BGP Next Hop Translation at IXPs

draft-marenamat-grow-route-server-nh-translation-00

Maria Matejka (CZ.NIC), Daniel Wagner (DE-CIX)

GROW @ IETF 124, Montréal 2025-11-06

Motivation (1)

- RFC8950[0] is there – yay!

[0] https://www.rfc-editor.org/rfc/rfc8950

Motivation (1)

- RFC8950[0] is there – yay!
- Run IXPs on IPv6 only
  - Some RIRs have reduced their standard allocation sizes for IXP peering LANs[1]
  - No more renumbering

[0] https://www.rfc-editor.org/rfc/rfc8950
[1] https://www.ripe.net/publications/docs/ripe-826/

Motivation (1)

- RFC8950[0] is there – yay!
- Run IXPs on IPv6 only
  - Some RIRs have reduced their standard allocation sizes for IXP peering LANs[1]
  - No more renumbering
- ...only if all BGP speakers support it

[0] https://www.rfc-editor.org/rfc/rfc8950
[1] https://www.ripe.net/publications/docs/ripe-826/

Motivation (2)

- Euro-IX started the RFC8950 WG[2]

[2] https://www.euro-ix.net/en/services/blog/servicesblogexploring-rfc8950-insights-from-working-group-chairs/

Motivation (2)

- Euro-IX started the RFC8950 WG [2]
- Goal:
  - Ensure coordinated transition
  - Provide IXP members with necessary configuration snippets enabling RFC8950 on their platform

[2] https://www.euro-ix.net/en/services/blog/servicesblogexploring-rfc8950-insights-from-working-group-chairs/

Motivation (2)

- Euro-IX started the RFC8950 WG [2]
- Goal:
    - Ensure coordinated transition
    - Provide IXP members with necessary configuration snippets enabling RFC8950 on their platform
- Side-product:
    - List of vendors (not…) supporting RFC8950 [3]

[2] https://www.euro-ix.net/en/services/blog/servicesblogexploring-rfc8950-insights-from-working-group-chairs/
[3] https://github.com/euro-ix/rfc8950-ixp/tree/main/examples

Problem Statement

- Some IXP members operate hardware that lacks support for RFC8950
- -> Roadblock to v6-only IXP peering LANs

Why an RFC?

- Got to update RFC7947[4]

[4] https://www.rfc-editor.org/rfc/rfc7947

Why an RFC?

- Got to update RFC7947[4]

- Lots of IXPs out there
- All facing the same problem
- One solution "done right" appreciated

[4] https://www.rfc-editor.org/rfc/rfc7947

Technical Proposal (-00 version)

- Terminology
  - Legacy Speaker (v4-only)
  - Supporting Speaker (dual-stack)
  - Unnumbered Speaker (v6-only)

- Main idea
  - Set up a BGP / ARP / ND snooper in the peering LAN that translates (next-hop) addresses based on a SLAT

Technical Proposal (-00 version)

- Translation table (Specific Local Address Table, SLAT)
  - Every speaker has a MAC address

```
SLAT
     +===================+
     | MAC Address       |
     |                   |
     |                   |
     +-------------------+
     | 00-00-5E-00-53-10 |
     +-------------------+
     | 00-00-5E-00-53-20 |
     +-------------------+
     | ...               |
     +-------------------+
```

Technical Proposal (-00 version)

- Translation table (Specific Local Address Table, SLAT)
  - Every speaker has a MAC address
  - Every speaker is assigned a LLA and GUA by the IXP

```
SLAT
     +===================+========+============+
     | MAC Address       |Link    |Global      |
     |                   |Local   |Unicast     |
     |                   |Address |Address     |
     +-------------------+--------+------------+
     | 00-00-5E-00-53-10 |FE80::10|2001:db8::10|
     +-------------------+--------+------------+
     | 00-00-5E-00-53-20 |FE80::20|2001:db8::20|
     +-------------------+--------+------------+
     | ...               |...     |...         |
     +-------------------+--------+------------+
```

Technical Proposal (-00 version)

- Translation table (Specific Local Address Table, SLAT)
  - Every speaker has a MAC address
  - Every speaker is assigned a LLA and GUA by the IXP
  - Every legacy speaker negotiates with the IXP a free, large enough IPv4 prefix

SLAT

MAC Address	Link Local Address	Global Unicast Address	CSLP 1	CLSP 2	...
00-00-5E-00-53-10	FE80::10	2001:db8::10	10.0.0.10	192.0.2.10	...
00-00-5E-00-53-20	FE80::20	2001:db8::20	10.0.0.20	192.0.2.20	...
...	...	...	...	...	...

Technical Proposal (-00 version)

- CSLP (Client-specific local prefix)
  - Can be the same for different legacy speakers
  - Can be any reserved / private v4 prefix short enough to house all BGP speakers in the peering LAN

Technical Proposal (-00 version)

- CSLP (Client-specific local prefix)
  - Can be the same for different legacy speakers
  - Can be any reserved / private v4 prefix short enough to house all BGP speakers in the peering LAN

  - Best case: get a /8 prefix from the reserved range ("IXP Interconnection Space")
  - "Ok" case: a (preferably small) number of v4 prefixes can be negotiated with all legacy speakers
  - Worst case: at least one legacy speaker cannot use any reserved / private v4 prefix for their addressing

Technical Proposal (-00 version)

- ARP / ND snooping
  - Done by some IXPs already to cope with BUM, ARP storms, etc...
  - E.g. with EVPN[5]

- BGP next hop translation
  - Can be done by the IXP Route Server (RS)

[5] https://blog.apnic.net/2023/08/16/peering-lan-2-0-introduction-of-evpn-at-de-cix/

Technical Proposal (-00 version)

- Security considerations
  - Snooping improves peering LAN security and control
  - MAC address registration & next hop translation should be checked by the IXP periodically and automatically
  - Mistakes in next hops / addressing can be detected and mitigated
  - In the worst case, an announcement gets dropped (better than propagating)

Technical Proposal (-00 version)

- Operational and management considerations
  - Step-by-step rollout (for existing IXPs)
    - Enable snooping, if not yet deployed
    - Enable NH translation at RSes (this draft), could be a new RS
    - Use current IPv4 prefix as first CSLP until its exhaustion
    - Negotiate additional CSLPs if necessary

Technical Proposal (-00 version)

- Operational and management considerations
  - Bilateral peerings between legacy and unnumbered speakers
    - Some peers do not use the RS
    - We cannot help them to translate their next hops (translation at the IXP switches?)
    - Publish SLAT
    - Or introduce BGP community telling the RS to exclude these routes from the Looking Glass (another draft?)

Technical Proposal (-00 version)

- Operational and management considerations
  - Bilateral peerings between legacy and unnumbered speakers
    - Some peers do not use the RS
    - We cannot help them to translate their next hops (translation at the IXP switches?)
    - Publish SLAT
    - Or introduce BGP community telling the RS to exclude these routes from the Looking Glass (another draft?)
  - Looking Glass enhancement
    - Show the original and translated next hops

Technical Proposal (-00 version)

- IANA considerations
  - This draft requests a /8 block from the 240.0.0.0/4 range
  - Termed "IXP Interconnection Space"
  - Should™ be usable by all peers as CSLP to keep the SLAT small

Thank you!

- Request for WG adoption
- Todos:
  - Fix some typos
  - Elaborate cases when IXPs can't do snooping
  - Get some running code

SLAT

```
+===================+========+============+=========+==========+===+
| MAC Address       |Link    |Global      |CSLP 1   |CLSP 2    |...|
|                   |Local   |Unicast     |         |          |   |
|                   |Address |Address     |         |          |   |
+-------------------+--------+------------+---------+----------+---+
| 00-00-5E-00-53-10 |FE80::10|2001:db8::10|10.0.0.10|192.0.2.10|...|
+-------------------+--------+------------+---------+----------+---+
| 00-00-5E-00-53-20 |FE80::20|2001:db8::20|10.0.0.20|192.0.2.20|...|
+-------------------+--------+------------+---------+----------+---+
```