BGP NH encoding discussion

IETF 105

Summary

- Shunwan pointed that RFC5549 (VPNv4 over IPv6) does not follow the NH encoding style from RFC4364 (VPNv4 over IPv4)

https://mailarchive.ietf.org/arch/msg/bess/LH2UeYZQxbm61Bj_JzztZqyU9JA

- His opinion was that the IPv6 NHs in RFC5549 should be encoded as a VPNv6 address (RD=0 + IPv6 address) rather than just regular IPv6 adresses

Discussion summary

- RFC5549 is very clear on how the NH should be encoded (no RD). However it is not consistent with RFC4364/RFC4659
- RD is an NLRI property (uniqueness), not a NH property
- RFC4364/4659 authors were trying to match the NLRI format in the NH (comes from RFC2858 ?)
- When two IPv6 NHs are used in the same NH field, which one should be used ?
- Do all implementations use the NH length field to parse the NH ?

Non exhaustive status

RFC#	Purpose	NH len	NH encoding	NH address length check	What the text says
RFC4760	MP-BGP definition	N/A	Identified by the AFI/SAFI	N/A	The Network Layer protocol associated with the Network Address of the Next Hop is identified by a combination of <AFI, SAFI> carried in the attribute.
RFC4364	BGP/MPLS IPVPN	12 bytes	<RD=0:IPv4>	N	When a PE router distributes a VPN-IPv4 route via BGP, it uses its own address as the "BGP next hop". This address is encoded as a VPN-IPv4 address with an RD of 0. ([BGP-MP]("->RFC2858") requires that the next hop address be in the same address family as the Network Layer Reachability Information (NLRI).)
RFC4659	BGP/MPLS IPv6VPN	24/48 bytes	<RD=0:IPv4-mapped IPv6> <RD=0:IPv6> <RD=0:IPv6><RD:IPv6>	Y	The value of the Length of the Next Hop Network Address field in the MP_REACH_NLRI attribute shall be set to 24 when only a global address is present, and to 48 if a link-local address is also included in the Next Hop field.
RFC4684	RT-constraint	4/16 bytes	<IPv4> <IPv6>	Y	The Next Hop field of MP_REACH_NLRI attribute shall be interpreted as an IPv4 address whenever the length of NextHop address is 4 octets, and as a IPv6 address whenever the length of the NextHop address is 16 octets.
RFC4798	6PE	16 bytes	<IPv4-mapped IPv6>	N	The IPv4 address of the egress 6PE router MUST be encoded as an IPv4-mapped IPv6 address in the BGP Next Hop field.
RFC5549	IPv4 NLRI with IPv6 NH (applies to SAFI 1,2,4,128)	16/32 bytes	<IPv6> <IPv6><IPv6>	Y	The BGP speaker receiving the advertisement MUST use the Length of Next Hop Address field to determine which network-layer protocol the next hop address belongs to. When the Length of Next Hop Address field is equal to 16 or 32, the next hop address is of type IPv6. **The RFC introduces the extended NH encoding capability which tells that for one or more AFI/SAFI, NH length is determined by the value of the Length of NH field. It only applies to family 1, SAFI (1,2,4,128).**

Non exhaustive status

RFC#	Purpose	NH len	NH encoding	NH address length check	What the text says
RFC6074	Prov, Disc, Sig of L2VPNs	4 bytes	<IPv4> <IPv6>	Not clear	a BGP next hop equal to the loopback address of the PE
RFC6514	BGP MVPN encoding	4/16 bytes	<IPv4> <IPv6>	Not clear (clarified by RFC6515)	C-mcast routes: The Next Hop field of the MP_REACH_NLRI attribute MUST be set to a routable IP address of the local PE. SA routes: The Next Hop field of the MP_REACH_NLRI attribute MUST be set to the IP address that the PE places in the Global Administrator field of the VRF Route Import Extended Community of the VPN-IP routes advertised by the PE from the MVPN's VRF. Other routes: The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router's IP Address field. **SAFI 129 is defined but without real NH encoding rules**
RFC6515	Clarifies infra address encoding in BGP mVPN			Y	"Network Address of Next Hop" field in the MP_REACH_NLRI attribute, as defined in Section 3 of [BGP-MP]. This field is preceded by a "length of next hop address" field. Hence, it is always clear whether the address is an IPv4 address (length is 4) or an IPv6 address (length is 16).
RFC7432	EVPN	4/16 bytes	<IPv4> <IPv6>	Not clear	The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

Robert's proposal on IDR list

- Do an implementation survey on how next hop parsing is handled:
  - Is the implementation capable of inferring NH format from NH len ?
  - If yes, does this apply to any AFI/SAFI or subset ?
  - Does the implementation support encoding of two IPv6 nexthops in the same NH field, how is it used ?

Open mic

- Is there something to do on this topic ?
  - Do we need to be more precise in the next specifications ? Define guidelines ?
  - Do we have to fix some specifications ?
  - Is the survey proposed by Robert worth doing ?
  - Any other poing ?