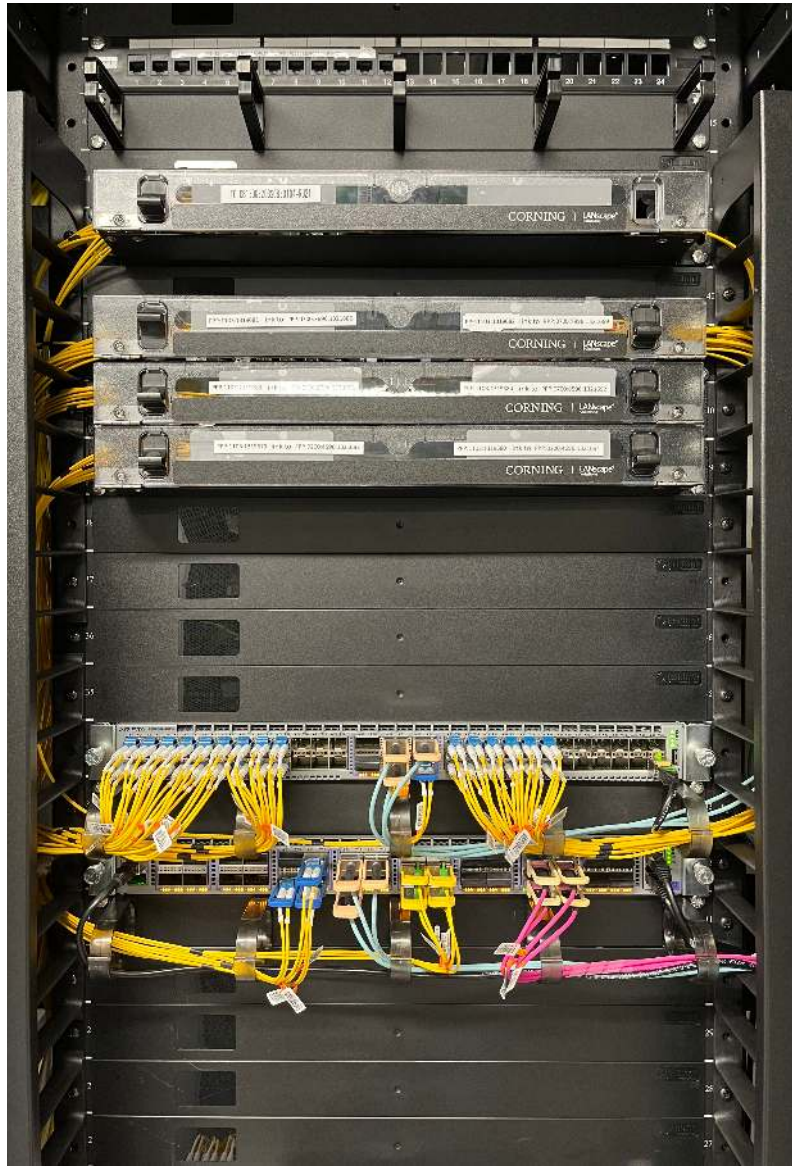


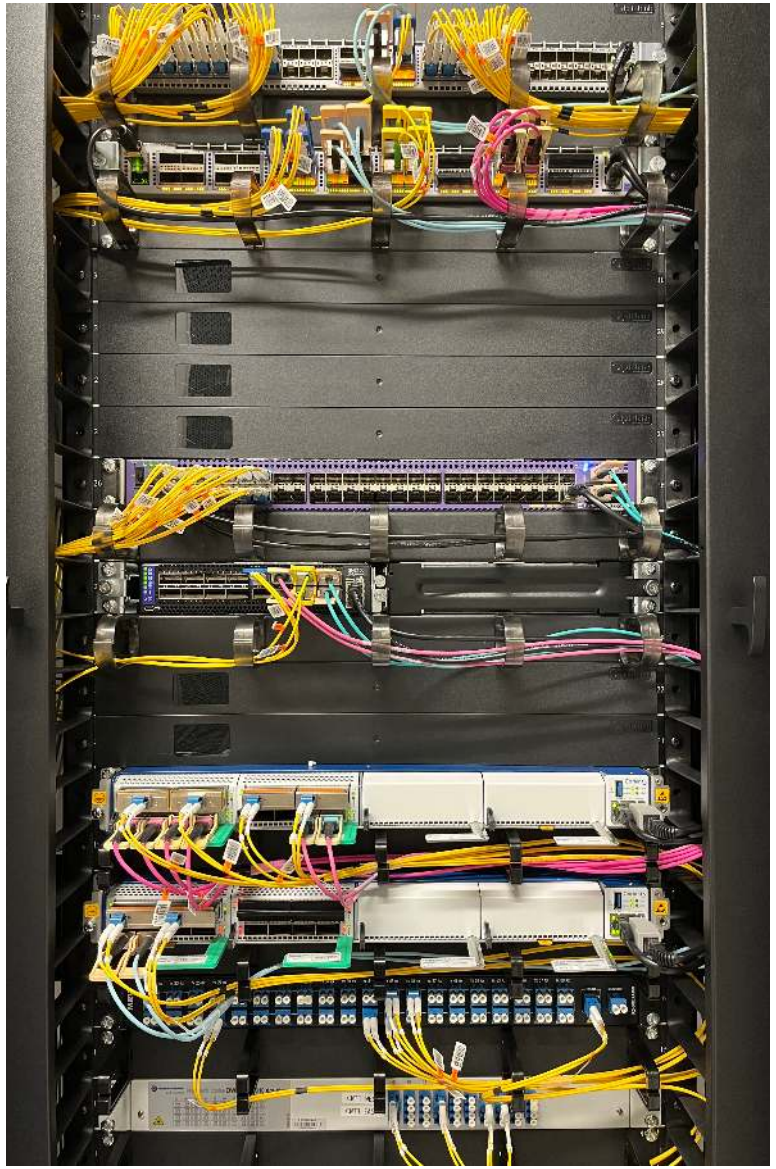
Mellanox SN2100 (now Nvidia)

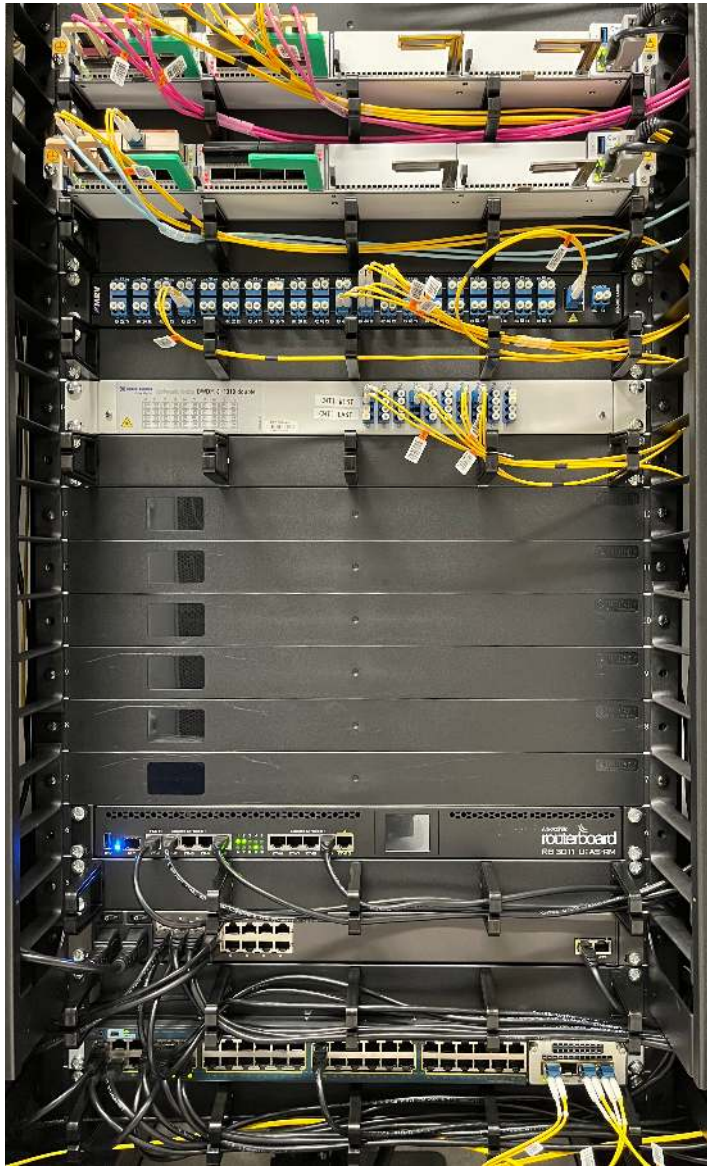


Extreme X670-G2

Equinix DB1

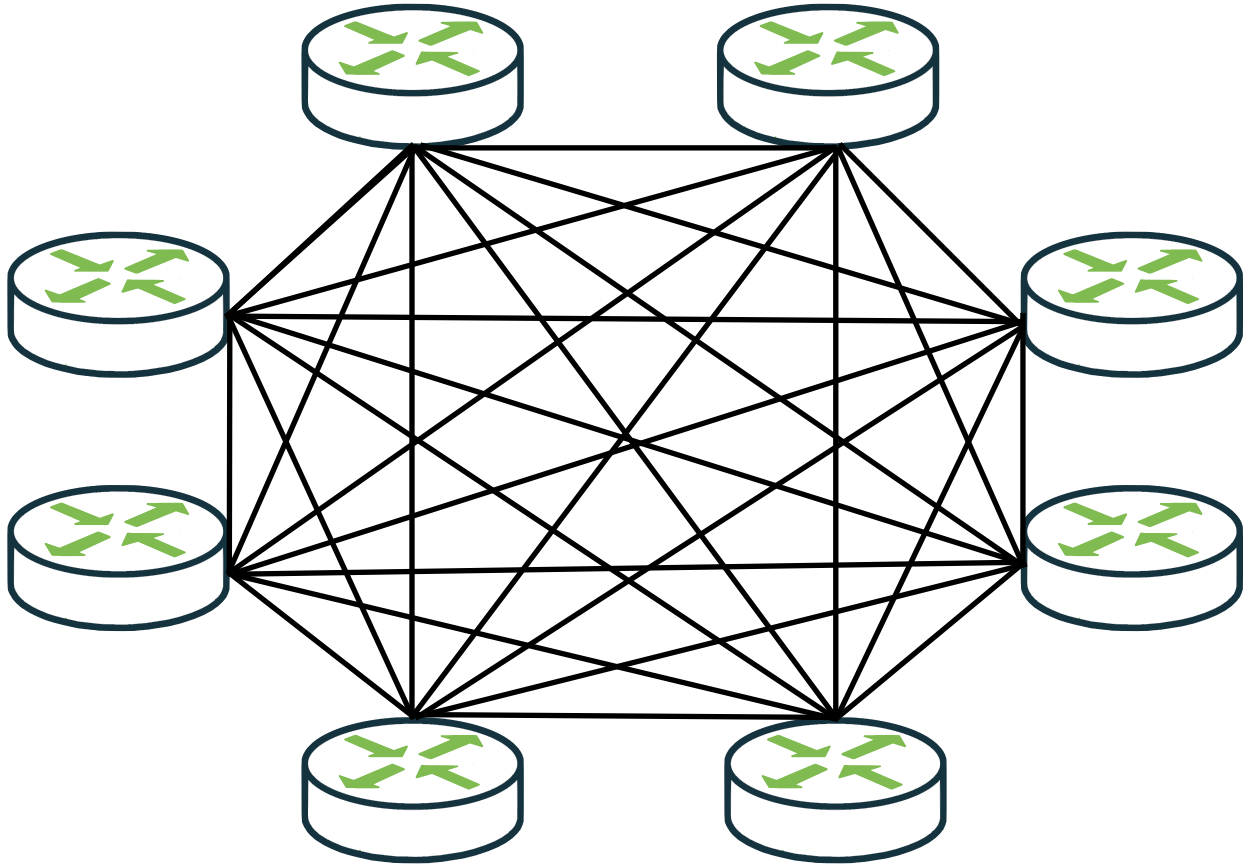






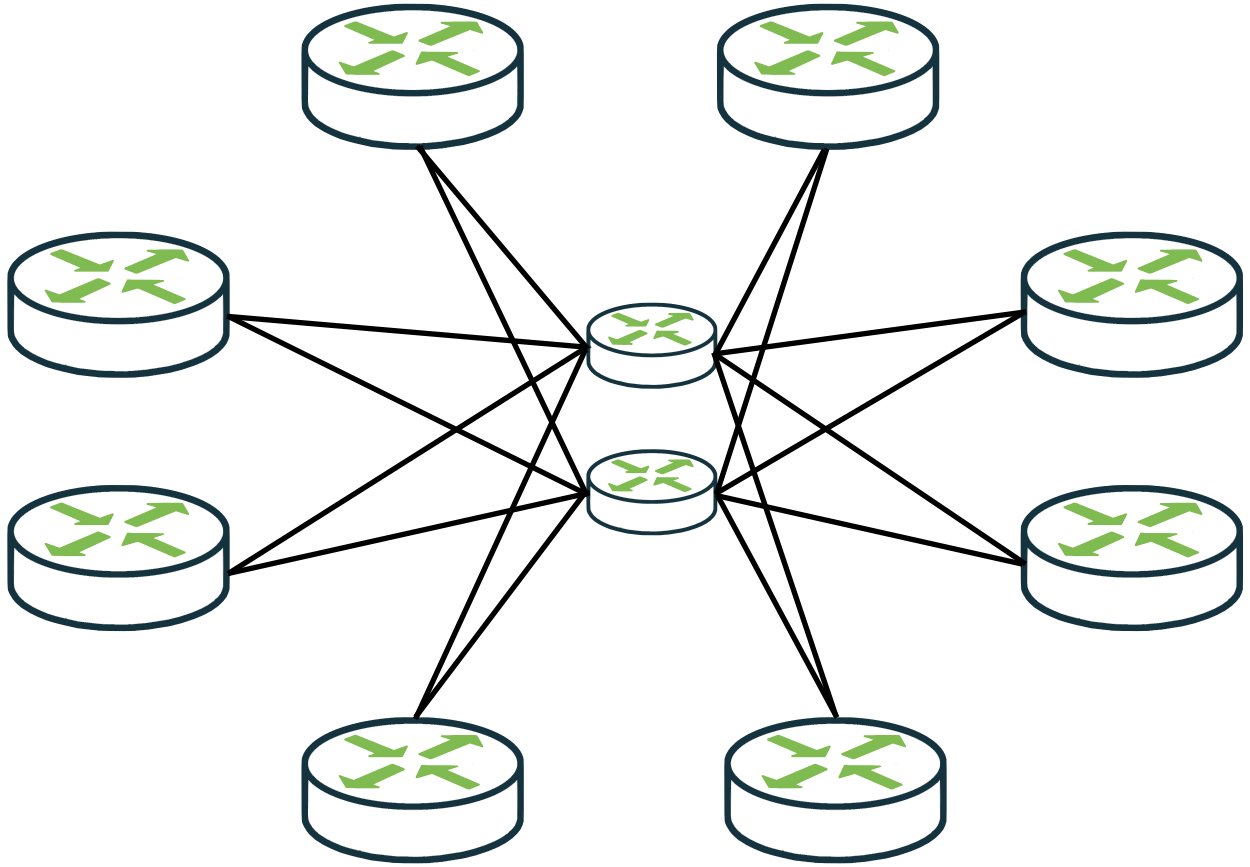
Route Servers
And Route Collectors

An IXP Without Route Servers



- $\frac{n(n-1)}{2}$ bilateral sessions
- 8 members - 28 sessions
- 100 members - 4,950 sessions

An IXP Without Route Servers



- RFC7947 - *multilateral interconnection using a third-party brokering system.*
- Unlike route reflectors, route servers operate with EBGP.
- Attribute and AS path transparency - ‘acts as if it’s doesn’t exist’.
- Critical infrastructure at an IXP.

Advantages of Route Servers



Immediate benefit of / for new members



Security and assurance for routes exchanged



Avoids email tennis, typos, mis-understandings etc.



No production changes, simpler configurations

At INEX: LAN1 - 86%. LAN2 - 90%. Cork - 100%.

Route Collectors

- Very similar configuration to route servers
  - Except: exports zero prefixes - learns only.
- A mandatory peering session at INEX and used for:
  - Onboarding and quarantine
  - Monitoring
  - Looking glass (inc. members)
  - Diagnostics
- Also a useful testbed for route servers (no consequences as not used for routing; all members; bigger routing tables)

Getting the Most From INEX?

What does ‘getting the most’ look like?

- The obvious: uncongested low latency exchange of traffic within the island at a fixed value-for-money price.

- Exchanging as much traffic as possible.
- Reliably exchanging valuable traffic.
- Resilience.
- Access to essential internet services.
- Part of a technical community.
- EOLAS.

NB: make sure your network is registered on PeeringDB!

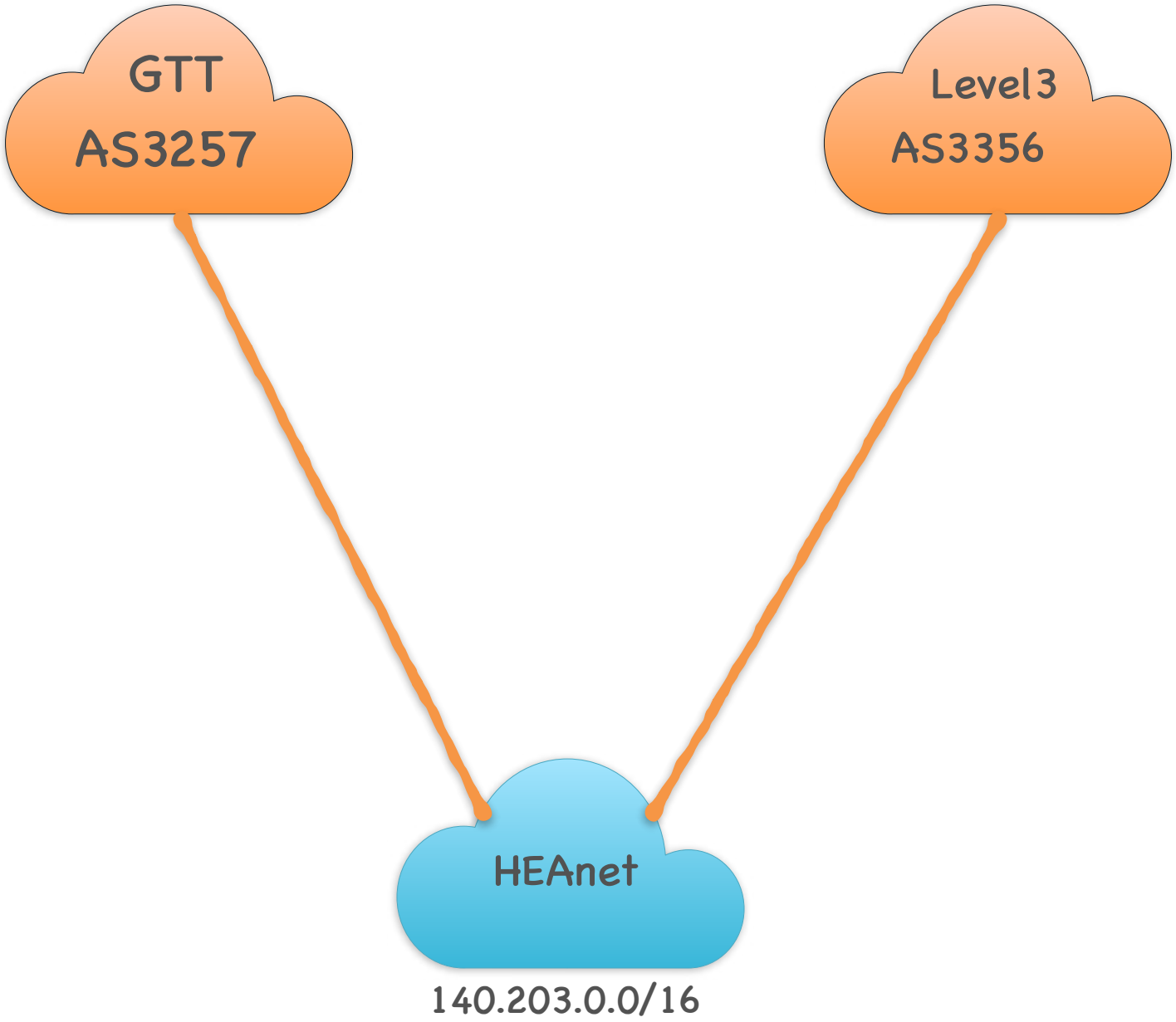


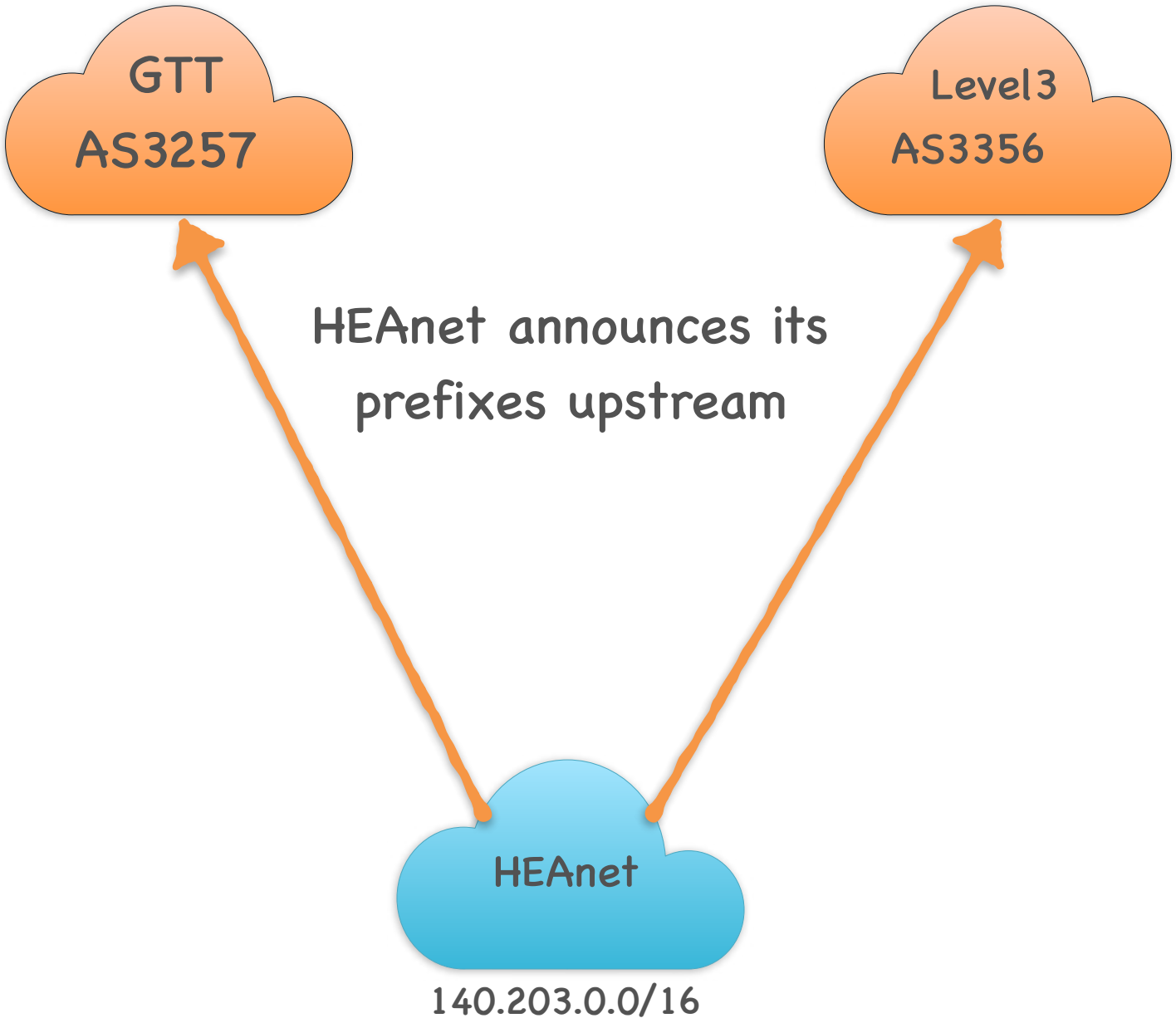
BGP

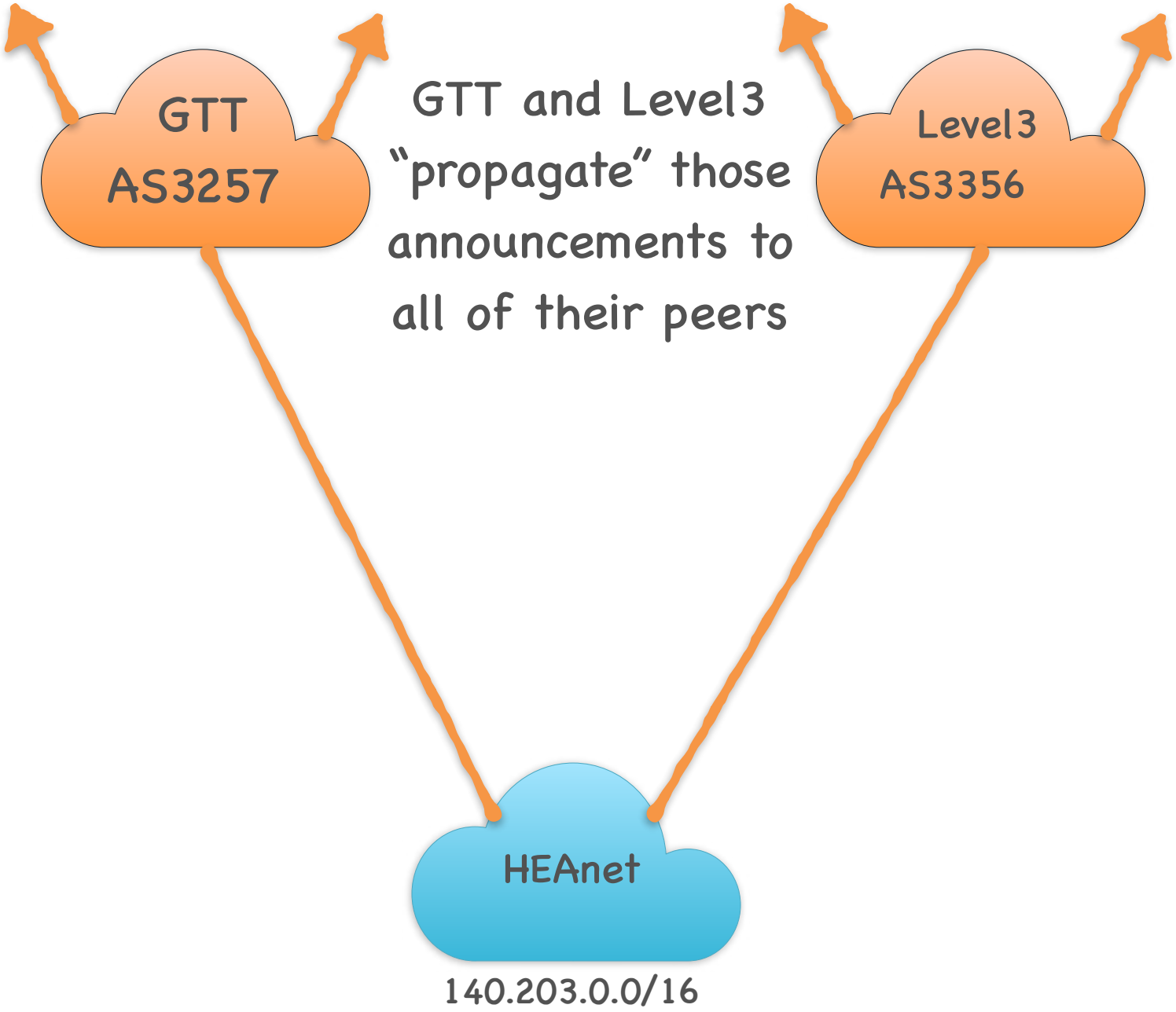
Prefix Propagation

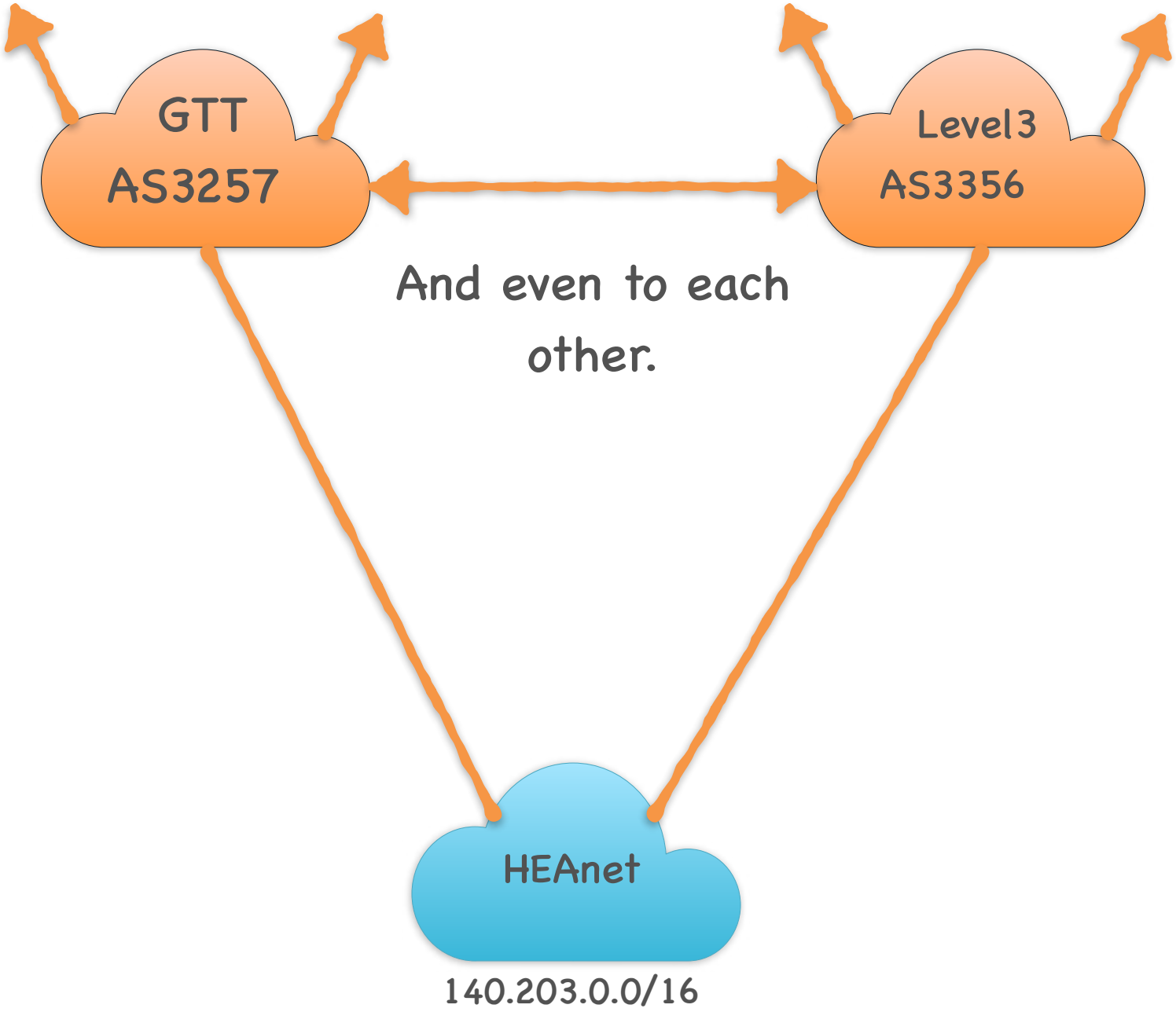
How does every other network on the internet learn what prefixes my network (AS) has and how to get to my network?

INEX


GTT
AS3257
Level3
AS3356
HEAnet
140.203.0.0/16

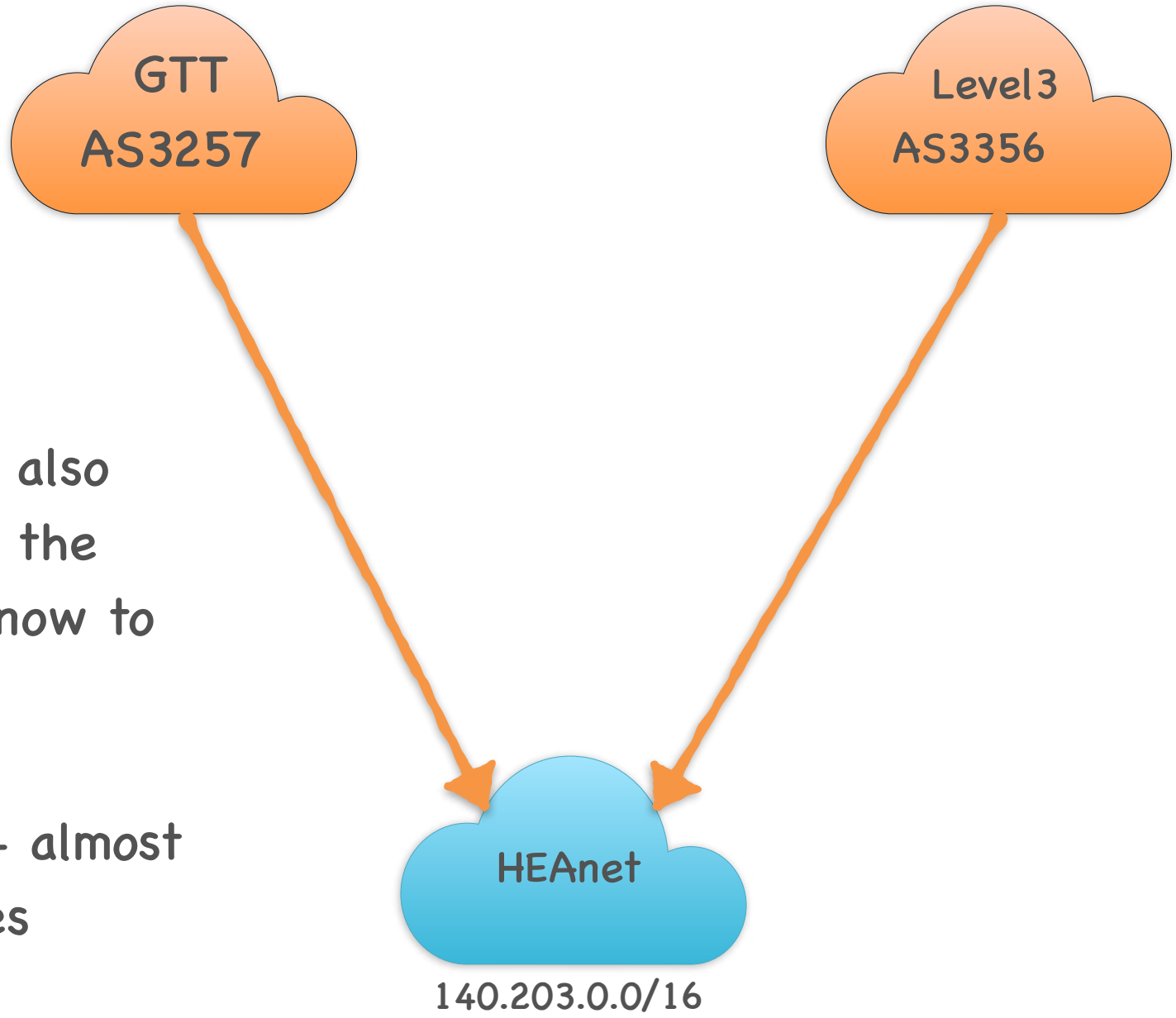

GTT
AS3257
Level3
AS3356
HEAnet announces its
prefixes upstream
HEAnet
140.203.0.0/16

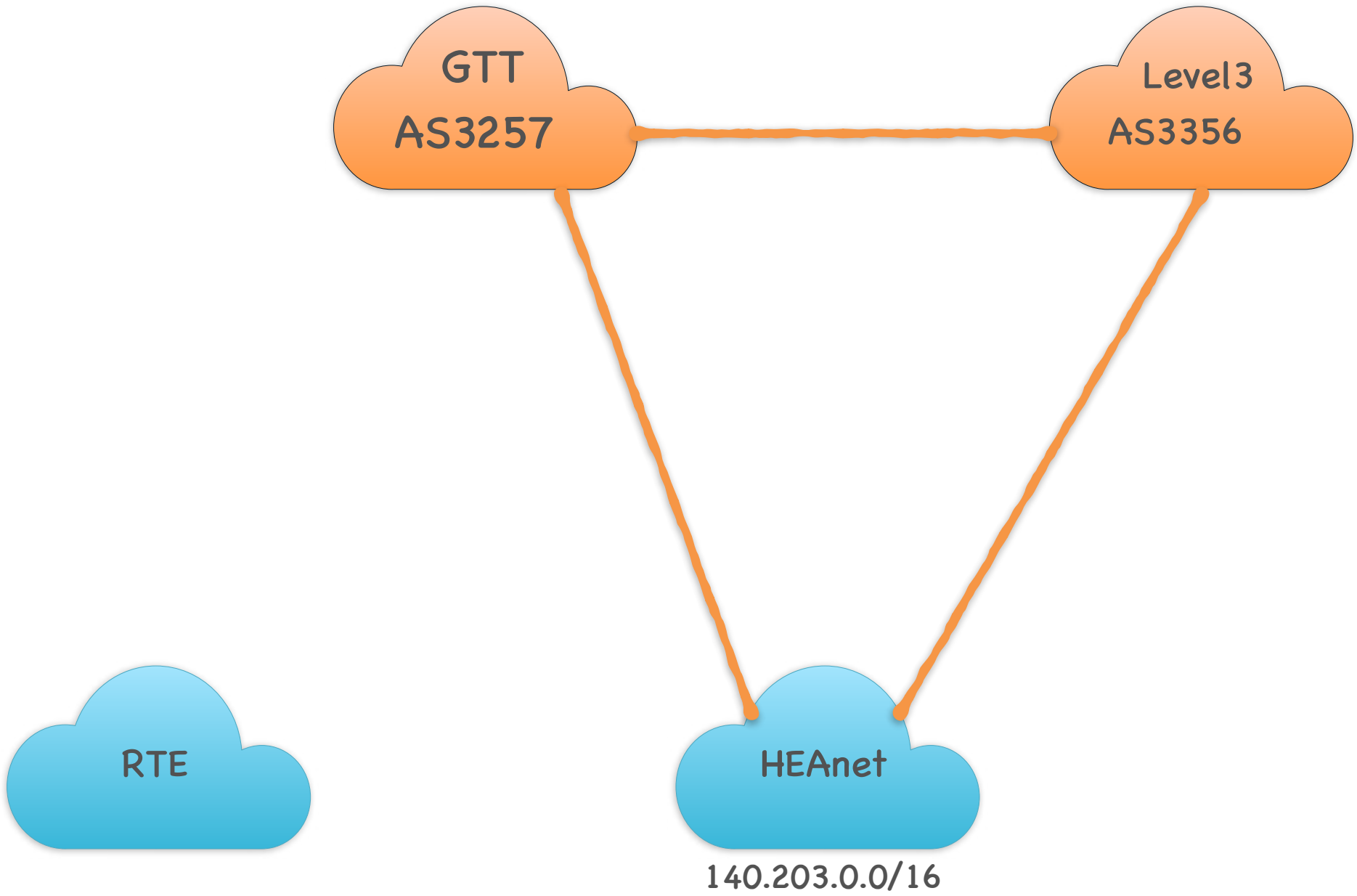

GTT
AS3257
GTT and Level3 "propagate" those announcements to all of their peers
Level3
AS3356
HEAnet
140.203.0.0/16

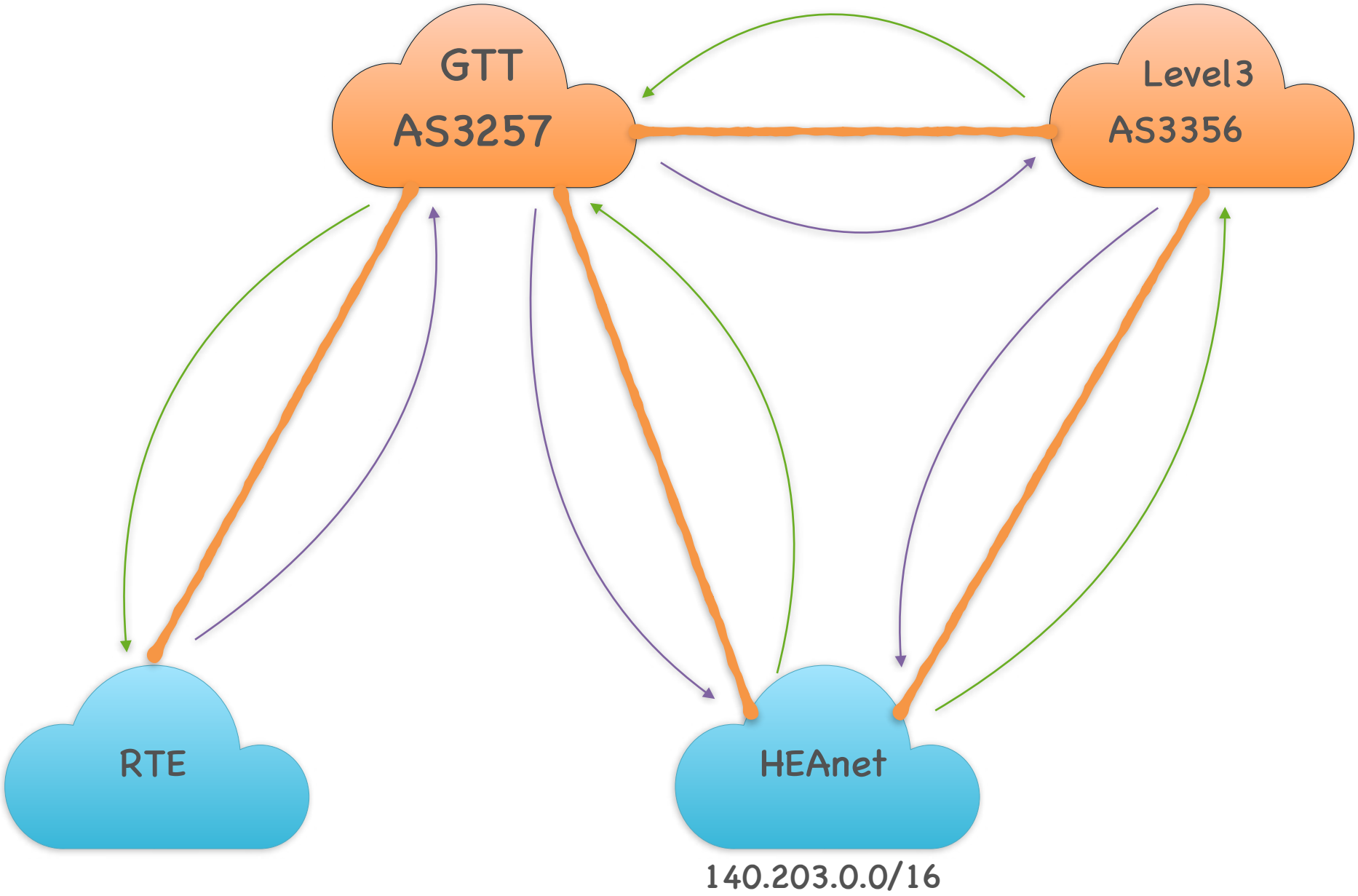

GTT
AS3257
Level3
AS3356
And even to each other.
HEAnet
140.203.0.0/16

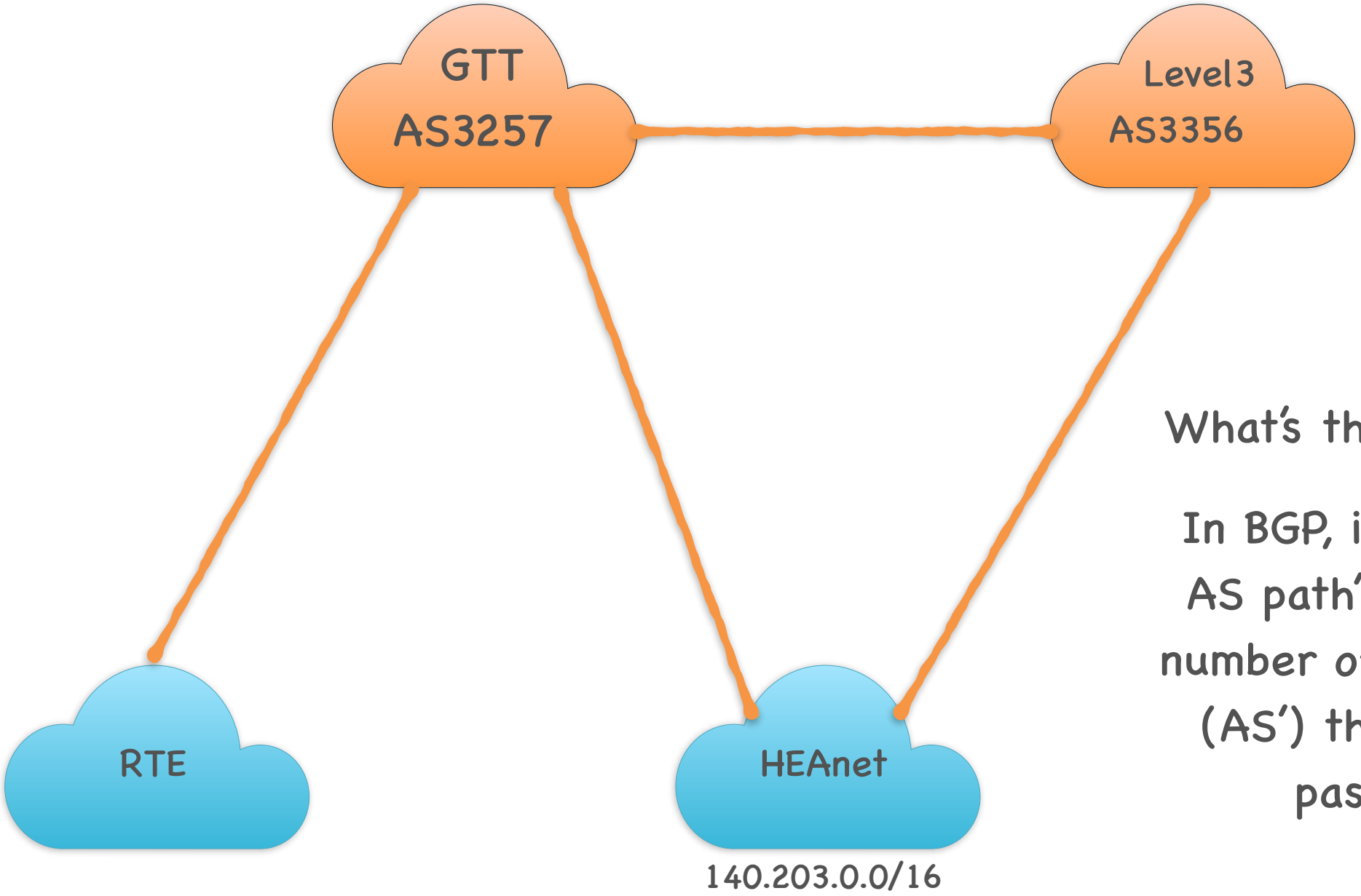
GTT & Level3 also propagate all the prefixes they know to HEAnet.

This is the DFZ - almost 1m prefixes



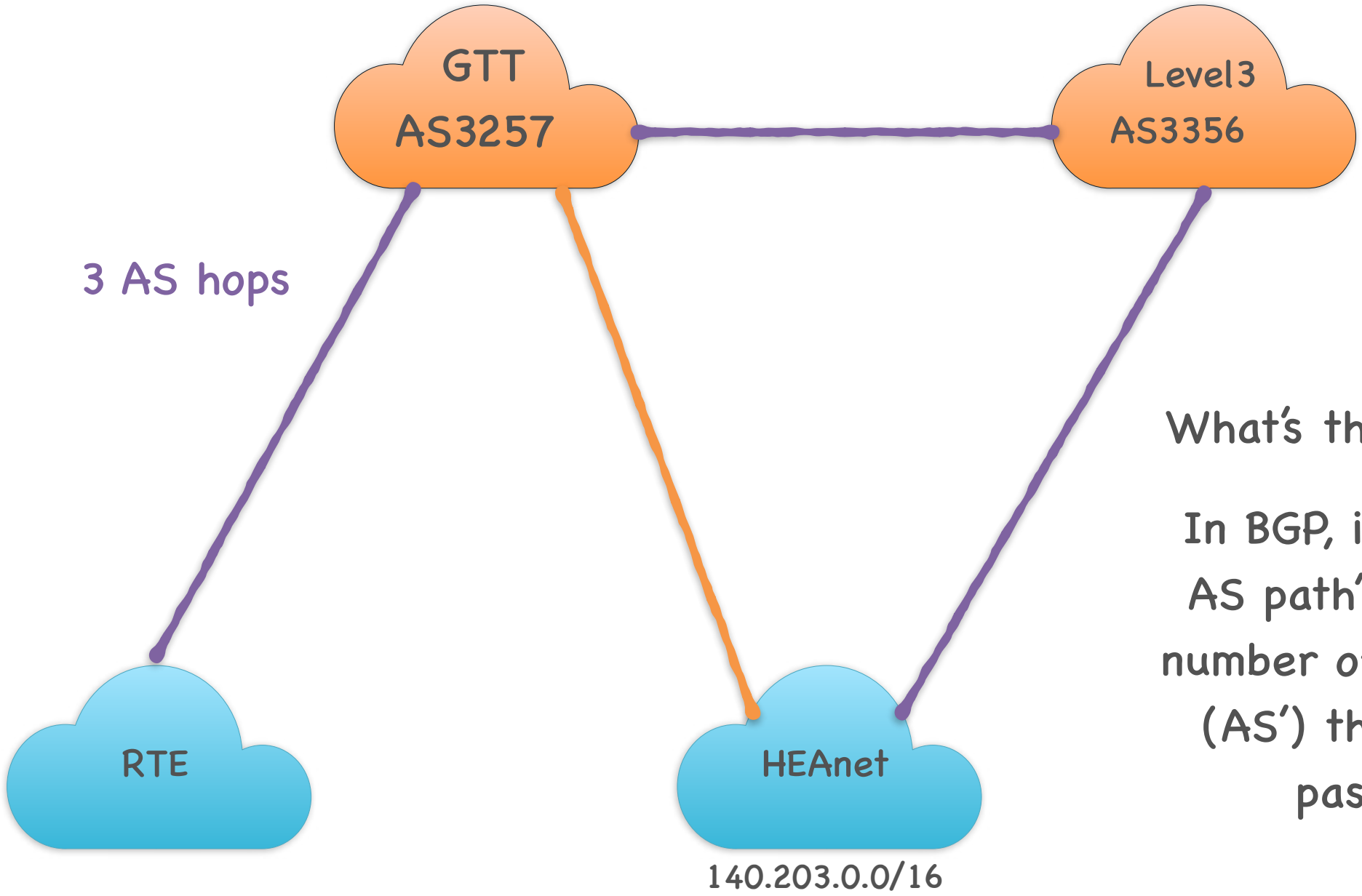

GTT
AS3257
Level3
AS3356
RTE
HEAnet
140.203.0.0/16


GTT
AS3257
Level3
AS3356
RTE
HEAnet
140.203.0.0/16



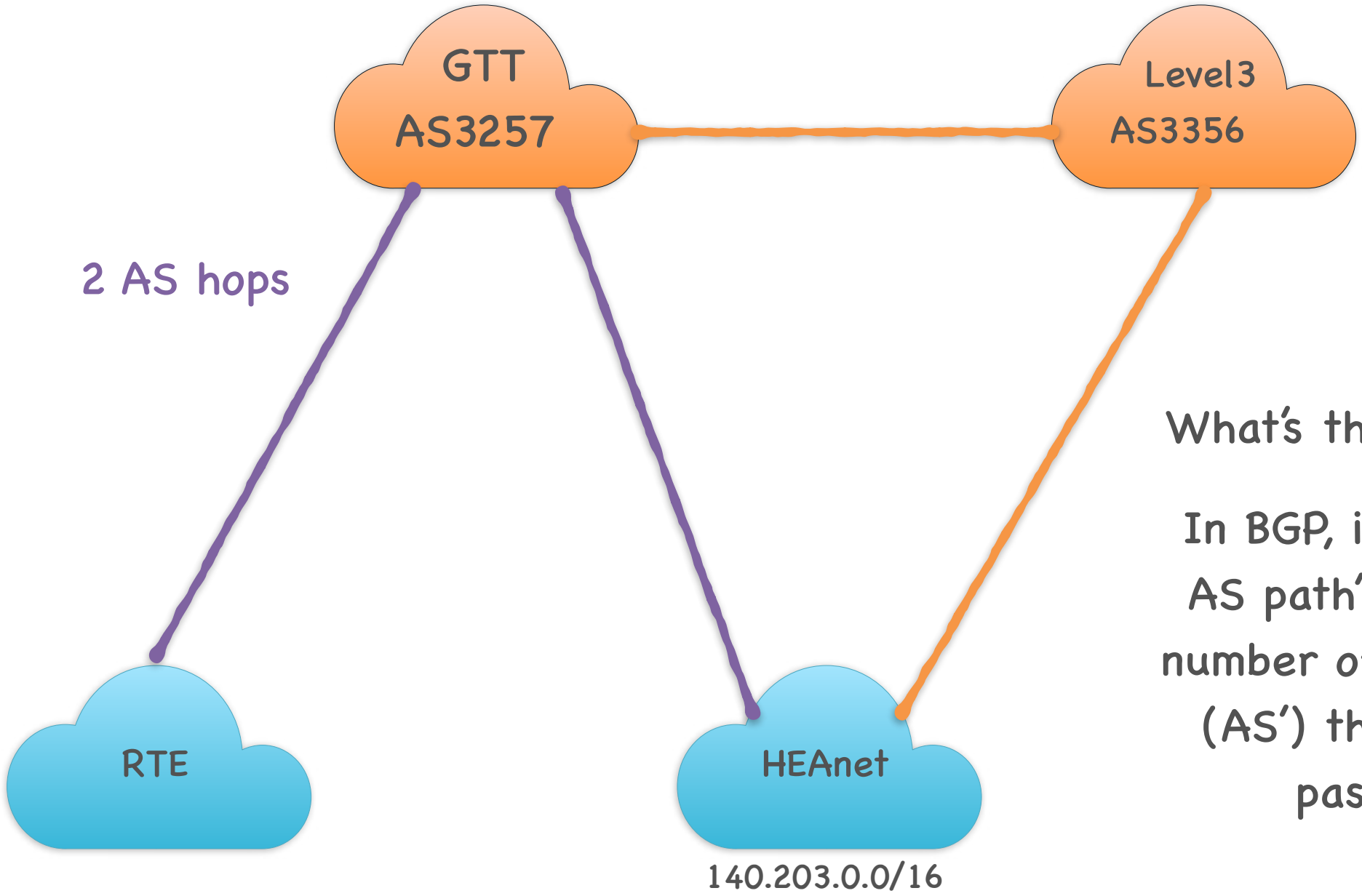
What's the quickest path?

In BGP, it's the "shortest AS path" - i.e. the least number of other networks (AS') that you have to pass through.



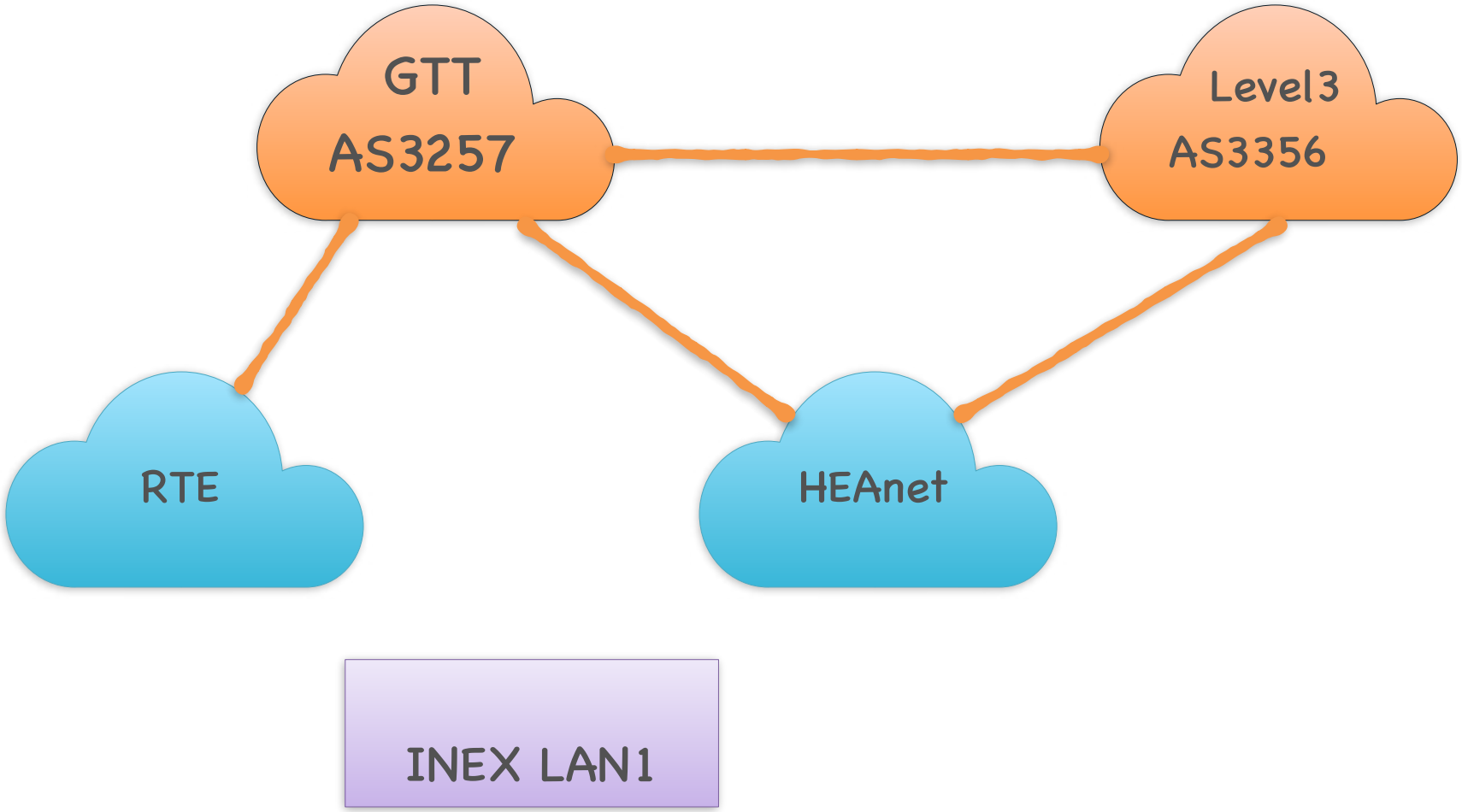
What's the quickest path?

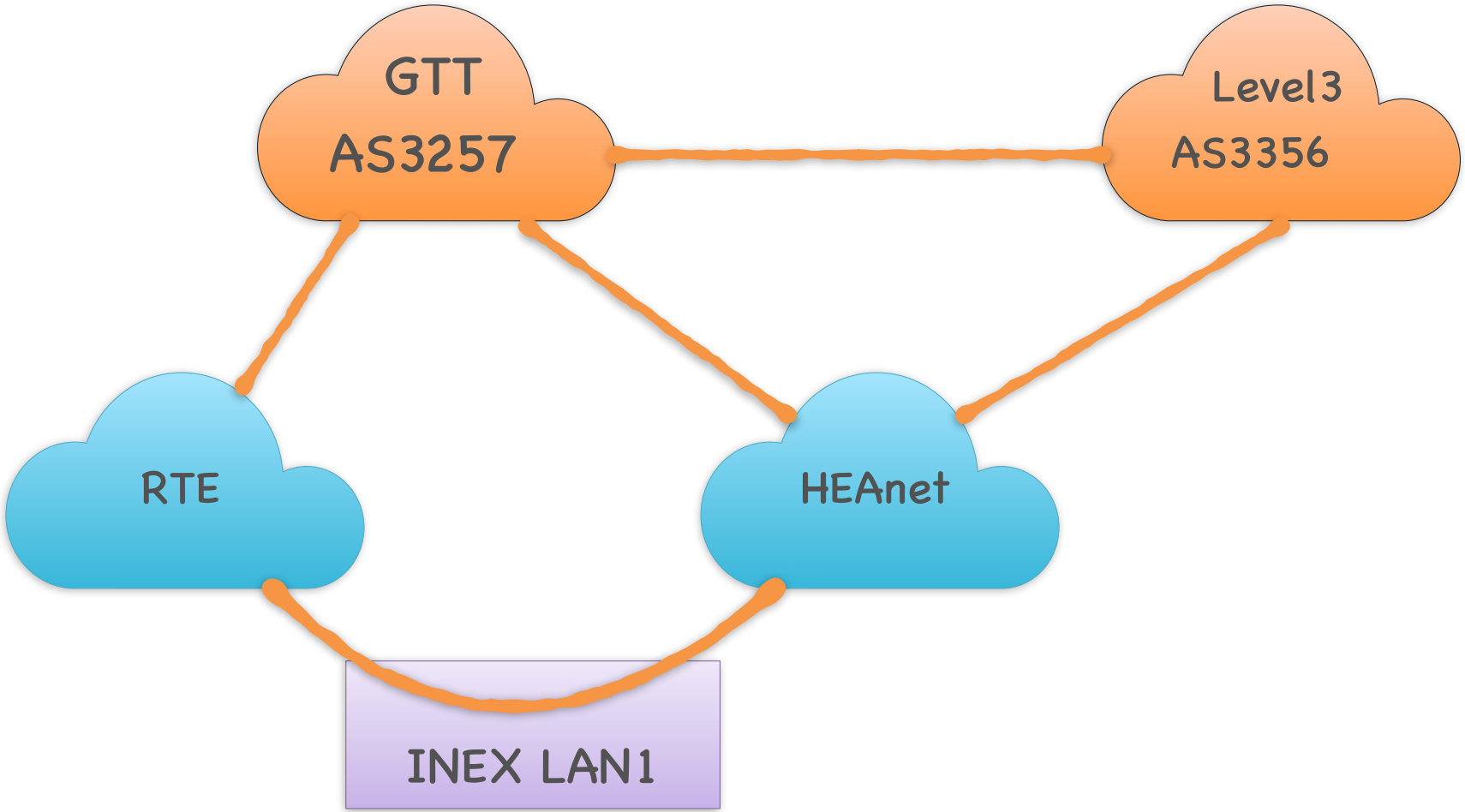
In BGP, it's the "shortest AS path" - i.e. the least number of other networks (AS') that you have to pass through.

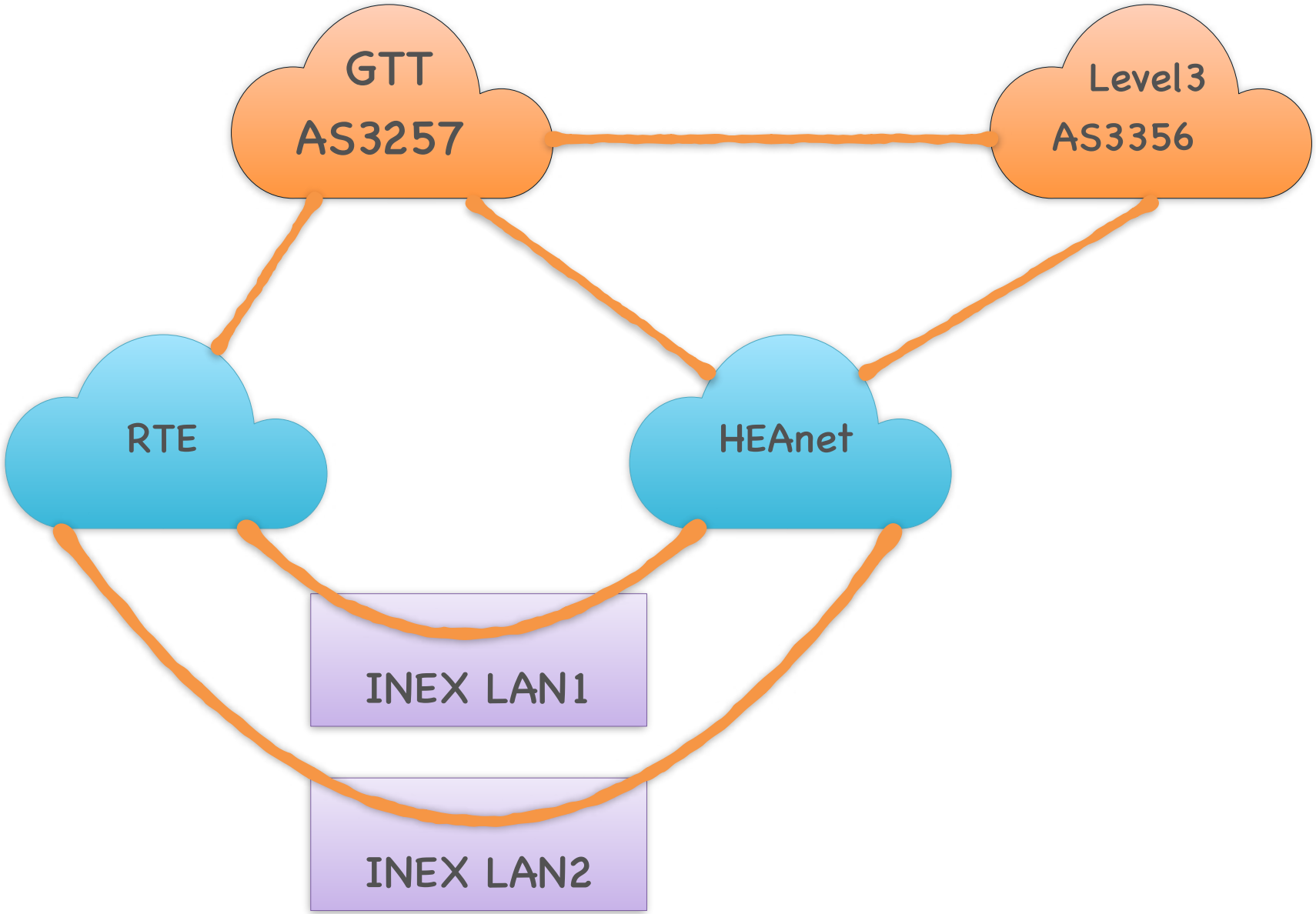


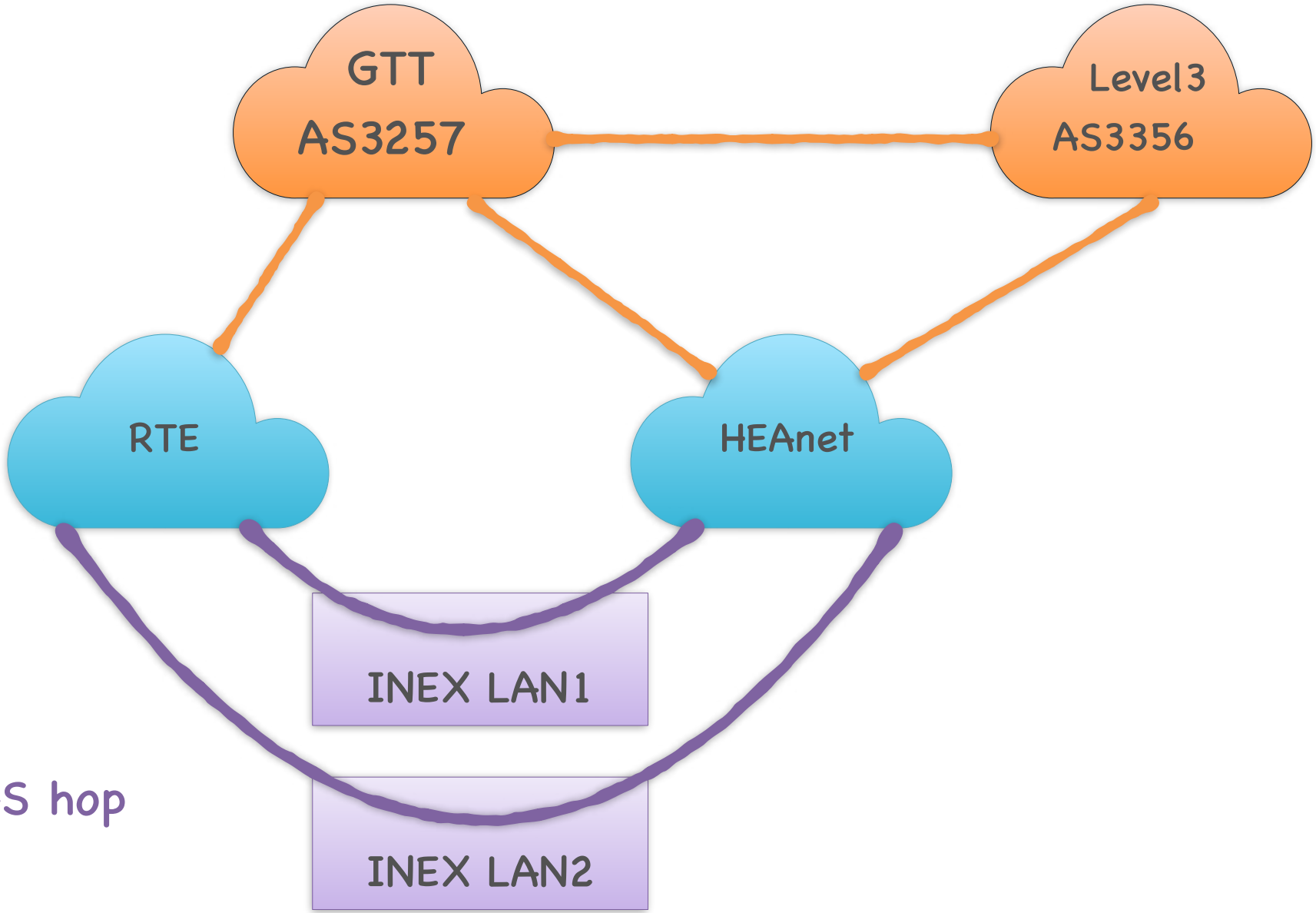
What's the quickest path?

In BGP, it's the "shortest AS path" - i.e. the least number of other networks (AS') that you have to pass through.


GTT
AS3257
Level3
AS3356
RTE
HEAnet
INEX LAN1


GTT
AS3257
Level3
AS3356
RTE
HEAnet
INEX LAN1


GTT
AS3257
Level3
AS3356
RTE
HEAnet
INEX LAN1
INEX LAN2



1 AS hop

BGP

Best Path Selection Algorithm (on prefix length)

- Prefer the path with the highest LOCAL_PREF (def: 100)
- Prefer the path with the shortest AS_PATH
- Prefer the path with the lowest MED
- Prefer the oldest path

Tie-breakers:

- Prefer the path from the router with the lower router-id
- Prefer the path that comes from the lowest neighbor address

(vendor specific and more esoteric decisions omitted)

INEX

BGP

Best Path Selection Algorithm (on prefix length)

- **Prefer the path with the highest LOCAL_PREF (def: 100)**
- **Prefer the path with the shortest AS_PATH**
- **Prefer the path with the lowest MED**
- Prefer the oldest path

Tie-breakers:

- Prefer the path from the router with the lower router-id
- Prefer the path that comes from the lowest neighbor address

Typical default decision. What you can effect.

INEX

BGP

Can we use BGP to exchange more traffic over INEX?

- **Prefer the path with the highest LOCAL_PREF (def: 100)**
  - Effects outbound traffic. Can lead to asymmetric routing if you don't also seek to effect the inbound traffic.
- **Prefer the path with the shortest AS_PATH**
  - Prepend your announcements to transit providers to help ensure INEX also looks closer.

Should you do this?



It's really traffic behind INEX members that you're affecting here…

(Yes, I purposely didn't mention announcing more specifics)

INEX

More traffic via bilateral peering sessions

- Not all members use the route servers.
  - **Caution -** some may peer with them but not actually use them!

You should peer directly (bilateral) with:

- Amazon AWS - peering-emea@amazon.com
  - https://aws.amazon.com/peering/policy/ and PeeringDB

- Apple - peering@group.apple.com (usually efficient++)

More traffic via bilateral peering sessions

- Edgio (Limelight) - *peering@llnw.com* (Jack and Ben)
  - https://www.edg.io/peering/ - 1Gb minimum

- Google - will be switching to bilateral only.
  - Currently more traffic available on bilats.
  - https://peering.google.com/#/options/peering

- Microsoft - https://www.microsoft.com/peering
  - *Must be requested and managed via Azure but works well.*
  - Are on route servers but do not use them for routing.

Tips for bilateral sessions

- You’re unlikely to have the resources to automate dynamic prefix filtering of bilateral BGP sessions like INEX does on the route servers.
- Use sensible import and export policies.

Export:

- Easy: just export your own prefixes and-nothing-else

Tips for bilateral sessions

Import:

- Reject bogons (*)
- Reject small prefixes (</24 ipv4 / </48 ipv6)
- Reject default route
- Reject your own prefixes
- ***Use sensible max-prefix setting!***

** for ipv6 accept only 2000::/3 prefix-length-range /16-/48*

Other networks not on the route servers

- BT AS5400 (S)
- C&W (S)
- Convergence Group (O)
- GEANT (S)
- NORDUnet (S)
- Servecentric (O)
- ServiceNow (O)
- Verizon Business (S)
- Virgin Media (S)
- Zayo (S)

O = open peering policy; S = selective peering policy

Biggest sources of traffic?

- Akamai
- Amazon
- Apple
- Cloudflare
- Edgio
- Fastly
- Google
- Meta
- Microsoft
- Netflix
- RTE
- Yahoo!

In alphabetical order.

Resilience

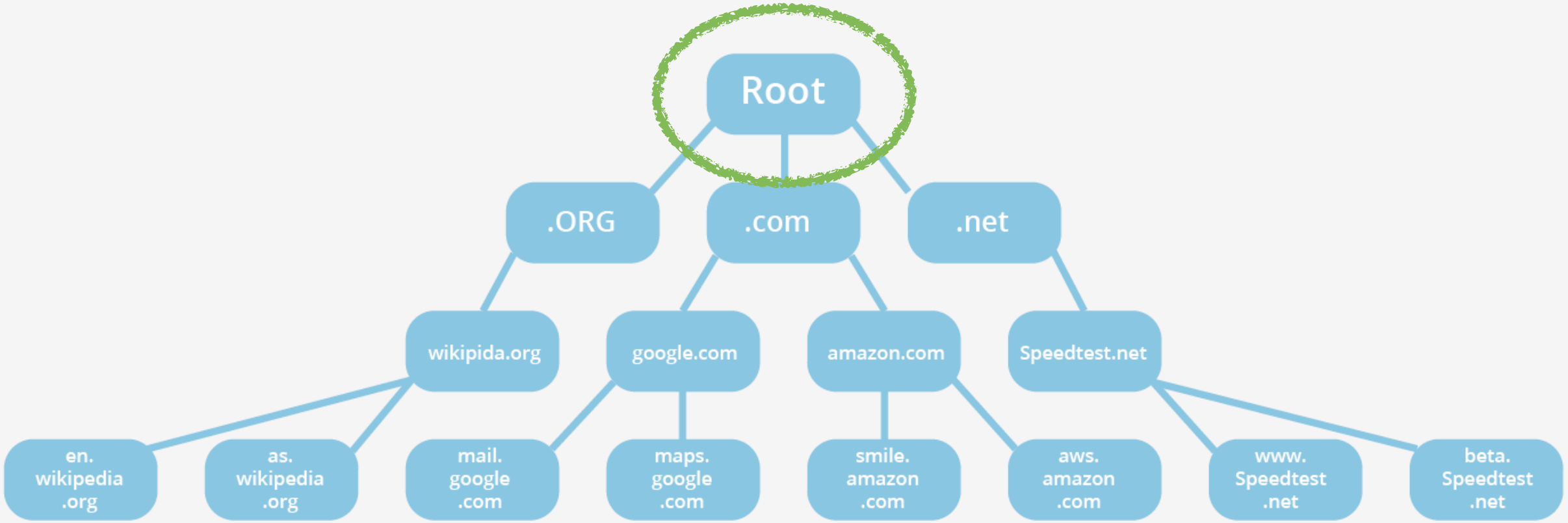
Resilient connectivity to INEX

- Capacity vs resilient?
  - Capacity means lagging additional ports to the same LAN

- We incentivise resilient connections to INEX via:
  - Free 1Gb or 10Gb port on INEX LAN2
  - Subsequent port pricing (80%) applies to 100Gb ports


INEX
Essential Internet Services
Over INEX

DNS Tree


Root
.ORG
.com
.net
wikipida.org
google.com
amazon.com
Speedtest.net
en.
wikipedia
.org
as.
wikipedia
.org
mail.
google
.com
maps.
google
.com
smile.
amazon
.com
aws.
amazon
.com
www.
Speedtest
.net
beta.
Speedtest
.net

DNS Root Zone

- Handled via 13 IP addresses
- Hardcoded:

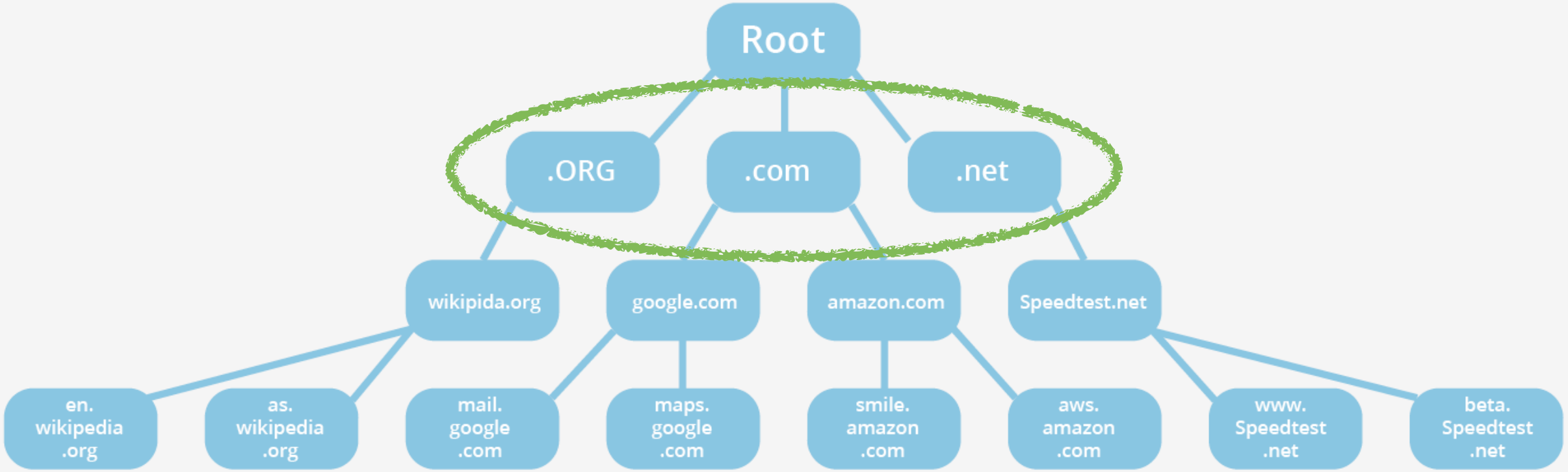
```
$ cat /usr/share/dns/root.hints

; OPERATED BY RIPE NCC
;
.                       3600000      NS    K.ROOT-SERVERS.NET.
K.ROOT-SERVERS.NET.     3600000      A     193.0.14.129
K.ROOT-SERVERS.NET.     3600000      AAAA  2001:7fd::1
```

DNS Root Zone at INEX

Server	INEX LAN1	INEX LAN2	INEX Cork
E (192.203.230.10)	PCH AS42 & RS		
F (192.5.5.241)	HE AS6939 (via 6939 1280 3557)		
J (192.58.128.30)	Verisign & RS	Verisign & RS	
K (193.0.14.129)	euNetworks; HE; RETN; Zayo; GEANT; VM; Nova; Colt; NORDUnet; & RS	euNetworks; HE; RETN; Zayo; VM; & RS	RIPE & RS
L (199.7.83.42)	HE; Zayo; GEANT; RETN; Equinix; NORDUnet.	HE; Zayo; RETN; Equinix; & RS	
M (202.12.27.33)	HE	HE & RS	

Bilateral peering sessions with Cloudflare should also deliver the F-root server.

DNS Tree


Root
.ORG
.com
.net
wikipida.org
google.com
amazon.com
Speedtest.net
en.
wikipedia
.org
as.
wikipedia
.org
mail.
google
.com
maps.
google
.com
smile.
amazon
.com
aws.
amazon
.com
www.
Speedtest
.net
beta.
Speedtest
.net

DNS TLDs at INEX

- .IE - multiple sources including IEDR, BT, HEAnet, NORDUnet
- .com / .net - via Verisign
- Packet Clearing House (PCH) - 126 ccTLDs plus 44 gov.cc and 23 mil.cc. Brand TLDs also.
- 8.8.8.8 / 8.8.4.4 / 9.9.9.9 all over INEX

RIPE Atlas Anchors

- Atlas Anchors are ‘super-probes’ - bigger machines and used as targets by the standard probes.
- INEX hosts one in Dublin *(currently offline)* and Cork.
- Two others in Ireland - HEAnet and Amazon *(also offline).*

AS112

- Intercepts DNS inverse requests for rfc1918 (and other address space).
  - Provides low-latency responses.
  - Soaks up queries leaked from networks.

- On the route servers but better to have bilateral sessions.

- https://www.as112.net/



Tools Available on IXP Manager

Cross Connect Records

Overview | Details | Ports | Cross Connects | Filtered Prefixes » | Peering Manager » | Statistics » | Peer to Peer Traffic »

Cross Connect

Name	Colocation Circuit Ref	State	Location	Assigned At	Chargeable	Owned By
IE.DUB.DUB1.2B.R03.01.U46 FXX/FXX (Fibre, duplex port: XX)	IExxxxxxxx	Connected	Digital Realty DUB1	2020-02-13	No	Customer
IE.DUB.DUB1.2B.R03.01.U41 FXX/FXX (Fibre, duplex port: XX)	IExxxxxxxx	Connected	Digital Realty DUB1	2020-02-13	No	Customer

Filtered Prefixes

Prefix	Filtered Because	Filtered On Router(s)
137.71.230.0/24	IRRDB PREFIX FILTERED	rs1-lan2-ipv4, rs2-lan2-ipv4, rs1-lan1-ipv4, rs2-lan1-ipv4
137.71.249.0/24	IRRDB PREFIX FILTERED	rs1-lan2-ipv4, rs2-lan2-ipv4, rs1-lan1-ipv4, rs2-lan1-ipv4
185.242.238.0/23	NEXT HOP NOT PEER IP	rs1-lan1-ipv4, rs2-lan1-ipv4

Peering Manager

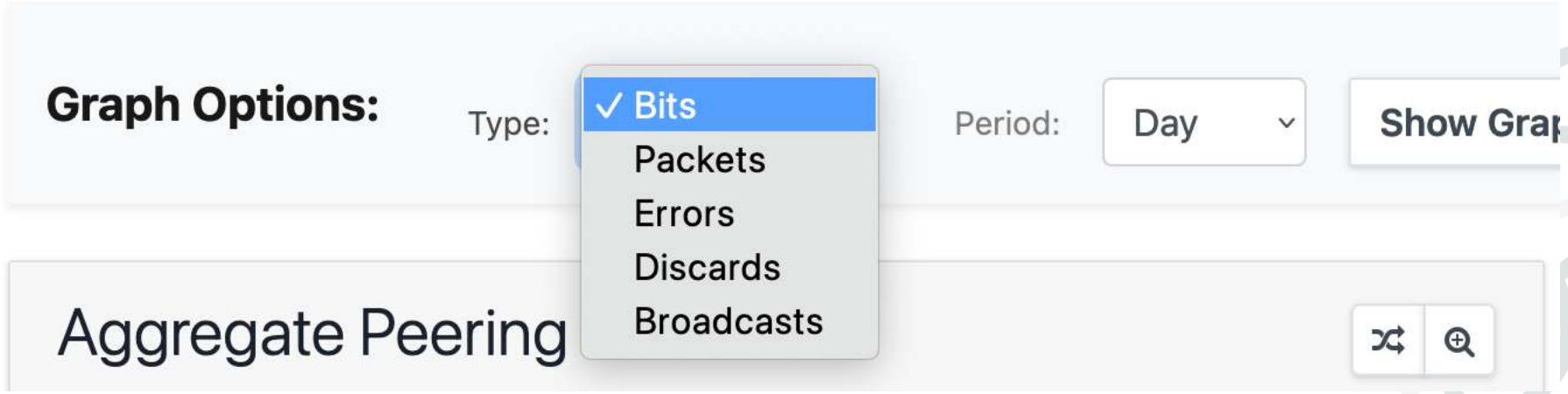
Potential Peers | Potential Bilateral Peers | Peers | Rejected / Ignored Peers

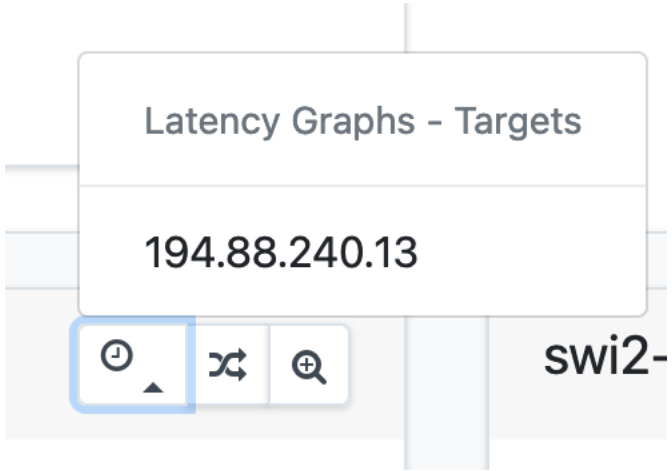
You currently do not exchange any routes in any way with the following members of the exchange **over the highlighted - in red - protocol(s) and LAN(s)** because:

- either you, they or both of you are not route server clients; and
- we have not detected that you have a bilateral peering session with them.

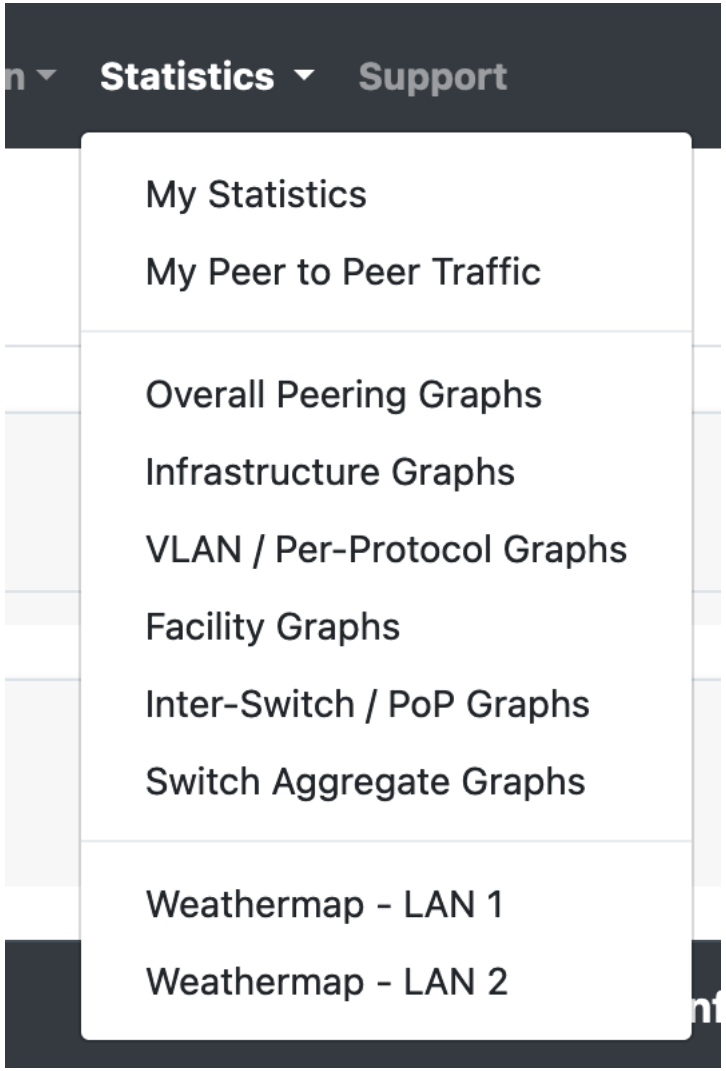
Member	ASN	Policy	INEX LAN1	INEX LAN2	
Cable & Wireless Worldwide	1273	selective	IPv4 IPv6		Request Peering Notes
Microsoft	8075	open	IPv4 IPv6	IPv4 IPv6	Request Peering Notes
ServiceNow Ireland Limited	16839	open	IPv4	IPv4	Request Peering Notes
Zayo	6461	selective	IPv4 IPv6	IPv4 IPv6	Request Peering Notes

Other Interface Graphing Options

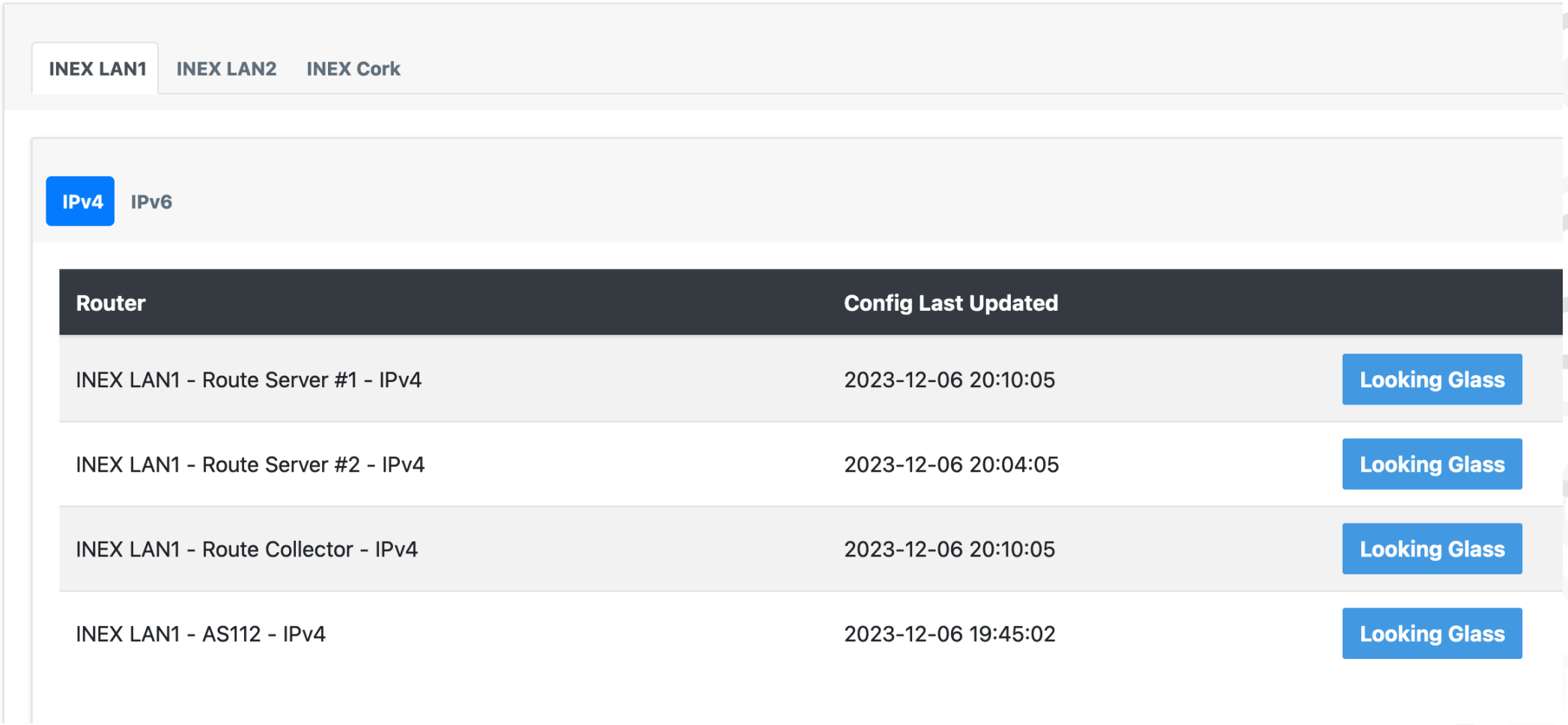




Public Graphs



Looking Glass



https://www.inex.ie/ixp/lg

Route Server Filtering

Rules in Production

Peer	LAN	Protocol	Advertised Prefix	Advertise Action	Received Prefix	Receive Action	Enabled	Order
Hurricane Electric	All	Both	*	Do Not Advertise	*	Do Not Receive (Drop)	Yes	1
Swisscom	All	Both	*	Do Not Advertise	*	Do Not Receive (Drop)	Yes	2

https://www.inex.ie/technical/route-servers/

Route Server Filtering

INEX

Your filters are not in sync with our production configuration. You can continue editing or: Revert Commit

Staged Rules (Deploy via Commit above)

Peer	LAN	Protocol	Advertised Prefix	Advertise Action	Received Prefix	Receive Action	Enabled	Order	Actions
Hurricane Electric	All	Both	*	Do Not Advertise	*	Do Not Receive (Drop)	Yes	1	

Rules in Production

Peer	LAN	Protocol	Advertised Prefix	Advertise Action	Received Prefix	Receive Action	Enabled	Order
Hurricane Electric	All	Both	*	Do Not Advertise	*	Do Not Receive (Drop)	Yes	1
Swisscom	All	Both	*	Do Not Advertise	*	Do Not Receive (Drop)	Yes	2

Other Tools

Data Sources

- PeeringDB - https://www.peeringdb.com/
- RIPE NCC's Whois Database
  - https://ftp.ripe.net/pub/stats/ripencc/membership/alloclist.txt
- RIPE NCC's RIS *("Routing Information Service")*
- PCH's MRT Routing Updates
  - https://www.pch.net/resources/Raw_Routing_Data/
  - https://www.pch.net/tools/looking_glass/
- https://www.routeviews.org/routeviews/
- https://ixpdb.euro-ix.net

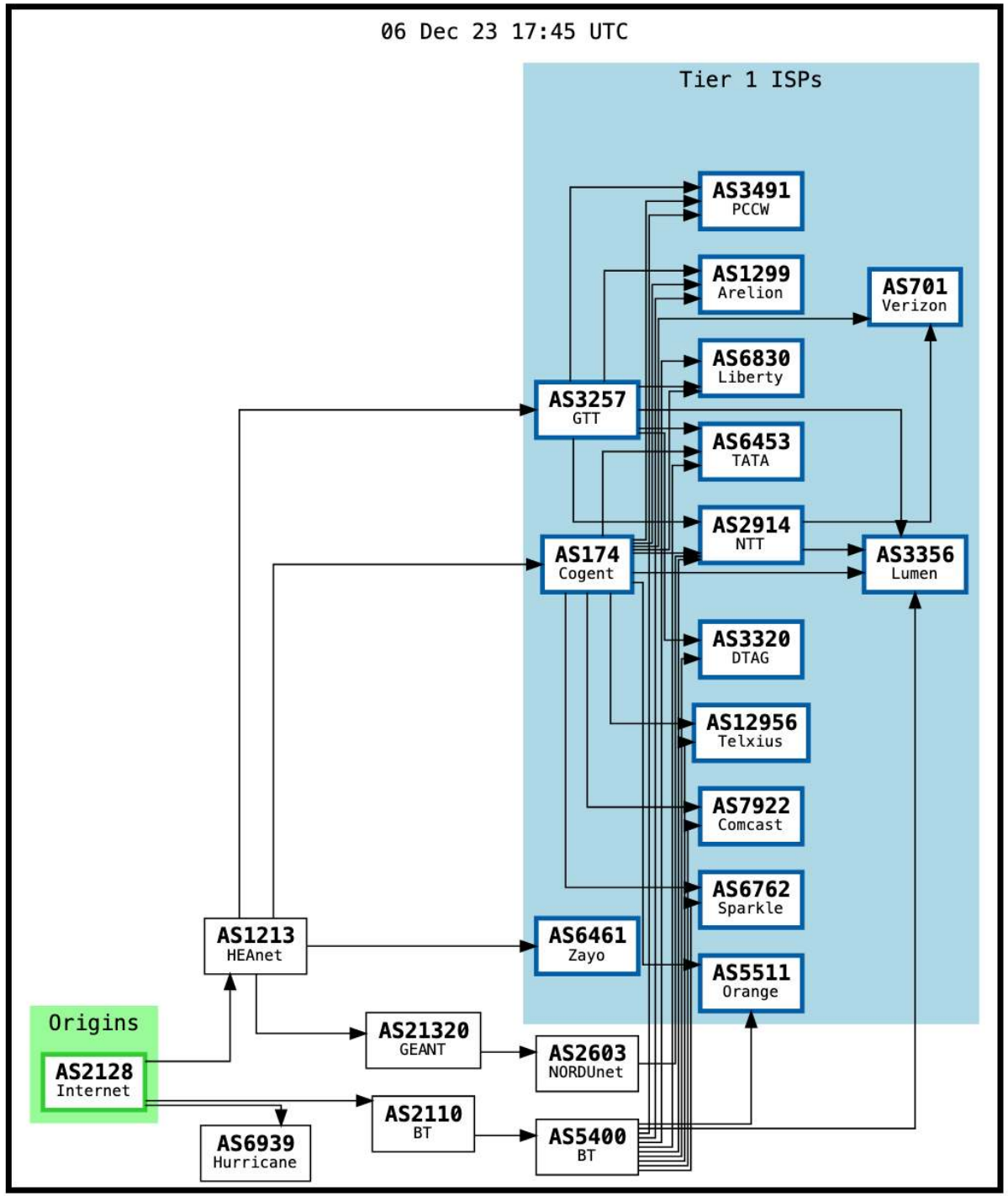
Reachability

- https://ping.pe/ - really useful and simple.
- RIPE Atlas - can specify many parameters
- https://www.jsdelivr.com/globalping - nice UI and FOSS
- https://mtr.tools/ - nice UI also

BGP Info and Diagnostics

- https://bgp.tools/ - very useful for visualising route propagation
- RIPE Stat - https://stat.ripe.net
  - https://stat.ripe.net/ui2013/ still exists!
- https://radar.cloudflare.com/ - security focused
- https://bgp.he.net/ - lots of info here (not all live)
- http://routing.he.net/ - pro tip: stick in an AS number.
- https://irrexplorer.nlnog.net/ - IRRDB and RPKI

https://bgp.tools/



Speedtesting

- http://speed.cloudflare.com/ - very nice UI and info
- https://fast.com/ (Netflix)
- https://www.speedtest.net/ - careful of the server choice
  - Blacknight have a 10Gb interface - good choice.
  - https://www.speedtest.net/apps/cli

```
$ speedtest -s 4604
    Server: Blacknight - Dublin (id: 4604)
    Download:   936.68 Mbps (data used: 422.7 MB)
    Upload:   938.67 Mbps (data used: 422.5 MB)
```


Thank you
INEX
INTERCONNECTING NETWORKS
AND PEOPLE FOR OVER 25 YEARS


INEX
INTERCONNECTING NETWORKS
AND PEOPLE FOR OVER 25 YEARS