Network Working Group
Internet-Draft
Updates: 6514 (if approved)
Intended status: Standards Track
Expires: 24 May 2024

F. Duan
J. Xie
S. Chen
Huawei Technologies

21 November 2023

BGP MVPN in IPv6 Infrastructure Networks: Problems and Solution Approaches

draft-duan-bess-mvpn-ipv6-infras-05

Abstract

MVPN deployment faces some problems while used in provider's IPv6 infrastructure networks. This document describes these problems, and corresponding solutions.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

BGP MVPN procedure is defined in [RFC6514]. As a mature MVPN technology, it has been accepted by most operators and vendors. In [RFC6515], BGP MVPN is updated for IPv6 infrastructure networks. However, the deployment of BGP MVPN in IPv6 network still faces some problems. This document describes these problems and corresponding solutions.

2. Terminology

Readers of this document are assumed to be familiar with the terminology and concepts of the documents listed as Normative References.

3. Problems and Solutions

3.1. Problems

In [RFC6514] and [RFC6515], the following issues are critical for IPv6 infrastructure scenarios while a non-segmented inter-AS P-tunnel is being established between root PE and its leaf PEs, of which the reason is that the Source AS field (4 octets) of C-multicast route cannot hold a IPv6 address (16 octets).

1. In order to distinguish the C-multicast routes for a specific multicast c-flow (C-S, C-G) sent to different root PEs, section 11.1.3 of [RFC6514] proposed that "To support non-segmented inter-AS tunnels, the Source AS field in the C-multicast route is set to value of the Originating Router's IP Address field of the found Intra-AS I-PMSI A-D route". However, in NLRI of C-multicast route, Source AS field is 4 octets in length, while the Originating Router's IP Address field of Intra-AS I-PMSI A-D route is 16 octets length in provider's IPv6 networks. The 4-octet Source AS field cannot hold the Originating Router's IP Address in IPv6 network.

2. In order to control the propagation of C-multicast routes between different ASes, section 11.2 of [RFC6514] proposed that "Instead of matching the RD and Source AS carried in the C-multicast route against the RD and Source AS of an Inter-AS I-PMSI A-D route, the ASBR should match it against the RD and the Originating Router's IP Address of the Intra-AS I-PMSI A-D routes". However, Source AS field in NLRI of C-multicast route cannot be translated to the Originating Router's IP Address of the Intra-AS I-PMSI A-D routes in provider's IPv6 networks, because of the mismatch of their field length.

In the process of evolution to IPv6, IPv4 and IPv6 infrastructure addresses will coexist in the provider's network. The following figure is an example of BGP MVPN evolution to IPv6.

```
 +-----------+           +---+     +---+           +-----------+
 | +-------+ |           |   |     |   |           | +-------+ |
 | | O-MVRF| |=BGP Peer4=|   |     |   |=BGP Peer4=| | O-MVRF| |
 | +-------+ |           |   |     |   |           | +-------+ |
 |           |           |   |     |   |           |           |
 |    PE1    |           |RR1| ... |RRn|           |    PE2    |
 |           |           |   |     |   |           |           |
 | +-------+ |           |   |     |   |           | +-------+ |
 | | N-MVRF| |=BGP Peer6=|   |     |   |=BGP Peer6=| | N-MVRF| |
 | +-------+ |           |   |     |   |           | +-------+ |
 +-----------+           +---+     +---+           +-----------+
 O-MVRF = Old Multicast VRF using IPv4 infrastructure addresses
 N-MVRF = New Multicast VRF using IPv6 infrastructure addresses
```

Figure 1: BGP MVPN Evolution to IPv6 Infrastructure

During the evolution process, IPv4 and IPv6 parallel BGP sessions are established between Provider Edge routers and route reflectors(RR). If the BGP MVPN routes are sent to all IPv4 and IPv6 BGP peers without any control, the number of the PATHs of these routes will be doubled with each reflection while BGP ADD-PATH [RFC7911] procedure is enabled on route reflectors.

3.2. Modification of C-Multicast Route NLRI

The solution to distinguish the C-multicast routes sent to different root PEs is related to the way to distinguish UMH routes for a specific multicast source (C-S) sent from different root PEs, which the later is not a problem of IPv6 infrastructure specific. In [RFC6514], it recommended that the RDs of root PEs of a same MVPN were configured distinctly to perform selective forwarding selection, which was broken by GTM procedures defined in [RFC7716] because the UMH routes sent from different root PEs through BGP SAFI 1 or SAFI 2 lack RD informations. There are also some MVPN deployment cases that the RDs of root PEs may be configuered with a same value for provisioning reasons. According to above description, whether the RDs of PEs of a MVPN are same or not are two different deployment cases. This document addresses the C-multicast routes distinguishing issue for both cases. How to distinguish UMH routes in the cases of root PEs with same RD is out of the scope of this document, because it is not IPv6 infrastructure specific.

To support non-segmented inter-AS tunnels in IPv6 infrastructure network, the C-multicast route NLRI is redefined as follow:

```
+-----------------------------------+
|      RD   (8 octets)              |
+-----------------------------------+
|   Root Distinguisher (4 octets)   |
+-----------------------------------+
| Multicast Source Length (1 octet) |
+-----------------------------------+
|   Multicast Source (variable)     |
+-----------------------------------+
|  Multicast Group Length (1 octet) |
+-----------------------------------+
|  Multicast Group   (variable)     |
+-----------------------------------+
```

In the above figure, the Root Distinguisher field replaces the Source As field defined in [RFC6514]. When constructing a C-multicast route, leaf PE follows the following specification:

1. For the cases of IPv4 infrastructure or Intra-AS P-tunnel establishment in IPv6 infrastructure, the Root Distinguisher field MUST be treated as Source AS field and section 11.1.3 of [RFC6514] MUST be fully followed.

2. For non-segmented Inter-AS P-tunnel establishment in IPv6 infrastructure scenarios, if the RDs of ingress PEs are distinct (which can be detected from UMH routes), the Root Distinguisher field MUST be filled with the number of ingress AS.

3. For non-segmented Inter-AS P-tunnel establishment in IPv6 infrastructure scenarios, if the RDs of ingress PEs are same, a 4-octet distinct value MUST be assigned by leaf PE for each root PE. For example, each leaf PE uses a same well-known / configured hash algorithm to transform the IPv6 root IP to 4-octet distinct value for each ingress PE, or a provisioning method is used to globally assign different 4-octet IDs for each ingress PE. The Root Distinguisher field in C-multicast NLRI is filled with this value and a distinct C-multicast route will be sent to individual upstream root PE.

The solution to control the propagation of C-multicast route between different ASes is to use the IPv6 address included in IPv6 VRF Route Import Extended Community insteading of Source AS field of C-multicast NLRI while locating Intra-AS AD route of the corresponding root PE the C-multicast sent to on ASBRs. This document recommends that the Local Administrator field of IPv6 VRF Route Import Extended Community is set to a non-zero value by root PEs even in GTM scenarios, of which the value is local assigned distinctly by root PE for both each MVPN and GTM instance. Accordingly, the IPv6 root address of a C-multicast route can be extracted from the only IPv6 VRF Route Import Extended Community carring a non-zero Local Administrator field.

When receiving a C-multicast route from E-BGP neighbors, the ASBR checks whether an IPv6 VRF Route Import Extended Community with a non-zero Local Administrator field is included in this route and takes following actions:

1. If the IPv6 VRF Route Import Extended Community does not exist in the C-multicast route, the ASBR will treat the Root Distinguisher field as Source AS field and follows the description in section 11.2 of [RFC6514].

2. If the IPv6 VRF Route Import Extended Community does exist in the C-multicast route, the ASBR will match the IPv6 address carried in this extended community and the RD in C-multicast route NLRI against the Originating Router’s IP Address and the RD of the

Intra-AS I-PMSI A-D routes. If the corresponding Intra-AS I-PMSI A-D route exists, the ASBR will propagate the C-multicast route in its local AS.

3.3. Route Reflection Control

To reduce BGP MVPN routes in Parallel IPv4 and IPv6 BGP sessions scenario, the following actions should be taken by sender PEs:

1. For Intra-AS I-PMSI A-D Route, S-PMSI A-D Route and Leaf A-D Route, if the Originating Router's IP Address field in the route is filled with an IPv6 address, it will be sent to the IPv6 BGP neighbors; otherwise, it will be sent to the IPv4 BGP neighbors.

2. For Inter-AS I-PMSI A-D Route and Source Active A-D Route, it is sent to both IPv6 BGP neighbors and IPv4 BGP neighbors.

3. For C-multicast route, if the IPv6 VRF Route Import Extended Community exists in the route, it will be sent to the IPv6 BGP neighbors; otherwise, it will be sent to the IPv4 BGP neighbors.

In the route reflectors, the part of routes which are received from IPv6 BGP neighbors will be reflected to other IPv6 BGP neighbors and the other part of routes which are received from IPv4 BGP neighbors will be reflected to other IPv4 BGP neighbors.

4. Security Considerations

This document introduces no new security considerations beyond those already specified in [RFC6514] and [RFC6515].

5. IANA Considerations

This document contains no actions for IANA.

6. Acknowledgements

Your name here

7. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC6515] Aggarwal, R. and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, DOI 10.17487/RFC6515, February 2012, <https://www.rfc-editor.org/info/rfc6515>.

[RFC7716] Zhang, J., Giuliano, L., Rosen, E., Ed., Subramanian, K., and D. Pacella, "Global Table Multicast with BGP Multicast VPN (BGP-MVPN) Procedures", RFC 7716, DOI 10.17487/RFC7716, December 2015, <https://www.rfc-editor.org/info/rfc7716>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Authors' Addresses

Fanghong Duan
Huawei Technologies
Email: duanfanghong@huawei.com

Jingrong Xie
Huawei Technologies
Email: xiejingrong@huawei.com

Siyu Chen
Huawei Technologies
Email: chensiyu27@huawei.com

BESS Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: 21 November 2023

G. Mishra
Verizon Inc.
M. Mishra
Cisco Systems
J. Tantsura
Microsoft, Inc.
S. Madhavi
Juniper Networks, Inc.
Q. Yang
Arista Networks
A. Simpson
Nokia
S. Chen
Huawei Technologies
20 May 2023

IPv6-Only PE Design for IPv4-NLRI with IPv6-NH
draft-ietf-bess-ipv6-only-pe-design-04

Abstract

As Enterprises and Service Providers upgrade their brown field or green field MPLS/SR core to an IPv6 transport, Multiprotocol BGP (MP-BGP)now plays an important role in the transition of their Provider (P) core network as well as Provider Edge (PE) Edge network from IPv4 to IPv6. Operators must be able to continue to support IPv4 customers when both the Core and Edge networks are IPv6-Only.

This document details an important External BGP (eBGP) PE-CE Edge and Inter-AS IPv6-Only peering design that leverages the MP-BGP capability exchange by using IPv6 peering as pure transport, allowing both IPv4 Network Layer Reachability Information (NLRI) and IPv6 Network Layer Reachability Information (NLRI)to be carried over the same (Border Gateway Protocol) BGP TCP session. The design change provides the same Dual Stacking functionality that exists today with separate IPv4 and IPv6 BGP sessions as we have today. With this design change from a control plane perspective a single IPv6 is required for both IPv4 and IPv6 routing updates and from a data plane forwarindg perspective an IPv6 address need only be configured on the PE and CE interface for both IPv4 and IPv6 packet forwarding.

This document provides a much needed solution for Internet Exchange Point (IXP) that are facing IPv4 address depletion at large peering points. With this design, IXP can now deploy PE-CE IPv6-Only eBGP Edge or Inter-AS peering design to eliminate IPv4 provisioning at the Edge. This core and edge IPv6-Only peering design paradigm change can apply to any eBGP peering, public internet or private, which can

be either Core networks, Data Center networks, Access networks or can be any eBGP peering scenario. This document provides vendor specific test cases for the IPv6-Only peering design as well as test results for the five major vendors stakeholders in the routing and switching indusrty, Cisco, Juniper, Arista, Nokia and Huawei. With the test results provided for the IPv6-Only Edge peering design, the goal is that all other vendors around the world that have not been tested will begin to adopt and implement this new Best Current Practice for eBGP IPv6-Only Edge peering.

As this issue with IXP IPv4 address depletion is a critical issue around the world, it is imperative for an immediate solution that can be implemented quickly. This Best Current Practice IPv6-only eBGP peering design specification will help proliferate IPv6-Only deployments at the eBGP Edge network peering points to starting immediately at a minimum with operators around the world using Cisco, Juniper, Arista, Nokia and Huawei. As other vendors start to implement this Best Current Practice, the IXP IPv4 address depletion gap will eventually be eliminated.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

As Enterprises and Service Providers upgrade their brown field or green field MPLS/SR core to an IPv6 transport such as MPLS LDPv6, SR-MPLSv6 or SRv6, Multiprotocol BGP (MP-BGP) now plays an important role in the transition of the Provider (P) core networks and Provider Edge (PE) edge networks from IPv4 to IPv6. Operators have a requirement to support IPv4 customers and must be able to support IPv4 address family and Sub-Address-Family Virtual Private Network (VPN)-IPv4, and Multicast VPN IPv4 customers.

IXP are also facing IPv4 address depletion at their peering points, which are large Layer 2 transit backbones that service providers peer and exchange IPv4 and IPv6 Network Layer Reachability Information (NLRI). Today, these transit exchange points are Dual Stacked. With this IPv6-only BGP peering design, only IPv6 is configured on the PE-CE interface, the Provider Edge (PE) - Customer Edge (CE), or Inter-AS ASBR (Autonomous System Boundary Router) to ASBR (Autonomous System Boundary Router) PE-PE Provider Edge (PE) - Provider Edge (PE), the IPv6 BGP peer is now used to carry IPv4 (Network Layer Reachability Information) NLRI over an IPv6 next hop using IPv6 next hop encoding defined in [RFC8950], while continuing to forward both IPv4 and IPv6 packets. In the framework of this design the PE is no longer Dual Stacked. However in the case of the CE, PE-CE link CE side of the link is no longer Dual Stacked, however all other internal links within the CE domain may or maynot be Dual stacked. In the Inter-AS case the ASBR-ASBR PE-PE peering all peerings would be now IPv6-Only for all Inter-AS options Peering Option-A, Option-B, Option-AB and Option-C per [RFC4364]. We now refer to this PE as an "IPv6-Only PE" using the IPv6-Only PE Design framework.

MP-BGP specifies that the set of usable next-hop address families is determined by the Address Family Identifier (AFI) and the Subsequent Address Family Identifier (SAFI). Historically the AFI/SAFI definitions for the IPv4 address family only have provisions for

advertising a Next Hop address that belongs to the IPv4 protocol when advertising IPv4 or VPN-IPv4. [RFC8950] specifies the extensions necessary to allow advertising IPv4 NLRI, Virtual Private Network Unicast (VPN-IPv4) NLRI, Multicast Virtual Private Network (MVPN-IPv4) NLRI with a Next Hop address that belongs to the IPv6 protocol. This comprises of an extended next hop encoding MP-REACH BGP capability exchange to allow the address of the Next Hop for IPv4 NLRI, VPN-IPv4 NLRI and MVPN-IPv4 NLRI to also belong to the IPv6 Protocol. [RFC8950] defines the encoding of the Next Hop to determine which of the protocols the address actually belongs to, and a new BGP Capability allowing MP-BGP Peers to discover dynamically whether they can exchange IPv4 NLRI and VPN-IPv4 NLRI with an IPv6 Next Hop.

The current specification for carrying IPv4 NLRI of a given address family via a Next Hop of a different address family is now defined in [RFC8950], and specifies the extended next hop encoding MP-REACH capability extension necessary to do so. This comprises an extension of the AFI/SAFI definitions to allow the address of the Next Hop for IPv4 NLRI or VPN-IPv4 NLRI to belong to either the IPv4 or the IPv6 protocol, the encoding of the Next Hop information to determine which of the protocols the address belongs to, and a new BGP Capability allowing MP-BGP peers to dynamically discover whether they can exchange IPv4 NLRI and VPN- IPv4 NLRI with an IPv6 Next Hop.

With the new extensions defined in [RFC8950] supporting NLRI and next hop address family mismatch, the BGP peer session can now be treated as a pure TCP transport and carry both IPv4 and IPv6 NLRI at the Provider Edge (PE) - Customer Edge (CE) or Inter-AS ASBR to ASBR PE-PE over a single IPv6 TCP session. This allows for the elimination of dual stack from the PE-CE and Inter-AS ASBR-ASBR PE-PE peering point, and now enable the peering to be IPv6-ONLY. The elimination of IPv4 on the PE-CE and Inter-AS ASBR-ASBR PE-PE peering points translates into OPEX expenditure savings of point-to-point infrastructure links as well as /31 address space savings and administration and network management of both IPv4 and IPv6 BGP peers. This reduction decreases the number of PE-CE BGP peers by fifty percent, which is a tremendous cost savings for operators.

While the savings exists at the Edge eBGP PE-CE peering, on the core side PE to Route Reflector (RR) peering carrying <AFI/SAFI> IPv4 <1/1>, VPN-IPV4 <1/128>, and Multicasat VPN <1/129>, there is no savings as the Provider (P) Core is IPv6 Only and thus can only have an IPv6 peer and must use [RFC8950] extended next hop encoding to carrying IPv4 NLRI IPV4 <2/1>, VPN-IPV4 <2/128>, and Multicasat VPN <2/129> over an IPv6 next hop.

This IPv6-Only PE design is applicable to both PE-CE Edge over a IPv4-Only Core, IPv6-Only Core as well as Global table or VPN overlay scenario as well as all Inter-AS Options Option-A, Option-B, Option-AB and Option-C The following Address Family (AFI) / Subsequent Address Family (SAFI) will be tested with both IPv4-Only Core, IPv6-Only Core and Global Routing Table (GRT) and IP Virtual Private Network (VPN) [RFC4364]. <AFI/SAFI> IPv4 <1/1>, VPN-IPV4 <1/128>, and Multicasat VPN <1/129>.

This document provides a much needed solution for Internet Exchange Point (IXP) that are facing IPv4 address depletion at large peering points. With this design, IXP can now use deploy PE-CE IPv6-Only eBGP Edge and Inter-AS peering design to eliminate IPv4 provisioning at the PE Edge as well as PE Inter-AS. This core and edge IPv6-Only peering design paradigm change can apply to any eBGP peering, public internet or private, which can be either Core networks, Data Center networks, Access networks or can be any eBGP peering scenario. This document provides detailed vendor specific test cases and test results for the IPv6-Only peering design as well as successful test results between five major vendors stakeholders in the routing and switching indusrty, Cisco, Juniper, Arista, Nokia and Huawei. With the test results provided for the IPv6-Only Edge peering design, the goal is that all other vendors around the world that have not been tested will begin to adopt and implement this new best practice for eBGP IPv6-Only Edge peering. This will give confidence to operators to start the proliferation of this IPv6-Only PE design.

As this issue with IXP address depletion is a critical issue around the world, it is imperative for an immediate solution that can be implemented quickly. This best practice IPv6-only eBGP peering design specification will help proliferate IPv6-Only deployments at the eBGP Edge and Inter-AS network peering points starting immediately at a minimum with operators around the world using Cisco, Juniper, Arista, Nokia and Huawei.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

Terminolgoy used in defining the IPv6-Only Edge specification.

AFBR: Address Family Border Router Provider Edge (PE).

Edge: PE-CE Edge Network Provider Edge - Customer Edge

Core: P Core Network Provider (P)

4to6 Softwire : IPv4 edge over an IPv6-Only core

6to4 Softwire: IPv6 edge over an IPv4-Only core

E2E: End to End

4. IPv6-Only Edge Peering Architecture

4.1. Problem Statement

This specification addresses a real world issue that has been discussed at many operator groups around the world related to IXP major peering points where hundreds of AS's have both IPv4 and IPv6 dual stacked peering. IPv4 address depletion have been a major issue issue for many years now. Operators around the world are clamoring for a solution that can help solve issues related to IPv4 address depletion at these large IXP peering points. With this solution IXPs as well as all infrastructure networks such as Core networks, DC networks, Access networks as well as any PE-CE public or private network can now utilize this IPv6-Only Edge solution and reap the benefits immediately on IPv4 address space saving.

```
                    IXP Problem Statement

          Dual Stacked            Dual Stacked
               CE                      PE

            +-------+ IPv4 BGP Peer +-------+
            |       |---------------|       |
            |   CE  | IPv6 BGP Peer |  PE   |
            |       |---------------|       |
            +-------+               +-------+
           IPv4 forwarding           IPv4 forwarding
           IPv6 forwarding           IPv6 forwarding
```

Figure 1: Problem Statement - IXP Dual Stack Peering

```
                                  _________
    Dual Stacked     _____     /         \                Dual Stacked
      PE / CE       /     \__/           \___              PE / CE
  +----+  +----+   /                        \        +------+   +-----+
  |    |  |    |  |0====VPN Overlay Tunnel ==0|      |      |   |     |
  |    |  |    |  |                            \     |      |   |     |
  | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
  |    |  |    |    \0=========Underlay =======0|     |      |   |     |
  +----+  +----+     \                        __/    +------+   +-----+
  IPv4 IPv6 BGP peer  \ IP / MPLS / SR domain /     IPv4 and IPv6 BGP peer
  IPv4 forwarding      \__          __        /        IPv4 forwarding
  IPv6 forwarding         \_______/  \_____/           IPv6 forwarding
```

Figure 2: Problem Statement - E2E Dual Stack Edge

4.2. IPv6-Only PE-CE Design Solution

The IPv6-Only Edge design solution provides a means of E2E single protocol design solution extension of [RFC5565] Softwire Mesh framework from the PE-CE Edge to the Core from ingres so egress through the entire operators domain. This solution eliminates all IPv4 addressing from end to end while still providing the same Dual Stack functionality of IPv4 and IPv6 packet forwarding from a data plane perspective by leveraging the [RFC8950] extended next hop encoding so that IPv4 NLRI can be advertised over a single IPv6 pure transport TCP session. This IPv6-Only E2E architecture eliminates all IPv4 peering and IPv4 addressing E2E from the ingress CE to ingress PE to egress PE to egress CE and all hops along the operator E2E path.

```
                    Solution applicable to
     any Edge peering scenario - IXP, Core, DC, Access, etc

      +-------+                +-------+
      |       |  IPv6 Only     |       |
      |   CE  |----------------|  PE   |
      |       |  IPv6 BGP Peer |       |
      +-------+                +-------+
     IPv4 forwarding            IPv4 forwarding
     IPv6 forwarding            IPv6 forwarding
```

Figure 3: IPv6-Only Solution Applicability

```
                                  _________
      IPv6-Only       _____    /         \                IPv6-Only
       PE / CE       /     \__/           \___             PE / CE
   +----+  +----+   /                        \         +------+   +-----+
   |    |  |    |  |0====VPN Overlay Tunnel ==0|       |      |   |     |
   |    |  |    |  |                            \      |      |   |     |
   | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay ===== ==0     |      |   |     |
   +----+  +----+     \                          __/    +------+   +-----+
   IPv6 BGP peer        \IP / MPLS / SR domain /        IPv6 BGP peer
   IPv4 forwarding      \__          __        /         IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/           IPv6 forwarding
```

Figure 4: E2E VPN Solution

4.3. IPv6-Only Edge Peering Design

4.3.1. IPv6-Only Edge Peering Packet Walk

The IPv6-Only Edge Peering design utilizes two key E2E Softwire Mesh Framework scenario's, 4to6 softwire and 6to4 softwire. The Softwire mesh framework concept is based on the overlay and underlay MPLS or SR based technology framework, where the underlay is the transport layer and the overlay is a Virtual Private Network (VPN) layer, and is the the tunneled virtualization layer containing the customer payload. The concept of a 6to4 Softwire is based on transmission of IPv6 packets at the edge of the network by tunneling the IPv6 packets over an IPv4-Only Core. The concept of a 4to6 Softwire is also based on transmission of IPv4 packets at the edge of the network by tunneling the IPv4 packets over an IPv6-Only Core.

This document describes End to End (E2E) test scenarios that follow a packet flow from IPv6-Only attachment circuit from ingress PE-CE to egress PE-CE tracing the routing protocol control plane and data plane forwarding of IPv4 packets in a 4to6 softwire or 6to4 softwire within the IPv4-Only or IPv6-Only Core network. In both secneario we are focusing on IPv4 packets and the control plane and data plane forwarding aspects of IPv4 packets from the PE-CE Edge network over an IPv6-Only P (Provider) core network or IPv4-Only P (Provider) core network. With this IPv6-Only Edge peering design, the Softwire Mesh Framework is not extended beyond the Provider Edge (PE) and continues to terminate on the PE router.

4.3.2. 6to4 Softwire IPv4-Only Core packet walk

6to4 softwire where IPv6-Edge eBGP IPv6 peering where IPv4 packets at network Edge traverse a IPv4-Only Core

In the scenario where IPv4 packets originating from a PE-CE edge are tunneled over an MPLS or Segment Routing IPv4 underlay core network, the PE and CE only have an IPv6 address configured on the interface. In this scenario the IPv4 packets that ingress the CE from within the CE AS are over an IPv6-Only interface and are forwarded to an IPv4 NLRI destination prefix learned from the Pure Transport Single IPv6 BGP Peer. In the IPv6-Only Edge peering architecture the PE is IPv6-Only as all PE-CE interfaces are IPv6-Only. However, on the CE, the PE-CE interface is the only interface that is IPv6-Only and all other interfaces may or may not be IPv6-Only. Following the data plane packet flow, IPv4 packets are forwarded from the ingress CE to the IPv6-Only ingress PE where the VPN label imposition push per prefix, per-vrf, per-CE occurs and the labeled packet is forwarded over a 6to4 softwire IPv4-Only core, to the egress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the egress CE. In the reverse direction IPv4 packets are forwarded from the egress CE to egress PE where the VPN label imposition per prefix, per-vrf, per-CE push occurs and the labeled packet is forwarded back over the 6to4 softwire IPv4-Only core, to the ingress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the ingress CE. . The functionality of the IPv4 forwarding plane in this scenario is identical from a data plane forwarding perspective to Dual Stack IPv4 forwarding scenario.

```
                            +--------+   +--------+
                            |  IPv4  |   |  IPv4  |
                            | Client |   | Client |
                            | Network|   | Network|
                            +--------+   +--------+
                                |   \     /   |
                                |    \   /    |
                                |     \ /     |
                                |      X      |
                                |     / \     |
                                |    /   \    |
                                |   /     \   |
                            +--------+   +--------+
                            |  AFBR  |   |  AFBR  |
                        +--| IPv4/6 |---| IPv4/6 |--+
                        |  +--------+   +--------+  |
          +--------+    |                           |     +--------+
          |  IPv4  |    |                           |     |  IPv4  |
          | Client |    |                           |     | Client |
          | Network|------|            IPv4           |-------| Network|
          +--------+    |            only           |     +--------+
                        |                           |
                        |  +--------+   +--------+  |
                        +--|  AFBR  |---|  AFBR  |--+
                           | IPv4/6 |   | IPv4/6 |
                           +--------+   +--------+
                              |   \     /   |
                              |    \   /    |
                              |     \ /     |
                              |      X      |
                              |     / \     |
                              |    /   \    |
                              |   /     \   |
                            +--------+   +--------+
                            |  IPv6  |   |  IPv4  |
                            | Client |   | Client |
                            | Network|   | Network|
                            +--------+   +--------+
```

Figure 5: 6to4 Softwire - IPv6 Edge over an IPv4-Only Core

4.3.3. 4to6 Softwire IPv6-Only Core packet walk

4to6 softwire where IPv6-Edge eBGP IPv6 peering where IPv4 packets at network Edge traverse a IPv6-Only Core

In the scenario where IPv4 packets originating from a PE-CE edge are tunneled over an MPLS or Segment Routing IPv4 underlay core network, the PE and CE only have an IPv6 address configured on the interface. In this scenario the IPv4 packets that ingress the CE from within the CE AS are over an IPv6-Only interface and are forwarded to an IPv4 NLRI destination prefix learned from the Pure Transport Single IPv6 BGP Peer. In the IPv6-Only Edge peering architecture the PE is IPv6-Only as all PE-CE interfaces are IPv6-Only. However, on the CE, the PE-CE interface is the only interface that is IPv6-Only and all other interfaces may or may not be IPv6-Only. Following the data plane packet flow, IPv4 packets are forwarded from the ingress CE to the IPv6-Only ingress PE where the VPN label imposition push per prefix, per-vrf, per-CE occurs and the labeled packet is forwarded over a 4to6 softwire IPv6-Only core, to the egress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the egress CE. In the reverse direction IPv4 packets are forwarded from the egress CE to egress PE where the VPN label imposition per prefix, per-vrf, per-CE push occurs and the labeled packet is forwarded back over the 4to6 softwire IPv6-Only core, to the ingress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the ingress CE. . The functionality of the IPv4 forwarding plane in this scenario is identical from a data plane forwarding perspective to Dual Stack IPv4 forwarding scenario.

```
                          +--------+   +--------+
                          |  IPv4  |   |  IPv4  |
                          | Client |   | Client |
                          | Network|   | Network|
                          +--------+   +--------+
                              |   \     /   |
                              |    \   /    |
                              |     \ /     |
                              |      X      |
                              |     / \     |
                              |    /   \    |
                              |   /     \   |
                          +--------+   +--------+
                          |  AFBR  |   |  AFBR  |
                       +--| IPv4/6 |---| IPv4/6 |--+
                       |  +--------+   +--------+  |
          +--------+   |                           |   +--------+
          |  IPv6  |   |                           |   |  IPv6  |
          | Client |   |                           |   | Client |
          | Network|------            IPv6           |-------| Network|
          +--------+   |            only           |   +--------+
                       |                           |
                       |  +--------+   +--------+  |
                       +--|  AFBR  |---|  AFBR  |--+
                          | IPv4/6 |   | IPv4/6 |
                          +--------+   +--------+
                             |   \     /   |
                             |    \   /    |
                             |     \ /     |
                             |      X      |
                             |     / \     |
                             |    /   \    |
                             |   /     \   |
                           +--------+   +--------+
                           |  IPv4  |   |  IPv4  |
                           | Client |   | Client |
                           | Network|   | Network|
                           +--------+   +--------+
```

Figure 6: 4to6 Softwire - IPv4 Edge over an IPv6-Only Core

4.4. RFC5549 and RFC8950 Applicability

4.4.1. IPv6-Only Edge Peering design next-hop encoding

This section describes [RFC8950] next hop encoding updates to [RFC5549] applicability to this specification. IPv6-only eBGP Edge PE-CE peering to carry IPv4 Unicast NLRI <AFI/SAFI> IPv4 <1/1> over an IPv6 next hop BGP capability extended hop encoding IANA capability codepoint value 5 defined is applicable to both [RFC5549] and [RFC8950] as IPv4 Unicast NLRI <AFI/SAFI> IPv4 <1/1> does not change in the RFC updates.

IPv4 packets over an IPv6-Only core 4to6 Softwire E2E packet flow is part of the IPv6-Only design vendor interoperaiblity test cases and in that respect is applicable as [RFC8950] updates [RFC5549] for <AFI/SAFI> VPN-IPV4 <1/128>, and Multicasat VPN <1/129>

4.4.2. RFC8950 updates to RFC5549 applicability

This section describes the [RFC8950] next hop encoding updates to [RFC5549]

In [RFC5549] when AFI/SAFI 1/128 is used, the next-hop address is encoded as an IPv6 address with a length of 16 or 32 bytes. This document modifies how the next-hop address is encoded to accommodate all existing implementations and bring consistency with VPNv4oIPv4 and VPNv6oIPv6. The next-hop address is now encoded as a VPN-IPv6 address with a length of 24 or 48 bytes [RFC8950] (see Sections 3 and 6.2 of this document). This change addresses Erratum ID 5253 (Err5253). As all known and deployed implementations are interoperable today and use the new proposed encoding, the change does not break existing interoperability. Updates to [RFC8950] is applicable to the IPv6-Only PE-CE edge design for the IPv6 next hop encoding E2E test case of IPv4 packets over and IPv6-Only core 4to6 Softwire. In this test case IPv4 Unicast NLRI <AFI/SAFI> IPv4 <1/1> is advertised over the PE to RR core peering 4to6 softwire in <AFI/SAFI> VPN-IPV4 <1/128>. In this test case label allocation mode comes into play which is discussed in section 8.9.

[RFC5549] next hop encoding of MP_REACH_NLRI with:

* NLRI= NLRI as per current AFI/SAFI definition

Advertising with [RFC4760] MP_REACH_NLRI with:

* AFI = 1

* SAFI = 128 or 129

* Length of Next Hop Address = 16 or 32

* NLRI= NLRI as per current AFI/SAFI definition

[RFC8950] next hop encoding of MP_REACH_NLRI with:

* NLRI= NLRI as per current AFI/SAFI definition

Advertising with [RFC4760] MP_REACH_NLRI with:

* AFI = 1

* SAFI = 128 or 129

* Length of Next Hop Address = 24 or 48

* Next Hop Address = VPN-IPv6 address of next hop with an 8-octet RD set to zero (potentially followed by the link-local VPN-IPv6 address of the next hop with an 8-octet RD is set to zero).

* NLRI= NLRI as per current AFI/SAFI definition

5. IPv6-Only PE Design Edge and Inter-AS Options E2E Test Cases

Proof of conept interoperability testing of the 4 test cases between the 5 vendors Cisco, Juniper, Arista, Nokia and Huawei.

Cisco, Juniper, Arista, Nokia, Huawei, platform, code revision and test results for all use cases

Cisco: Edge Router- XR ASR 9910 IOS XR 7.4.1, Core Router- NCS 6000 7.2.2, CRS-X 6.7.4

Juniper: Edge Router- MX platform MX480, MX960, Core Router- PTX Platform PTX5000, PTC10K8 (JUNOS and EVO) Release 20.4R2

Tested v4 edge over v6 core in a virtual setup using vMX platforrm and 20.4R2 and LDPv6 as underlay, but there were some data plane forwarding issues. Tested same setup on latest release 21.4 and it worked. Investigating what the minimum version is for this setup to work.

Tested on above Juniper platforms. Completed IPv6-Only PE design functionality test with PE-CE IPv6 peer carrying IPv4 and IPv6 prefixes control plane validation and data plane forwarding plane validation and verified end to end reachability CE to CE forwarding plane with Default Per-CE label allocation mode. Tested with IPv4-Only Core and IPv6-Only Core and proved that the IPv6-Only PE design solution works. Both IPv4 and IPv6 packets were forwarded identical functionality of Dual Stack without having IPv4 address configured.

Nokia: Edge and Core-7750 Service Router, Release R21

Huawei: Edge and Core-VRPv8, Release VRP-V800R020C10

Arista:

Intra-AS tests PE-CE Edge Peering IPv4-Only Core, IPv6-Only Core, Global Table (GRT) and IP VPN

AFI/SAFI IPv4-Unicast SAFI IPv6-Unicast SAFI

IPv4 Core:

Test-1 Global table (6PE)

Test-2 IP VPN

Global table IPv6

IPv6 Core:

Test-3 Global table

Test-4 IP VPN

Inter-AS Options tests IPv4-Only Core, IPv6-Only Core, Global Table (GRT) and IP VPN

AFI/SAFI VPN and MVPN

IPv4-Only Core

Test-5 Global table 6PE Option-B (Segmented LSP stitched IPv4 Core - Inter-AS Link IPv6-Only PE - IPv4 Core)

Test-6 Global table 6PE Option-C (Redistribute IPv4 Loopbacks into BGP-LU AFI/SAFI 2/6)

Test-7 IP VPN Inter AS Option-B (Segmented LSP stitched IPv4 Core - Inter-AS Link IPv6-Only PE - IPv4 Core)

Test-8 IP VPN Inter AS Option-C (Redistribute IPv4 Loopbacks into BGP-LU AFI/SAFI 2/6)

IPv6-Only Core

Test-9 Global table Option-B

Test-10 Global table Option-C

Test-11 IP VPN Inter AS Option-B

Test-12 IP VPN Inter AS Option-C

5.1. Test-1 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core(6PE), 6to4 softwire

```
                                 ________
   IPv6-Only        _____     /        \                IPv6-Only
    PE / CE        /     \__/          \___             PE / CE
+----+  +----+   /                        \        +------+   +-----+
|    |  |    |  |                          |_      |      |   |     |
|    |  |    |  |                            \     |      |   |     |
| CE |--| PE |--\         IPv4-Only Core      |----|  PE  |---|  CE |
|    |  |    |    \0=========Underlay =======0|    |      |   |     |
+----+  +----+     \                         __/    +------+   +-----+
IPv6 BGP peer       \     MPLS / SR domain /         IPv6 BGP peer
IPv4 forwarding      \__          __        /         IPv4 forwarding
IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 7: Test-1 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core (6PE)

5.2. Test-2 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core, 6to4 Softwire

```
                                  ________
      IPv6-Only       _____     /        \                IPv6-Only
       PE / CE       /     \__/          \___             PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  | 0====VPN Overlay Tunnel ==0|      |      |   |     |
   |    |  |    |  |                            \      |      |   |     |
   | CE |--| PE |--\         IPv4-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|      |      |   |     |
   +----+  +----+     \                         __/    +------+   +-----+
   IPv6 BGP peer       \   MPLS / SR domain   /         IPv6 BGP peer
   IPv4 forwarding      \__          __       /          IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 8: Test-2 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core

5.3. Test-3 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core, 4to6 Softwire

```
                                  ________
      IPv6-Only       _____     /        \                IPv6-Only
       PE / CE       /     \__/          \___             PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  |                          |_      |      |   |     |
   |    |  |    |  |                            \      |      |   |     |
   | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|      |      |   |     |
   +----+  +----+     \                         __/    +------+   +-----+
   IPv6 BGP peer       \     MPLS / SR domain /         IPv6 BGP peer
   IPv4 forwarding      \__          __       /          IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 9: Test-3 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core

5.4. Test-4 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core, 4to6 Softwire

```
                                  ________
      IPv6-Only       _____    /        \                IPv6-Only
       PE / CE       /     \__/          \___            PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  | 0====VPN Overlay Tunnel ==0|     |      |   |     |
   |    |  |    |  |                            \     |      |   |     |
   | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|    |      |   |     |
   +----+  +----+     \                        __/    +------+   +-----+
   IPv6 BGP peer       \    MPLS / SR domain  /         IPv6 BGP peer
   IPv4 forwarding      \__          __       /          IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 10: Test-4 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core

5.5. Test-5 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core(6PE), 6to4 softwire -Inter-AS Option-B

```
                   Inter-AS ASBR-ASBR link is IPv6-Only PE
     IPv6-Only       __________          __________      IPv6-Only
      PE / CE       /          \        /          \      PE / CE
     +--+ +----+   /            \      /            \    +--+ +--+
     |  | |    |  |    AS 1      \    |    AS 2      \   |  | |  |
     |  | |    |  |               \IPv6|              \  |  | |  |
     |CE|-| PE |--| IPv4-Only Core|----|IPv4-Only Core|--|PE|-|CE|
     |  | |    |  |0=Underlay==0  |    |0==Underlay==0|  |  | |  |
     +--+ +----+   \              /     \              /   +--+ +--+
     IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
     IPv4 forwarding \_________/         \_________/    IPv4 forwarding
     IPv6 forwarding                                    IPv6 forwarding
```

Figure 11: Test-5 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core (6PE) - Inter-AS Option-B

5.6. Test-6 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core(6PE), 6to4 softwire -Inter-AS Option-C

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
IPv6-Only       __________          __________      IPv6-Only
 PE / CE       /          \        /          \      PE / CE
+--+ +----+   /            \      /            \    +--+ +--+
|  | |    |  |    AS 1      \    |    AS 2     \   |  | |  |
|  | |    |  |               \IPv6|              \  |  | |  |
|CE|-| PE |--| IPv4-Only Core|----|IPv4-Only Core|--|PE|-|CE|
|  | |    |  |0=Underlay==0  |    |0==Underlay==0|  |  | |  |
+--+ +----+   \             /     \             /   +--+ +--+
IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
IPv4 forwarding \_________/         \_________/    IPv4 forwarding
IPv6 forwarding                                    IPv6 forwarding
```

Figure 12: Test-6 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core (6PE) - Inter-AS Option-C

5.7. Test-7 E2E IPv6-Only PE-CE, VPN over IPv4-Only, 6to4 softwire - Inter-AS Option-B

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
IPv6-Only       __________          __________      IPv6-Only
 PE / CE       /          \        /          \      PE / CE
+--+ +----+   /            \      /            \    +--+ +--+
|  | |    |  |    AS 1      \    |    AS 2     \   |  | |  |
|  | |    |  |               \IPv6|              \  |  | |  |
|CE|-| PE |--| IPv4-Only Core|----|IPv4-Only Core|--|PE|-|CE|
|  | |    |  |0=Overlay===0  |    |0==Overlay===0|  |  | |  |
+--+ +----+   \             /     \             /   +--+ +--+
IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
IPv4 forwarding \_________/         \_________/    IPv4 forwarding
IPv6 forwarding                                    IPv6 forwarding
```

Figure 13: Test-7 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core - Inter-AS Option-B

5.8. Test-8 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core, 6to4 softwire -Inter-AS Option-C

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
     IPv6-Only       __________          __________      IPv6-Only
      PE / CE       /          \        /          \      PE / CE
     +--+ +----+   /            \      /            \    +--+ +--+
     |  | |    |  |    AS 1      \    |    AS 2      \   |  | |  |
     |  | |    |  |               \IPv6|              \  |  | |  |
     |CE|-| PE |--| IPv4-Only Core|----|IPv4-Only Core|--|PE|-|CE|
     |  | |    |  |0=Overlay===0  |    |0==Overlay===0|  |  | |  |
     +--+ +----+   \             /     \             /   +--+ +--+
     IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
     IPv4 forwarding \_________/         \_________/    IPv4 forwarding
     IPv6 forwarding                                    IPv6 forwarding
```

Figure 14: Test-8 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core - Inter-AS Option-C

5.9. Test-9 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core, 4to6 softwire -Inter-AS Option-B

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
     IPv6-Only       __________          __________      IPv6-Only
      PE / CE       /          \        /          \      PE / CE
     +--+ +----+   /            \      /            \    +--+ +--+
     |  | |    |  |    AS 1      \    |    AS 2      \   |  | |  |
     |  | |    |  |               \IPv6|              \  |  | |  |
     |CE|-| PE |--| IPv6-Only Core|----|IPv6-Only Core|--|PE|-|CE|
     |  | |    |  |0=Underlay==0  |    |0==Underlay==0|  |  | |  |
     +--+ +----+   \             /     \             /   +--+ +--+
     IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
     IPv4 forwarding \_________/         \_________/    IPv4 forwarding
     IPv6 forwarding                                    IPv6 forwarding
```

Figure 15: Test-9 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core - Inter- AS Option-B

5.10. Test-10 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core, 4to6 softwire -Inter-AS Option-C

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
     IPv6-Only       __________          __________      IPv6-Only
      PE / CE       /          \        /          \      PE / CE
     +--+ +----+   /            \      /            \    +--+ +--+
     |  | |    |  |    AS 1      \    |    AS 2     \   |  | |  |
     |  | |    |  |               \IPv6|              \  |  | |  |
     |CE|-| PE |--| IPv6-Only Core|--- |IPv6-Only Core|--|PE|-|CE|
     |  | |    |  |0=Underlay==0  |    |0==Underlay==0|  |  | |  |
     +--+ +----+   \             /     \             /   +--+ +--+
     IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
     IPv4 forwarding \_________/         \_________/    IPv4 forwarding
     IPv6 forwarding                                    IPv6 forwarding
```

Figure 16: Test-10 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core - Inter-AS Option-C

5.11. Test-11 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core, 4to6 softwire -Inter-AS Option-B

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
     IPv6-Only       __________          __________      IPv6-Only
      PE / CE       /          \        /          \      PE / CE
     +--+ +----+   /            \      /            \    +--+ +--+
     |  | |    |  |    AS 1      \    |    AS 2     \   |  | |  |
     |  | |    |  |               \IPv6|              \  |  | |  |
     |CE|-| PE |--| IPv6-Only Core|--- |IPv6-Only Core|--|PE|-|CE|
     |  | |    |  |0=Overlay===0  |    |0==Overlay===0|  |  | |  |
     +--+ +----+   \             /     \             /   +--+ +--+
     IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
     IPv4 forwarding \_________/         \_________/    IPv4 forwarding
     IPv6 forwarding                                    IPv6 forwarding
```

Figure 17: Test-11 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core - Inter-AS Option-B

5.12. Test-12 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core, 4to6 softwire -Inter-AS Option-C

```
                 Inter-AS ASBR-ASBR link is IPv6-Only PE
     IPv6-Only       __________          __________      IPv6-Only
      PE / CE       /          \        /          \      PE / CE
     +--+ +----+   /            \      /            \    +--+ +--+
     |  | |    |  |    AS 1      \    |    AS 2     \   |  | |  |
     |  | |    |  |               \IPv6|              \  |  | |  |
     |CE|-| PE |--| IPv6-Only Core|--- |IPv6-Only Core|--|PE|-|CE|
     |  | |    |  |0=Overlay===0  |    |0==Overlay===0|  |  | |  |
     +--+ +----+   \             /     \             /   +--+ +--+
     IPv6 BGP peer  \ MPLS/SR   /       \ MPLS/SR   /   IPv6 BGP peer
     IPv4 forwarding \_________/         \_________/    IPv4 forwarding
     IPv6 forwarding                                    IPv6 forwarding
```

Figure 18: Test-12 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core - Inter-AS Option-C

5.13. IPv6-Only PE-CE Operational Considerations Testing

Ping CE to PE when destination prefix is withdrawn
Traceroute CE to PE and test all ICMPv4 and ICMPv6 type co
des

```
                 +-------+                +-------+
                 |       |  IPv6 Only     |       |
                 |   CE  |----------------|  PE   |
                 |       |  IPv6 BGP Peer |       |
                 +-------+                +-------+
                IPv4 forwarding           IPv4 forwarding
                IPv6 forwarding           IPv6 forwarding
```

Figure 19: Ping and Trace Test Case

6. Operational Considerations

With a single IPv6 Peer carrying both IPv4 and IPv6 NLRI there are some operational considerations in terms of what changes and what does not change.

What does not change with a single IPv6 transport peer carrying IPv4 NLRI and IPv6 NLRI below:

Routing Policy configuration is still separate for IPv4 and IPv6 configured by capability as previously.

Layer 1, Layer 2 issues such as one-way fiber or fiber cut will impact both IPv4 and IPv6 as previously.

If the interface is in the Admin Down state, the IPv6 peer would go down, and IPv4 NLRI and IPv6 NLRI would be withdrawn as previously.

Changes resulting from a single IPv6 transport peer carrying IPv4 NLRI and IPv6 NLRI below:

Physical interface is no longer dual stacked.

Any change in IPv6 address or DAD state will impact both IPv4 and IPv6 NLRI exchange.

Single BFD session for both IPv4 and IPv6 NLRI fate sharing as the session is now tied to the transport, which now is only IPv6 address family.

Both IPv4 and IPv6 peer now exists under the IPv6 address family configuration.

Fate sharing of IPv4 and IPv6 address family from a logical perspective now carried over a single physical IPv6 peer.

From an operations perspective, prior to elimination of IPv4 peers, an audit is recommended to identify and IPv4 and IPv6 peering incongruencies that may exist and to rectify them. No operational impacts or issues are expected with this change.

With MPLS VPN overlay, per-CE next-hop label allcoation mode where both IPv4 and IPv6 prefixes have the same label in no table lookup pop-n-forward mode should be taken into consideration.

7. IANA Considerations

There are not any IANA considerations.

8. Security Considerations

The extensions defined in this document allow BGP to propagate reachability information about IPv4 prefixes over an MPLS or SR IPv6-Only core network. As such, no new security issues are raised beyond those that already exist in BGP-4 and the use of MP-BGP for IPv6. Both IPv4 and IPv6 peers exist under the IPv6 address family configuration. The security features of BGP and corresponding security policy defined in the ISP domain are applicable. For the inter-AS distribution of IPv6 routes according to case (a) of Section 4 of this document, no new security issues are raised beyond

those that already exist in the use of eBGP for IPv6 [RFC2545].

9. Acknowledgments

Thanks to Kaliraj Vairavakkalai, Linda Dunbar, Aijun Wang, Eduardfor Vasilenko, Joel Harlpern, Michael McBride, Ketan Talaulikar for review comments.

10. Contributors

The following people contributed substantive text to this document:

Mohana Sundari
EMail: mohanas@juniper.net

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

11.2. Informative References

[I-D.ietf-idr-dynamic-cap] Chen, E. and S. R. Sangli, "Dynamic Capability for BGP-4", Work in Progress, Internet-Draft, draft-ietf-idr-dynamic-cap-16, 21 October 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-dynamic-cap-16>.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, DOI 10.17487/RFC4798, February 2007, <https://www.rfc-editor.org/info/rfc4798>.

[RFC4925] Li, X., Ed., Dawkins, S., Ed., Ward, D., Ed., and A. Durand, Ed., "Softwire Problem Statement", RFC 4925, DOI 10.17487/RFC4925, July 2007, <https://www.rfc-editor.org/info/rfc4925>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/info/rfc5549>.

[RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, DOI 10.17487/RFC5565, June 2009, <https://www.rfc-editor.org/info/rfc5565>.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <https://www.rfc-editor.org/info/rfc6074>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8950] Litkowski, S., Agrawal, S., Ananthamurthy, K., and K. Patel, "Advertising IPv4 Network Layer Reachability Information (NLRI) with an IPv6 Next Hop", RFC 8950, DOI 10.17487/RFC8950, November 2020, <https://www.rfc-editor.org/info/rfc8950>.

Authors' Addresses

Gyan Mishra
Verizon Inc.
Email: gyan.s.mishra@verizon.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS
Email: mankamis@cisco.com

Jeff Tantsura
Microsoft, Inc.
Email: jefftant.ietf@gmail.com

Sudha Madhavi
Juniper Networks, Inc.
Email: smadhavi@juniper.net

Qing Yang
Arista Networks
Email: qyang@arista.com

Adam Simpson
Nokia
Email: adam.1.simpson@nokia.com

Shuanglong Chen
Huawei Technologies
Email: chenshuanglong@huawei.com

Spring
Internet-Draft
Intended status: Standards Track
Expires: 4 September 2022

M. Jethanandani, Ed.
T. Murakami
Arrcus, Inc
3 March 2022

A YANG Model for SRv6 Mobile User Plane
draft-mahesh-spring-srv6-mobile-yang-00

Abstract

This document defines a YANG data model for configuration and management of SRv6 for the mobile network.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 4 September 2022.

Copyright Notice

Table of Contents

1. Introduction

In mobile networks, mobility systems provide connectivity over a wireless link to stationary and non-stationary nodes. The user-plane establishes a tunnel between the mobile node and its anchor node over IP-based backhaul and core networks.

When SRv6 is applied to mobile networks, it enables a source routing architecture, where operators get to explicitly specify a route for the packets to traverse both to and from a mobile node. The SRv6 Endpoint nodes serve as mobile user-plane anchors.

For example, in an Enhanced mode topology, the intermediate waypoints, SIDs, can be used for Traffic Engineering. For more details, see Segment Routing IPv6 for Mobile User Plane [I-D.ietf-dmm-srv6-mobile-uplane], Section 5.2. The gNB and UPF are SR-aware, and there are two service segments, one for traffic engineering to support a low latency path, and the other for service programming. In such a topology the operator routes the traffic through these SRv6 nodes, so they can perform their Endpoint functionality and forward the packet. Further, in the uplink direction, when the gNB receives a packet from a UE, it adds the segments of the SR policy to route the traffic through those two segments, while doing something similar in the downlink direction.

This document describes a YANG 1.1 [RFC7950] data model for the Segment Routing IPv6 (SRv6) user plane of mobile networks.

The model conforms to the NMDA [RFC8342] architecture.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

This document references terms defined in other documents. In particular, it imports definitions for the following terms from Segment Routing Architecture [RFC8402], and IPv6 Segment Routing Header (SRH) [RFC8754].

* Active Segment

* BGP-Prefix Segment

* Prefix SID

* Segment

* SID

* SRH

* SRv6

* SRv6 Endpoint nodes

* SRv6 SID

* Segment Routing domain (SR domain)

* SR Global Block (SRGB)

* SR Local Block (SRLB)

2.1. Acronyms

This document uses a few acronyms. Some of them are defined here for reference.

Acronym	Definition
gNB	gNodeB, a 5G Base Station using New Radio technology
MUP	Mobile User Plane
SR	Segment Routing
SRv6	Segment Routing over v6
UE	User Equipment
UPF	User Plane Function

Table 1: Acronyms

3. Tree Diagram

An abridged version of the tree diagram is shown here. Annotations used in the diagram are defined in YANG Tree Diagrams [RFC8340].

```
module: ietf-srv6-mobile

  augment /rt:routing/rt:control-plane-protocols
            /rt:control-plane-protocol/bgp:bgp/bgp:global:
    +--rw route-distinguisher?     rt-types:route-distinguisher
    +--rw label-allocation-mode?   identityref
    +--rw sid-allocation-mode?     identityref
    +--rw srv6
       +--rw locator?                 leafref
       +--ro sid-manager-connected?   boolean
       +--ro locator-registered?      boolean
       +--ro micro-segment-enabled?   boolean
       +--rw mobile
          +--rw encapsulation
          |     ...
          +--rw decapsulations
          |     ...
          +--rw decapsulation-source-prefix?   inet:ipv6-prefix
  augment /rt:routing/rt:control-plane-protocols
            /rt:control-plane-protocol/bgp:bgp/bgp:global
            /bgp:route-selection-options:
    +--rw selection-deferral-time?   uint16
    +--rw med-missing-as-worst?      boolean
    +--rw multipath-as-path-relax?   boolean
    +--rw multipath-nexthop-relax?   boolean
  augment /rt-pol:routing-policy/rt-pol:defined-sets
            /bp:bgp-defined-sets:
    +--rw n4-interface-sets
       +--rw interface-set* [name]
          +--rw name      string
          +--rw member*   identityref
  augment /rt-pol:routing-policy/rt-pol:policy-definitions
            /rt-pol:policy-definition/rt-pol:statements
            /rt-pol:statement/rt-pol:conditions/bp:bgp-conditions:
    +--rw match-n4-network-interface-set
       +--rw n4-network-instance-set?   leafref
       +--rw match-set-options?         match-set-options-type
  augment /rt-pol:routing-policy/rt-pol:policy-definitions
            /rt-pol:policy-definition/rt-pol:statements
            /rt-pol:statement/rt-pol:actions/bp:bgp-actions:
    +--rw set-network-interface
       +--rw apply-policy
          +--rw import-policy*           leafref
          +--rw default-import-policy?   default-policy-type
          +--rw export-policy*           leafref
          +--rw default-export-policy?   default-policy-type
```

Figure 1: Tree Diagram for SRv6 YANG Model

4. YANG Model

The YANG model is divided into two parts. The first part of the model augments the BGP model in BGP Model for Service Provider Network [I-D.ietf-idr-bgp-model] for the BGP configuration, while the second part augments the BGP Routing Policy model in BGP Model for Service Provider Network [I-D.ietf-idr-bgp-model].

The BGP model is augmented both at a global level to add SRv6 configuration, and at the route selection option. The BGP policy model is augemented to add a defined set, a set of match options, and a set of actions.

The model imports Common YANG Data Types [RFC6991], A YANG Data Model for Routing Management(NMDA Version) [RFC8349], A YANG Data Model for Routing Policy [RFC9067], YANG Data Model for Segment Routing [RFC9020], YANG Data Model for SRv6 Base and Static [I-D.ietf-spring-srv6-yang], and BGP Model for Service Provider Network [I-D.ietf-idr-bgp-model].

```
<CODE BEGINS> file "ietf-srv6-mobile@2022-03-03.yang"
module ietf-srv6-mobile {
  yang-version "1.1";
  namespace "urn:ietf:params:xml:ns:yang:ietf-srv6-mobile";
  prefix "srv6-mob";

  import ietf-inet-types {
    prefix "inet";
    reference
      "RFC 6991: Common YANG Data Types.";
  }
  import ietf-routing {
    prefix rt;
    reference
      "RFC 8349, A YANG Data Model for Routing Management
       (NMDA Version).";
  }
  import ietf-routing-types {
    prefix rt-types;
    reference
      "RFC 8294: Common YANG Types for the Routing Area.";
  }
  import ietf-routing-policy {
    prefix rt-pol;
    reference
      "RFC 9067: A YANG Data Model for Routing Policy.";
  }
  import ietf-bgp {
```

```
      prefix bgp;
      reference
        "I-D.ietf-idr-bgp-model: BGP Model for Service Provider
         Network.";
    }
    import ietf-bgp-policy {
      prefix bp;
      reference
        "I-D.ietf-idr-bgp-model: BGP Model for Service Provider
         Network.";
    }
    import ietf-bgp-types {
      prefix bt;
      reference
        "I-D.ietf-idr-bgp-model: BGP Model for Service Provider
         Network.";
    }
    import ietf-segment-routing {
      prefix sr;
      reference
        "RFC 9020: YANG Data Model for Segment Routing.";
    }
    import ietf-srv6-base {
      prefix srv6;
      reference
        "I-D.ietf-spring-srv6-yang: YANG Data Model for SRv6 Base
         and Static.";
    }
    import ietf-srv6-types {
      prefix srv6-types;
      reference
        "RFC 9020: YANG Data Model for Segment Routing.";
    }

    organization
      "IETF SPRING Working Group";

    contact
      "WG Web:   <https://datatracker.ietf.org/wg/spring/about>
       WG List:  <spring@ietf.org>

       Editor: Mahesh Jethanandani (mjethanandani at gmail dot com)
       Author:  Tetsuya Murakami (tetsuya at arrcus dot com)";

    description
      "This module augments the BGP YANG model to add support for
       configuration in mobile networks.
```

```
       Copyright (c) 2022 IETF Trust and the persons identified as
       authors of the code.  All rights reserved.

       Redistribution and use in source and binary forms, with or
       without modification, is permitted pursuant to, and subject to
       the license terms contained in, the Simplified BSD License set
       forth in Section 4.c of the IETF Trust's Legal Provisions
       Relating to IETF Documents
       (https://trustee.ietf.org/license-info).

       This version of this YANG module is part of RFC XXXX
       (https://www.rfc-editor.org/info/rfcXXXX); see the RFC itself
       for full legal notices.

       The key words 'MUST', 'MUST NOT', 'REQUIRED', 'SHALL', 'SHALL
       NOT', 'SHOULD', 'SHOULD NOT', 'RECOMMENDED', 'NOT RECOMMENDED',
       'MAY', and 'OPTIONAL' in this document are to be interpreted as
       described in BCP 14 (RFC 2119) (RFC 8174) when, and only when,
       they appear in all capitals, as shown here.";

    revision "2022-03-03" {
      description
        "Initial Version.";
      reference
        "RFC XXXX, A YANG Model for BGP configuration in mobile
         networks.";
    }

   /*
    * Features
    */

    /*
     * Typedefs
     */
    typedef srv6-mobile-function-type {
      type union {
        type identityref {
          base "srv6-types:srv6-endpoint-type";
        }
        type uint16;
      }
      description
        "Type definition for SRV6 mobile function. Unknown
         values are represented as the numeric value.";
      reference
        "draft-ietf-dmm-srv6-mobile-uplane";
    }
```

```
/*
 * Identities
 */
identity srv6-mup {
  base bt:afi-safi-type;
  description
    "Segment Routing for Mobile User Plane (AFI,SAFI = ?,?)";
  reference
    "RFC XXXX: A YANG Model for BGP configuration in mobile
     networks.";
}

identity label-allocation-mode {
  description
    "Base identity to be used to express types of label allocation
     strategies to be used within a network instance";
}

identity label-per-prefix {
  base label-allocation-mode;
  description
    "A label is to be allocated per prefix entry in the RIB for the
     network instance";
}

identity label-per-nexthop {
  base label-allocation-mode;
  description
    "A label is to be allocated per nexthop entry in the RIB for
     the network instance";
}

identity label-instance {
  base label-allocation-mode;
  description
    "A single label is to be used for the instance";
}

identity sid-allocation-mode {
  description
    "Base identity to be used to express types of SRv6 segment ID
     allocation strategies to be used within a network instance.";
}

identity sid-per-nexthop {
  base sid-allocation-mode;
  description
    "A segment ID is to be allocated per nexthop entry in the RIB
```

```
         for the network instance";
    }

    identity sid-instance {
      base sid-allocation-mode;
      description
        "A single segment ID is to be used for the instance";
    }

    identity sid-per-nexthop-no-transportation {
      base sid-allocation-mode;
      description
        "A segment ID is to be allocated per nexthop entry in the
         RIB for the network instance";
    }

    identity sid-instance-no-transportation {
      base sid-allocation-mode;
      description
        "A single segment ID is to be used for the instance";
    }

    identity mobile-interface-type {
      description
        "Base identity for different mobile interfaces.";
    }

    identity n4 {
      base mobile-interface-type;
      description
        "N4 interface.";
    }

    /*
     * Groupings
     */

    /*
     * BGP configuration
     */
    augment "/rt:routing/rt:control-plane-protocols" +
            "/rt:control-plane-protocol/bgp:bgp/bgp:global" {

      description
        "Augmentation of the BGP global configuration to add srv6
         mobile configuration.";

      leaf route-distinguisher {
```

```
        type rt-types:route-distinguisher;
        description
          "The route distinguisher that should be used for the local
           VRF or VSI instance when it is signalled via BGP.";
      }

      leaf label-allocation-mode {
        type identityref {
          base label-allocation-mode;
        }
        must "not(../sid-allocation-mode)" {
          error-message "label-allocation-mode and sid-allocation-mode"
                      + "cannot co-exist";
        }
        must "not(/rt:routing/rt:control-plane-protocols" +
             "/rt:control-plane-protocol" +
             "/bgp:bgp/bgp:global/srv6/mobile/encapsulation/config" +
             "/locator | " +
             "/rt:routing/rt:control-plane-protocols" +
             "/rt:control-plane-protocol/bgp:bgp/bgp:global" +
             "/srv6/locator)" {
          error-message "SRv6 configurations must be removed first";
        }
        description
          "The label allocation mode to be used for L3 entries
           in the network instance";
      }

      leaf sid-allocation-mode {
        type identityref {
          base sid-allocation-mode;
        }
        must "not(../label-allocation-mode)" {
          error-message "label-allocation-mode and sid-allocation-mode "
                      + "cannot co-exist";
        }
        must "boolean(/rt:routing/rt:control-plane-protocols" +
             "/rt:control-plane-protocol" +
             "/bgp:bgp/bgp:global/srv6/mobile/encapsulation/locator |" +
             "/rt:routing/rt:control-plane-protocols" +
             "/rt:control-plane-protocol/bgp:bgp/bgp:global/srv6" +
             "/mobile/decapsulations/decapsulation/locator |" +
             "/rt:routing/rt:control-plane-protocols" +
             "/rt:control-plane-protocol/bgp:bgp/bgp:global/srv6" +
             "/locator)" {
          error-message "SRv6 locator name must be configured";
        }
        description
```

```
          "The segment ID allocation mode to be used for L3 entries
           in the network instance";
      }

      container srv6 {
        description
          "SRv6 mobile container.";

        leaf locator {
          type leafref {
            path "/rt:routing/sr:segment-routing/" +
                 "srv6:srv6/srv6:locators/srv6:locator/srv6:name";
          }
          description
            "Locator configuration.";
        }

        leaf sid-manager-connected {
          type boolean;
          config false;
          description
            "Connection with segment ID manager is active";
        }

        leaf locator-registered {
          type boolean;
          config false;
          description
            "Locator name is registered";
        }

        leaf micro-segment-enabled {
          type boolean;
          config false;
          description
            "Locator has enabled micro-segment behavior";
        }

        container mobile {
          when "derived-from-or-self(/rt:routing" +
               "/rt:control-plane-protocols" +
               "/rt:control-plane-protocol/bgp:bgp/bgp:global" +
               "/bgp:afi-safis/bgp:afi-safi/bgp:name, 'srv6-mup')" {
            description
              "This augmentation is valid only for a MUP SAFI.";
          }

          description
```

```
            "Mobile configuration of SRv6.";

          container encapsulation {
            description
              "Encapsulation configuration.";

            leaf locator {
              type leafref {
                path "/rt:routing/sr:segment-routing/" +
                     "srv6:srv6/srv6:locators/srv6:locator/srv6:name";
              }
              description
                "Reference to SRv6 locater key";
            }

            leaf function {
              type srv6-mobile-function-type;
              must "boolean(current()/../locator)" {
                error-message
                  "SRv6 Mobile Locator name must be configured";
              }
              description
                "One of the SRv6 function types.";
            }

            leaf source-adress {
              type inet:ipv4-address;
              description
                "GTP source IP address";
            }

            leaf source-position {
              type uint8;
              description
                "Bit position of GTP source IP address";
            }

            container n4-network-instance {
              description
                "Definitions for the N4 interface.";

              leaf routing-policy {
                type leafref {
                  path "/rt-pol:routing-policy/" +
                       "rt-pol:policy-definitions/" +
                       "rt-pol:policy-definition/rt-pol:name";
                }
                must "boolean(current()/../access)" {
```

```
                error-message
                  "SRv6 Mobile access instance name must be " +
                  "configured";
              }
              description
                "Reference to routing-policy";
            }

            leaf access {
              type string;
              description
                "Mobile access instance.";
            }
          }
        }

        container decapsulations {
          description
            "SRv6 mobile decapsulation configuration.";

          list decapsulation {
            key "id";
            description
              "SRv6 mobile Decapsulation config";

            leaf id {
              type uint16;
              description
                "SRv6 mobile decapsulation entry id";
            }

            leaf locator {
              type leafref {
              path "/rt:routing/sr:segment-routing/" +
                   "srv6:srv6/srv6:locators/srv6:locator/srv6:name";
              }
              description
                "Reference to SRv6 locater key";
            }

            leaf function {
              type srv6-mobile-function-type;
              must "boolean(current()/../locator)" {
                error-message
                  "SRv6 Mobile Locator name must be configured";
              }
              description
                "One of SRv6 function types.";
```

```
              }

              container n4-network-instance {
                description
                  "Definitions for the N4 interface.";

                leaf core {
                  type string;
                  description
                    "Core instance";
                }
              }
            }
          }
          leaf decapsulation-source-prefix {
            type inet:ipv6-prefix;
            description
              "IPv6 prefix for GTP source address";
          }
        }
      }
    }

    augment "/rt:routing/rt:control-plane-protocols" +
            "/rt:control-plane-protocol/bgp:bgp/bgp:global" +
            "/bgp:route-selection-options" {

      description
        "Augmentation of the BGP global configuration for
         route selection options to add srv6 mobile configuration.";

      leaf selection-deferral-time {
        type uint16 {
          range 1..3600;
        }
        default 300;
        description
          "An upper-bound on the time (in seconds) that the best-path
           selection is deferred";
      }

      leaf med-missing-as-worst {
        type boolean;
        description
          "A route without MED is treated as with highest MED value";
      }

      leaf multipath-as-path-relax {
```

```
        type boolean;
        default true;
        description
          "Paths with different AS-Path but of same length can form
           ECMP";
      }

      leaf multipath-nexthop-relax {
        type boolean;
        default false;
        description
          "Enable BGP multi-path for paths with same next-hop";
      }
    }

    augment "/rt-pol:routing-policy/rt-pol:defined-sets" +
            "/bp:bgp-defined-sets" {
      description
        "Augmentation of the Routing Policy module to add
         mobile interface defined sets.";

      container n4-interface-sets {
        description
          "Enclosing container for list of n4 interface sets.";

        list interface-set {
          key "name";
          description
            "List of defined interface sets.";

          leaf name {
            type string;
            description
              "Name of interface set. This is used to reference
               the set in match conditions.";
          }

          leaf-list member {
            type identityref {
              base "mobile-interface-type";
            }
            description
              "Members of interface set.";
          }
        }
      }
    }
```

```
    augment "/rt-pol:routing-policy/rt-pol:policy-definitions" +
            "/rt-pol:policy-definition/rt-pol:statements" +
            "/rt-pol:statement/rt-pol:conditions/bp:bgp-conditions" {
      description
        "Augmentation of the Routing Policy module to add conditions.";

      container match-n4-network-interface-set {
        description
          "Match a referenced network instance.";
        leaf n4-network-instance-set {
          type leafref {
            path "/rt-pol:routing-policy/rt-pol:defined-sets/"
               + "bp:bgp-defined-sets/n4-interface-sets/"
               + "interface-set/name";
          }
          description
            "References a defined community set.";
        }
        uses rt-pol:match-set-options-group;
      }
    }

    augment "/rt-pol:routing-policy/rt-pol:policy-definitions" +
            "/rt-pol:policy-definition/rt-pol:statements" +
            "/rt-pol:statement/rt-pol:actions/bp:bgp-actions" {
      description
        "Augmentation of the Routing Policy module to add actions.";

      container set-network-interface {
        description
          "Set a referenced network instance.";
        uses rt-pol:apply-policy-group;
      }
    }
  }
  <CODE ENDS>
```

Figure 2: SRv6 YANG Model for Mobile User Plane

5. IANA Considerations

This memo registers the following namespace URIs in the IETF XML in the "IETF XML Registry" [RFC3688]:

URI: urn:ietf:params:xml:ns:yang:ietf-srv6-mobile
Registrant Contact: The IESG.
XML: N/A; the requested URI is an XML namespace.

This document registers the following YANG modules in the "YANG Module Names" registry [RFC6020]:

Name: ietf-srv6-mobile
Namespace: urn:ietf:params:xml:ns:yang:ietf-srv6-mobile
Prefix: srv6-mob
Reference: RFC XXXX

6. Security Considerations

The YANG module specified in this document defines a schema for data that is designed to be accessed via network management protocols such as NETCONF [RFC6241] or RESTCONF [RFC8040]. The lowest NETCONF layer is the secure transport layer, and the mandatory-to-implement secure transport is Secure Shell (SSH) [RFC6242]. The lowest RESTCONF layer is HTTPS, and the mandatory-to-implement secure transport is TLS [RFC8446].

The Network Configuration Access Control Model (NACM) [RFC8341] provides the means to restrict access for particular NETCONF or RESTCONF users to a preconfigured subset of all available NETCONF or RESTCONF protocol operations and content.

There are a number of data nodes defined in this YANG module that are writable/creatable/deletable (i.e., config true, which is the default). These data nodes may be considered sensitive or vulnerable in some network environments. Write operations (e.g., edit-config) to these data nodes without proper protection can have a negative effect on network operations. These are the subtrees and data nodes and their sensitivity/vulnerability:

Some of the readable data nodes in this YANG module may be considered sensitive or vulnerable in some network environments. It is thus important to control read access (e.g., via get, get-config, or notification) to these data nodes. These are the subtrees and data nodes and their sensitivity/vulnerability:

Some of the RPC operations in this YANG module may be considered sensitive or vulnerable in some network environments. It is thus important to control access to these operations. These are the operations and their sensitivity/vulnerability:

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3688] Mealling, M., "The IETF XML Registry", BCP 81, RFC 3688, DOI 10.17487/RFC3688, January 2004, <https://www.rfc-editor.org/info/rfc3688>.

[RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <https://www.rfc-editor.org/info/rfc6020>.

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <https://www.rfc-editor.org/info/rfc6241>.

[RFC6242] Wasserman, M., "Using the NETCONF Protocol over Secure Shell (SSH)", RFC 6242, DOI 10.17487/RFC6242, June 2011, <https://www.rfc-editor.org/info/rfc6242>.

[RFC6991] Schoenwaelder, J., Ed., "Common YANG Data Types", RFC 6991, DOI 10.17487/RFC6991, July 2013, <https://www.rfc-editor.org/info/rfc6991>.

[RFC7950] Bjorklund, M., Ed., "The YANG 1.1 Data Modeling Language", RFC 7950, DOI 10.17487/RFC7950, August 2016, <https://www.rfc-editor.org/info/rfc7950>.

[RFC8040] Bierman, A., Bjorklund, M., and K. Watsen, "RESTCONF Protocol", RFC 8040, DOI 10.17487/RFC8040, January 2017, <https://www.rfc-editor.org/info/rfc8040>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8340] Bjorklund, M. and L. Berger, Ed., "YANG Tree Diagrams", BCP 215, RFC 8340, DOI 10.17487/RFC8340, March 2018, <https://www.rfc-editor.org/info/rfc8340>.

[RFC8341] Bierman, A. and M. Bjorklund, "Network Configuration Access Control Model", STD 91, RFC 8341, DOI 10.17487/RFC8341, March 2018, <https://www.rfc-editor.org/info/rfc8341>.

[RFC8342] Bjorklund, M., Schoenwaelder, J., Shafer, P., Watsen, K., and R. Wilton, "Network Management Datastore Architecture (NMDA)", RFC 8342, DOI 10.17487/RFC8342, March 2018, <https://www.rfc-editor.org/info/rfc8342>.

[RFC8349] Lhotka, L., Lindem, A., and Y. Qu, "A YANG Data Model for Routing Management (NMDA Version)", RFC 8349, DOI 10.17487/RFC8349, March 2018, <https://www.rfc-editor.org/info/rfc8349>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <https://www.rfc-editor.org/info/rfc8446>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC9020] Litkowski, S., Qu, Y., Lindem, A., Sarkar, P., and J. Tantsura, "YANG Data Model for Segment Routing", RFC 9020, DOI 10.17487/RFC9020, May 2021, <https://www.rfc-editor.org/info/rfc9020>.

[RFC9067] Qu, Y., Tantsura, J., Lindem, A., and X. Liu, "A YANG Data Model for Routing Policy", RFC 9067, DOI 10.17487/RFC9067, October 2021, <https://www.rfc-editor.org/info/rfc9067>.

[I-D.ietf-idr-bgp-model]
Jethanandani, M., Patel, K., Hares, S., and J. Haas, "BGP YANG Model for Service Provider Networks", Work in Progress, Internet-Draft, draft-ietf-idr-bgp-model-12, 25 October 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-idr-bgp-model-12>.

[I-D.ietf-spring-srv6-yang]
Raza, K., Agarwal, S., Liu, X., Hu, Z., Hussain, I., Shah, H., Voyer, D., Matsushima, S., Horiba, K., Rajamanickam, J., and A. AbdelSalam, "YANG Data Model for SRv6 Base and Static", Work in Progress, Internet-Draft, draft-ietf-spring-srv6-yang-01, 14 January 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-spring-srv6-yang-01>.

7.2. Informative References

[I-D.ietf-dmm-srv6-mobile-uplane]
Matsushima, S., Filsfils, C., Kohno, M., Garvia, P. C., Voyer, D., and C. E. Perkins, "Segment Routing IPv6 for Mobile User Plane", Work in Progress, Internet-Draft, draft-ietf-dmm-srv6-mobile-uplane-18, 18 February 2022, <https://datatracker.ietf.org/doc/html/draft-ietf-dmm-srv6-mobile-uplane-18>.

Appendix A. Appendix 1 Complete Tree Diagram

Here is a complete tree diagram for the configuration and operational part of the model.

```
module: ietf-srv6-mobile

  augment /rt:routing/rt:control-plane-protocols
            /rt:control-plane-protocol/bgp:bgp/bgp:global:
    +--rw route-distinguisher?     rt-types:route-distinguisher
    +--rw label-allocation-mode?   identityref
    +--rw sid-allocation-mode?     identityref
    +--rw srv6
       +--rw locator?                 leafref
       +--ro sid-manager-connected?   boolean
       +--ro locator-registered?      boolean
       +--ro micro-segment-enabled?   boolean
       +--rw mobile
          +--rw encapsulation
          |  +--rw locator?               leafref
          |  +--rw function?              srv6-mobile-function-type
          |  +--rw source-adress?         inet:ipv4-address
          |  +--rw source-position?       uint8
          |  +--rw n4-network-instance
          |     +--rw routing-policy?   leafref
          |     +--rw access?           string
          +--rw decapsulations
          |  +--rw decapsulation* [id]
          |     +--rw id                     uint16
          |     +--rw locator?               leafref
          |     +--rw function?
          |     |       srv6-mobile-function-type
          |     +--rw n4-network-instance
          |        +--rw core?   string
          +--rw decapsulation-source-prefix?   inet:ipv6-prefix
  augment /rt:routing/rt:control-plane-protocols
            /rt:control-plane-protocol/bgp:bgp/bgp:global
            /bgp:route-selection-options:
```

```
       +--rw selection-deferral-time?   uint16
       +--rw med-missing-as-worst?      boolean
       +--rw multipath-as-path-relax?   boolean
       +--rw multipath-nexthop-relax?   boolean
     augment /rt-pol:routing-policy/rt-pol:defined-sets
               /bp:bgp-defined-sets:
       +--rw n4-interface-sets
          +--rw interface-set* [name]
             +--rw name      string
             +--rw member*   identityref
     augment /rt-pol:routing-policy/rt-pol:policy-definitions
               /rt-pol:policy-definition/rt-pol:statements
               /rt-pol:statement/rt-pol:conditions/bp:bgp-conditions:
       +--rw match-n4-network-interface-set
          +--rw n4-network-instance-set?   leafref
          +--rw match-set-options?         match-set-options-type
     augment /rt-pol:routing-policy/rt-pol:policy-definitions
               /rt-pol:policy-definition/rt-pol:statements
               /rt-pol:statement/rt-pol:actions/bp:bgp-actions:
       +--rw set-network-interface
          +--rw apply-policy
             +--rw import-policy*           leafref
             +--rw default-import-policy?   default-policy-type
             +--rw export-policy*           leafref
             +--rw default-export-policy?   default-policy-type
```

Figure 3: Complete tree diagram

Acknowledgements

TBA

Contributors

Thanks to all of the contributors.

Authors' Addresses

Mahesh Jethanandani (editor)
Arrcus, Inc
Email: mjethanandani@gmail.com

Tetsuya Murakami
Arrcus, Inc
Email: tetsuya@arrcus.com

BESS Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: 28 August 2022

G. Mishra
Verizon Inc.
M. Mishra
Cisco Systems
J. Tantsura
Microsoft, Inc.
S. Madhavi
Juniper Networks, Inc.
Q. Yang
Arista Networks
A. Simpson
Nokia
S. Chen
Huawei Technologies
24 February 2022

IPv6-Only PE Design for IPv4-NLRI All SAFI over IPv6-NH
draft-mishra-bess-ipv4nlri-all-safi-ipv6nh-01

Abstract

As Enterprises and Service Providers upgrade their brown field or green field MPLS/SR core to an IPv6 transport, Multiprotocol BGP (MP-BGP)now plays an important role in the transition of their Provider (P) core network as well as Provider Edge (PE) Inter-AS peering network from IPv4 to IPv6. Operators must be able to continue to support IPv4 customers when both the Core and Edge networks are IPv6-Only.

This document details an important External BGP (eBGP) PE-PE Inter-AS IPv6-Only peering design that leverages the MP-BGP capability exchange by using IPv6 peering as pure transport, allowing both IPv4 Network Layer Reachability Information (NLRI) and IPv6 Network Layer Reachability Information (NLRI)to be carried over the same (Border Gateway Protocol) BGP TCP session for all Address Family Identifiers (AFI) and Subsequent Address Family Identifiers(SAFI). The design change provides the same Dual Stacking functionality that exists today with separate IPv4 and IPv6 BGP sessions as we have today. With this design change from a control plane perspective a single IPv6-Only peer is required for both IPv4 and IPv6 routing updates and from a data plane forwarindg perspective an IPv6 address need only be configured on the PE to PE Inter-AS peering interface for both IPv4 and IPv6 packet forwarding. This document extends the IPv6-Only PE-CE peering architecture defined in [I-D.ietf-bess-ipv6-only-pe-design] to PE-PE inter-as peering architecture where the 4to6 softwire is now extended to Inter-AS L3 VPN options Option-A, Option-AB and Option-C and now applies to all

AFI/SAFI ubiquitously. As service providers migrate to Segment Routing architecture SR-MPLS and SRv6, VPN overlay exsits as well, and thus Inter-AS options Option-A, Option-AB and Option-C are still applicable and thus this extension of IPv6-Only peering architecure extension to Inter-AS peering is very relevant to Segment Routing as well.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

As Enterprises and Service Providers upgrade their brown field or green field MPLS/SR core to an IPv6 transport such as MPLS LDPv6, SR-MPLSv6 or SRv6, Multiprotocol BGP (MP-BGP) now plays an important role in the transition of the Provider (P) core networks and Provider Edge (PE) edge networks from IPv4 to IPv6. Operators have a requirement to support IPv4 customers and must be able to support IPv4 address family and Sub-Address-Family Virtual Private Network (VPN)-IPv4, and Multicast VPN IPv4 customers.

IXP are also facing IPv4 address depletion at their peering points, which are large Layer 2 transit backbones that service providers peer and exchange IPv4 and IPv6 Network Layer Reachability Information (NLRI). Today, these transit exchange points are Dual Stacked. With this IPv6-only BGP peering design, only IPv6 is configured on the PE-PE inter-as peering interface, the Inter-AS Provider Edge (PE) - Provider Edge (PE), the IPv6 BGP peer is now used to carry IPv4 (Network Layer Reachability Information) NLRI over an IPv6 next hop using IPv6 next hop encoding defined in [RFC8950], while continuing

to forward both IPv4 and IPv6 packets. In the framework of this design the ASBRs providing Inter-AS options peering PE to PE extending L3 VPN services is now no longer Dual Stacked.

MP-BGP specifies that the set of usable next-hop address families is determined by the Address Family Identifier (AFI) and the Subsequent Address Family Identifier (SAFI). Historically the AFI/SAFI definitions for the IPv4 address family only have provisions for advertising a Next Hop address that belongs to the IPv4 protocol when advertising IPv4 or VPN-IPv4. [RFC8950] specifies the extensions necessary to allow advertising IPv4 NLRI, Virtual Private Network Unicast (VPN-IPv4) NLRI, Multicast Virtual Private Network (MVPN-IPv4) NLRI with a Next Hop address that belongs to the IPv6 protocol. This comprises of an extended next hop encoding MP-REACH BGP capability exchange to allow the address of the Next Hop for IPv4 NLRI, VPN-IPv4 NLRI and MVPN-IPv4 NLRI to also belong to the IPv6 Protocol. [RFC8950] defines the encoding of the Next Hop to determine which of the protocols the address actually belongs to, and a new BGP Capability allowing MP-BGP Peers to discover dynamically whether they can exchange IPv4 NLRI and VPN-IPv4 NLRI with an IPv6 Next Hop.

The current specification for carrying IPv4 NLRI of a given address family via a Next Hop of a different address family is now defined in [RFC8950], and specifies the extended next hop encoding MP-REACH capability extension necessary to do so. This comprises an extension of the AFI/SAFI definitions to allow the address of the Next Hop for IPv4 NLRI or VPN-IPv4 NLRI to belong to either the IPv4 or the IPv6 protocol, the encoding of the Next Hop information to determine which of the protocols the address belongs to, and a new BGP Capability allowing MP-BGP peers to dynamically discover whether they can exchange IPv4 NLRI and VPN- IPv4 NLRI with an IPv6 Next Hop.

With the new extensions defined in [RFC8950] supporting NLRI and next hop address family mismatch, the BGP peer session can now be treated as a pure TCP transport and carry both IPv4 and IPv6 NLRI at the Provider Edge (PE) - Customer Edge (CE) over a single IPv6 TCP session. This allows for the elimination of dual stack from the PE-PE Inter-AS peering point, and now enable the Inter-AS peering to be IPv6-ONLY. The elimination of IPv4 Inter Provider ASBR tie point, PE-PE Inter-AS peering points translates into OPEX expenditure savings of point-to-point infrastructure links as well as /31 address space savings and administration and network management of both IPv4 and IPv6 BGP peers. This reduction decreases the number of PE-PE Inter-AS options BGP peers by fifty percent, which is a tremendous cost savings for operators.

While the savings exists at the Edge eBGP PE-PE Inter-AS peering, on the core side PE to Route Reflector (RR) peering carrying <AFI/SAFI> IPv4 <1/1>, VPN-IPV4 <1/128>, and Multicasat VPN <1/129>, there is no savings as the Provider (P) Core is IPv6 Only and thus can only have an IPv6 peer and must use [RFC8950] extended next hop encoding to carrying IPv4 NLRI IPV4 <2/1>, VPN-IPV4 <2/128>, and Multicast VPN <2/129> over an IPv6 next hop.

This document details an important External BGP (eBGP) PE-PE Inter-AS IPv6-Only peering design that leverages the MP-BGP capability exchange by using IPv6 peering as pure transport, allowing both IPv4 Network Layer Reachability Information (NLRI) and IPv6 Network Layer Reachability Information (NLRI)to be carried over the same (Border Gateway Protocol) BGP TCP session for all Address Family Identifiers (AFI) and Subsequent Address Family Identifiers(SAFI). The design change provides the same Dual Stacking functionality that exists today with separate IPv4 and IPv6 BGP sessions as we have today. With this design change from a control plane perspective a single IPv6-Only peer is required for both IPv4 and IPv6 routing updates and from a data plane forwarindg perspective an IPv6 address need only be configured on the PE to PE Inter-AS peering interface for both IPv4 and IPv6 packet forwarding. This document extends the IPv6-Only PE-CE peering architecture defined in [I-D.ietf-bess-ipv6-only-pe-design] to PE-PE inter-as peering architecture where the 4to6 softwire is now extended to Inter-AS L3 VPN options Option-A, Option-AB and Option-C and now applies to all AFI/SAFI ubiquitously. As service providers migrate to Segment Routing architecture SR-MPLS and SRv6, VPN overlay exsits as well, and thus Inter-AS options Option-A, Option-AB and Option-C are still applicable and thus this extension of IPv6-Only peering architecure extension to Inter-AS peering is relevant to Segment Routing.

This document details an important External BGP (eBGP) PE-PE Inter-AS IPv6-Only peering design that leverages the MP-BGP capability exchange by using IPv6 peering as pure transport, allowing both IPv4 Network Layer Reachability Information (NLRI) and IPv6 Network Layer Reachability Information (NLRI)to be carried over the same (Border Gateway Protocol) BGP TCP session for the following Address Family Identifiers (AFI) and Subsequent Address Family Identifiers(SAFI) to be carried over IPv6-Only Inter-AS peerings described in detail in this document: <AFI/SAFI> IPv4 Unicast <1/1>, IPv4 Multicast <1/2>,VPN-IPV4 <1/128>, Multicasat VPN <1/129>, BGP-LU IPV4 (4PE) <1/4>, BGP-LU IPV4 <1/4>

This document details an important External BGP (eBGP) PE-PE Inter-AS IPv6-Only peering design that leverages the MP-BGP capability exchange by using IPv6 peering as pure transport, allowing both IPv4 Network Layer Reachability Information (NLRI) and IPv6 Network Layer

Reachability Information (NLRI)to be carried over the same (Border Gateway Protocol) BGP TCP session for all remaining Address Family Identifiers (AFI) and Subsequent Address Family Identifiers(SAFI) below as well that can be carried over IPv6-Only Inter-AS peerings: <AFI/SAFI> MCAST-VPN [RFC6514] <1/5>, NLRI Multi-Segment Pseudowires [RFC7267] <1/6>, BGP Tunnel Encapsulation SAFI [RFC9012] <1/7>, MCAST-VPLS [RFC7117] <1/8>, BGP SFC [RFC9015] <1/9>, Tunnel SAFI [I-D.nalawade-kapoor-tunnel-safi] <1/6>, Virtual Private LAN Service (VPLS) [RFC4761] and [RFC6074] <1/5>, BGP MDT SAFI [RFC6037] <1/66>, BGP 4to6 SAFI [RFC5747] <1/67>, BGP 6to4 SAFI draft xx <1/8>, Layer 1 VPN Auto-Discovery [RFC5195] <1/69>, BGP EVPNs [RFC7432] <1/70>, BGP-LS (VPLS) [RFC7752] <1/71>, BGP-LS-EVPN [RFC7752] <72/>, SR-TE Policy SAFI draftxx <1/73>, BGP 6to4 SAFI draft xx <1/8>, SDN WAN Capabilities draftxx <1/74>, Routing Policy SAFI draftxx <1/75>, Classful-Transport SAFI draftxx <1/76>, Tunneled Traffic FlowSpec draftxx <1/77>, MCAST-TREE SAFI draft xx <1/78>, Route Target Constraints [RFC4684] <1/132>, Dissemination of Flow Specification Rules [RFC8955] <1/133>, L3 VPN Dissemination of Flow Specification Rules [RFC8955] <1/1344>, VPN Auto-Discovery SAFI draftxx <1/140>

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

Terminolgoy used in defining the IPv6-Only Edge specification.

AFBR: Address Family Border Router Provider Edge (PE).

Edge: PE-CE Edge Network Provider Edge - Customer Edge

Core: P Core Network Provider (P)

4to6 Softwire : IPv4 edge over an IPv6-Only core

6to4 Softwire: IPv6 edge over an IPv4-Only core

E2E: End to End

4. IPv6-Only Edge Peering Architecture

4.1. Problem Statement

This specification addresses a real issue that has been discussed at many operator groups around the world related to IXP major peering points where hundreds of AS's have both IPv4 and IPv6 dual stacked peering. IPv4 address depletion have been a major issue issue for many years now. Operators around the world are clamoring for a solution that can help solve issues related to IPv4 address depletion at these large IXP peering points. With this solution IXPs as well as all infrastructure networks such as Core networks, DC networks, Access networks as well as any PE-CE public or private network can now utilize this IPv6-Only Edge solution and reap the benefits immediately on IPv4 address space saving.

```
                          IXP Problem Statement

                     Dual Stacked            Dual Stacked
                          CE                      PE

                       +-------+ IPv4 BGP Peer +-------+
                       |       |---------------|       |
                       |   CE  | IPv6 BGP Peer |  PE   |
                       |       |---------------|       |
                       +-------+               +-------+
                      IPv4 forwarding           IPv4 forwarding
                      IPv6 forwarding           IPv6 forwarding
```

Figure 1: Problem Statement - IXP Dual Stack Peering

```
                                 ________
   Dual Stacked     _____     /        \                Dual Stacked
     PE / CE       /     \__/          \___              PE / CE
 +----+  +----+   /                      \        +------+   +-----+
 |    |  |    |  |0====VPN Overlay Tunnel ==0|      |      |   |     |
 |    |  |    |  |                          \      |      |   |     |
 | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
 |    |  |    |    \0=========Underlay =======0|    |      |   |     |
 +----+  +----+     \                        __/    +------+   +-----+
 IPv4 IPv6 BGP peer  \ IP / MPLS / SR domain /     IPv4 and IPv6 BGP peer
 IPv4 forwarding      \__         __        /        IPv4 forwarding
 IPv6 forwarding         \_______/  \_____/          IPv6 forwarding
```

Figure 2: Problem Statement - E2E Dual Stack Edge

4.2. IPv6-Only PE-CE Design Solution

The IPv6-Only Edge design solution provides a means of E2E single protocol design solution extension of [RFC5565] Softwire Mesh framework from the PE-CE Edge to the Core from ingres so egress through the entire operators domain. This solution eliminates all IPv4 addressing from end to end while still providing the same Dual Stack functionality of IPv4 and IPv6 packet forwarding from a data plane perspective by leveraging the [RFC8950] extended next hop encoding so that IPv4 NLRI can be advertised over a single IPv6 pure transport TCP session. This IPv6-Only E2E architecture eliminates all IPv4 peering and IPv4 addressing E2E from the ingress CE to ingress PE to egress PE to egress CE and all hops along the operator E2E path.

```
                    Solution applicable to
          any Edge peering scenario - IXP, Core, DC, Access, etc

           +-------+                +-------+
           |       |  IPv6 Only     |       |
           |   CE  |----------------|  PE   |
           |       |  IPv6 BGP Peer |       |
           +-------+                +-------+
          IPv4 forwarding            IPv4 forwarding
          IPv6 forwarding            IPv6 forwarding
```

Figure 3: IPv6-Only Solution Applicability

```
                                 ________
   IPv6-Only       _____     /        \                 IPv6-Only
    PE / CE       /     \__/          \___              PE / CE
+----+  +----+   /                      \        +------+   +-----+
|    |  |    |  |0====VPN Overlay Tunnel ==0|      |      |   |     |
|    |  |    |  |                          \      |      |   |     |
| CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
|    |  |    |    \0=========Underlay ===== ==0    |      |   |     |
+----+  +----+     \                         __/    +------+   +-----+
IPv6 BGP peer        \IP / MPLS / SR domain /        IPv6 BGP peer
IPv4 forwarding      \__         __       /          IPv4 forwarding
IPv6 forwarding         \_______/  \_____/           IPv6 forwarding
```

Figure 4: E2E VPN Solution

4.3. IPv6-Only Edge Peering Design

4.3.1. IPv6-Only Edge Peering Packet Walk

The IPv6-Only Edge Peering design utilizes two key E2E Softwire Mesh Framework scenario's, 4to6 softwire and 6to4 softwire. The Softwire mesh framework concept is based on the overlay and underlay MPLS or SR based technology framework, where the underlay is the transport layer and the overlay is a Virtual Private Network (VPN) layer, and is the the tunneled virtualization layer containing the customer payload. The concept of a 6to4 Softwire is based on transmission of IPv6 packets at the edge of the network by tunneling the IPv6 packets over an IPv4-Only Core. The concept of a 4to6 Softwire is also based on transmission of IPv4 packets at the edge of the network by tunneling the IPv4 packets over an IPv6-Only Core.

This document describes End to End (E2E) test scenarios that follow a packet flow from IPv6-Only attachment circuit from ingress PE-CE to egress PE-CE tracing the routing protocol control plane and data plane forwarding of IPv4 packets in a 4to6 softwire or 6to4 softwire within the IPv4-Only or IPv6-Only Core network. In both secneario we are focusing on IPv4 packets and the control plane and data plane forwarding aspects of IPv4 packets from the PE-CE Edge network over an IPv6-Only P (Provider) core network or IPv4-Only P (Provider) core network. With this IPv6-Only Edge peering design, the Softwire Mesh Framework is not extended beyond the Provider Edge (PE) and continues to terminate on the PE router.

4.3.2. 6to4 Softwire IPv4-Only Core packet walk

6to4 softwire where IPv6-Edge eBGP IPv6 peering where IPv4 packets at network Edge traverse a IPv4-Only Core

In the scenario where IPv4 packets originating from a PE-CE edge are tunneled over an MPLS or Segment Routing IPv4 underlay core network, the PE and CE only have an IPv6 address configured on the interface. In this scenario the IPv4 packets that ingress the CE from within the CE AS are over an IPv6-Only interface and are forwarded to an IPv4 NLRI destination prefix learned from the Pure Transport Single IPv6 BGP Peer. In the IPv6-Only Edge peering architecture the PE is IPv6-Only as all PE-CE interfaces are IPv6-Only. However, on the CE, the PE-CE interface is the only interface that is IPv6-Only and all other interfaces may or may not be IPv6-Only. Following the data plane packet flow, IPv4 packets are forwarded from the ingress CE to the IPv6-Only ingress PE where the VPN label imposition push per prefix, per-vrf, per-CE occurs and the labeled packet is forwarded

over a 6to4 softwire IPv4-Only core, to the egress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the egress CE. In the reverse direction IPv4 packets are forwarded from the egress CE to egress PE where the VPN label imposition per prefix, per-vrf, per-CE push occurs and the labeled packet is forwarded back over the 6to4 softwire IPv4-Only core, to the ingress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the ingress CE. . The functionality of the IPv4 forwarding plane in this scenario is identical from a data plane forwarding perspective to Dual Stack IPv4 forwarding scenario.

```
                              +--------+   +--------+
                              |  IPv4  |   |  IPv4  |
                              | Client |   | Client |
                              | Network|   | Network|
                              +--------+   +--------+
                                  |   \     /   |
                                  |    \   /    |
                                  |     \ /     |
                                  |      X      |
                                  |     / \     |
                                  |    /   \    |
                                  |   /     \   |
                              +--------+   +--------+
                              |  AFBR  |   |  AFBR  |
                           +--| IPv4/6 |---| IPv4/6 |--+
                           |  +--------+   +--------+  |
          +--------+       |                           |       +--------+
          |  IPv4  |       |                           |       |  IPv4  |
          | Client |       |                           |       | Client |
          | Network|------ |            IPv4           |------- | Network|
          +--------+       |            only           |       +--------+
                           |                           |
                           |  +--------+   +--------+  |
                           +--|  AFBR  |---|  AFBR  |--+
                              | IPv4/6 |   | IPv4/6 |
                              +--------+   +--------+
                                |   \     /   |
                                |    \   /    |
                                |     \ /     |
                                |      X      |
                                |     / \     |
                                |    /   \    |
                                |   /     \   |
                             +--------+   +--------+
                             |  IPv6  |   |  IPv4  |
                             | Client |   | Client |
                             | Network|   | Network|
                             +--------+   +--------+
```

Figure 5: 6to4 Softwire - IPv6 Edge over an IPv4-Only Core

4.3.3. 4to6 Softwire IPv6-Only Core packet walk

4to6 softwire where IPv6-Edge eBGP IPv6 peering where IPv4 packets at network Edge traverse a IPv6-Only Core

In the scenario where IPv4 packets originating from a PE-CE edge are tunneled over an MPLS or Segment Routing IPv4 underlay core network, the PE and CE only have an IPv6 address configured on the interface. In this scenario the IPv4 packets that ingress the CE from within the CE AS are over an IPv6-Only interface and are forwarded to an IPv4 NLRI destination prefix learned from the Pure Transport Single IPv6 BGP Peer. In the IPv6-Only Edge peering architecture the PE is IPv6-Only as all PE-CE interfaces are IPv6-Only. However, on the CE, the PE-CE interface is the only interface that is IPv6-Only and all other interfaces may or may not be IPv6-Only. Following the data plane packet flow, IPv4 packets are forwarded from the ingress CE to the IPv6-Only ingress PE where the VPN label imposition push per prefix, per-vrf, per-CE occurs and the labeled packet is forwarded over a 4to6 softwire IPv6-Only core, to the egress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the egress CE. In the reverse direction IPv4 packets are forwarded from the egress CE to egress PE where the VPN label imposition per prefix, per-vrf, per-CE push occurs and the labeled packet is forwarded back over the 4to6 softwire IPv6-Only core, to the ingress PE where the VPN label disposition pop occurs and the native IPv4 packet is forwarded to the ingress CE. . The functionality of the IPv4 forwarding plane in this scenario is identical from a data plane forwarding perspective to Dual Stack IPv4 forwarding scenario.

```
                          +--------+   +--------+
                          |  IPv4  |   |  IPv4  |
                          | Client |   | Client |
                          | Network|   | Network|
                          +--------+   +--------+
                              |   \     /   |
                              |    \   /    |
                              |     \ /     |
                              |      X      |
                              |     / \     |
                              |    /   \    |
                              |   /     \   |
                          +--------+   +--------+
                          |  AFBR  |   |  AFBR  |
                       +--| IPv4/6 |---| IPv4/6 |--+
                       |  +--------+   +--------+  |
          +--------+   |                           |   +--------+
          |  IPv6  |   |                           |   |  IPv6  |
          | Client |   |                           |   | Client |
          | Network|------            IPv6           |-------| Network|
          +--------+   |            only           |   +--------+
                       |                           |
                       |  +--------+   +--------+  |
                       +--|  AFBR  |---|  AFBR  |--+
                          | IPv4/6 |   | IPv4/6 |
                          +--------+   +--------+
                            |   \     /   |
                            |    \   /    |
                            |     \ /     |
                            |      X      |
                            |     / \     |
                            |    /   \    |
                            |   /     \   |
                          +--------+   +--------+
                          |  IPv4  |   |  IPv4  |
                          | Client |   | Client |
                          | Network|   | Network|
                          +--------+   +--------+
```

Figure 6: 4to6 Softwire - IPv4 Edge over an IPv6-Only Core

4.4. RFC5549 and RFC8950 Applicability

4.4.1. IPv6-Only Edge Peering design next-hop encoding

This section describes [RFC8950] next hop encoding updates to [RFC5549] applicability to this specification. IPv6-only eBGP Edge PE-CE peering to carry IPv4 Unicast NLRI <AFI/SAFI> IPv4 <1/1> over an IPv6 next hop BGP capability extended hop encoding IANA capability codepoint value 5 defined is applicable to both [RFC5549] and [RFC8950] as IPv4 Unicast NLRI <AFI/SAFI> IPv4 <1/1> does not change in the RFC updates.

IPv4 packets over an IPv6-Only core 4to6 Softwire E2E packet flow is part of the IPv6-Only design vendor interoperaiblity test cases and in that respect is applicable as [RFC8950] updates [RFC5549] for <AFI/SAFI> VPN-IPV4 <1/128>, and Multicasat VPN <1/129>

4.4.2. RFC8950 updates to RFC5549 applicability

This section describes the [RFC8950] next hop encoding updates to [RFC5549]

In [RFC5549] when AFI/SAFI 1/128 is used, the next-hop address is encoded as an IPv6 address with a length of 16 or 32 bytes. This document modifies how the next-hop address is encoded to accommodate all existing implementations and bring consistency with VPNv4oIPv4 and VPNv6oIPv6. The next-hop address is now encoded as a VPN-IPv6 address with a length of 24 or 48 bytes [RFC8950] (see Sections 3 and 6.2 of this document). This change addresses Erratum ID 5253 (Err5253). As all known and deployed implementations are interoperable today and use the new proposed encoding, the change does not break existing interoperability. Updates to [RFC8950] is applicable to the IPv6-Only PE-CE edge design for the IPv6 next hop encoding E2E test case of IPv4 packets over and IPv6-Only core 4to6 Softwire. In this test case IPv4 Unicast NLRI <AFI/SAFI> IPv4 <1/1> is advertised over the PE to RR core peering 4to6 softwire in <AFI/SAFI> VPN-IPV4 <1/128>. In this test case label allocation mode comes into play which is discussed in section 8.9.

[RFC5549] next hop encoding of MP_REACH_NLRI with:

* NLRI= NLRI as per current AFI/SAFI definition

Advertising with [RFC4760] MP_REACH_NLRI with:

* AFI = 1

* SAFI = 128 or 129

* Length of Next Hop Address = 16 or 32

* NLRI= NLRI as per current AFI/SAFI definition

[RFC8950] next hop encoding of MP_REACH_NLRI with:

* NLRI= NLRI as per current AFI/SAFI definition

Advertising with [RFC4760] MP_REACH_NLRI with:

* AFI = 1

* SAFI = 128 or 129

* Length of Next Hop Address = 24 or 48

* Next Hop Address = VPN-IPv6 address of next hop with an 8-octet RD set to zero (potentially followed by the link-local VPN-IPv6 address of the next hop with an 8-octet RD is set to zero).

* NLRI= NLRI as per current AFI/SAFI definition

5. IPv6-Only PE Design Edge E2E Design for all AFI/SAFI

Proof of conept interoperability testing of the 4 test cases bet

5.1. Design Solution-1 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core(6PE), 6to4 softwire

```
                                 ________
      IPv6-Only       _____     /        \                IPv6-Only
       PE / CE       /     \__/          \___             PE / CE
   +----+  +----+   /                        \         +------+   +-----+
   |    |  |    |  |                          |_       |      |   |     |
   |    |  |    |  |                            \      |      |   |     |
   | CE |--| PE |--\         IPv4-Only Core      |----|  PE  |---|  CE |
   |    |  |    |     \0=========Underlay =======0|    |      |   |     |
   +----+  +----+      \                        __/    +------+   +-----+
   IPv6 BGP peer       \     MPLS / SR domain /         IPv6 BGP peer
   IPv4 forwarding      \__          __       /          IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/           IPv6 forwarding
```

Figure 7: Design Solution-1 E2E IPv6-Only PE-CE, Global Table over IPv4-Only Core (6PE)

Cisco, Juniper, Arista, Nokia, Huawei code and platform and test results.

Cisco: Edge Router- XR ASR 9910 IOS XR 7.4.1, Core Router- NCS 6000 7.2.2, CRS-X 6.7.4

Juniper: Edge Router- MX platform MX480, MX960, Core Router- PTX Platform PTX5000, PTC10K8 (JUNOS and EVO) Release 20.4R2

Tested v4 edge over v6 core in a virtual setup using vMX platforrm and 20.4R2 and LDPv6 as underlay, but there were some data plane forwarding issues. Tested same setup on latest release 21.4 and it worked. Investigating what the minimum version is for this setup to work.

Arista:

Nokia: Edge and Core-7750 Service Router, Release R21

Huawei: Edge and Core-VRPv8, Release VRP-V800R020C10

5.2. Design Solution-2 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core, 6to4 Softwire

```
                                  ________
      IPv6-Only       _____    /        \                IPv6-Only
       PE / CE       /     \__/          \___             PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  | 0====VPN Overlay Tunnel ==0|     |      |   |     |
   |    |  |    |  |                            \     |      |   |     |
   | CE |--| PE |--\         IPv4-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|     |      |   |     |
   +----+  +----+     \                        __/    +------+   +-----+
   IPv6 BGP peer       \   MPLS / SR domain   /         IPv6 BGP peer
   IPv4 forwarding      \__          __       /          IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 8: Design Solution-2 E2E IPv6-Only PE-CE, VPN over IPv4-Only Core

5.3. Design Solution-3 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core (4PE), 4to6 Softwire

```
                                  ________
      IPv6-Only       _____    /        \                IPv6-Only
       PE / CE       /     \__/          \___            PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  |                          |_      |      |   |     |
   |    |  |    |  |                            \     |      |   |     |
   | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|    |      |   |     |
   +----+  +----+     \                         __/    +------+   +-----+
   IPv6 BGP peer       \     MPLS / SR domain /         IPv6 BGP peer
   IPv4 forwarding      \__          __        /         IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 9: Design Solution-3 E2E IPv6-Only PE-CE, Global Table over IPv6-Only Core (4PE)

5.4. Design Solution-4 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core, 4to6 Softwire

```
                                  ________
      IPv6-Only       _____    /        \                IPv6-Only
       PE / CE       /     \__/          \___            PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  | 0====VPN Overlay Tunnel ==0|     |      |   |     |
   |    |  |    |  |                            \     |      |   |     |
   | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|    |      |   |     |
   +----+  +----+     \                         __/    +------+   +-----+
   IPv6 BGP peer       \    MPLS / SR domain  /         IPv6 BGP peer
   IPv4 forwarding      \__          __        /         IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 10: Design Solution-4 E2E IPv6-Only PE-CE, VPN over IPv6-Only Core

5.5. Design Solution-5 E2E Inter-AS Option B and AB, 4to6 Softwire

```
                                  ________
      IPv6-Only       _____    /        \                IPv6-Only
       PE / CE       /     \__/          \___             PE / CE
   +----+  +----+   /                        \        +------+   +-----+
   |    |  |    |  | 0====VPN Overlay Tunnel ==0|     |      |   |     |
   |    |  |    |  |                            \     |      |   |     |
   | CE |--| PE |--\         IPv6-Only Core      |----|  PE  |---|  CE |
   |    |  |    |    \0=========Underlay =======0|    |      |   |     |
   +----+  +----+     \                        __/    +------+   +-----+
   IPv6 BGP peer       \    MPLS / SR domain  /         IPv6 BGP peer
   IPv4 forwarding      \__          __       /          IPv4 forwarding
   IPv6 forwarding         \_______/  \_____/            IPv6 forwarding
```

Figure 11: Design Solution-5 E2E E2E Inter-AS Option B and AB
over IPv6-Only Core

5.6. IPv6-Only PE-CE Operational Considerations Testing

Ping CE to PE when destination prefix is withdrawn
Traceroute CE to PE and test all ICMPv4 and ICMPv6 type co
des

```
+-------+                +-------+
|       |  IPv6 Only     |       |
|   CE  |----------------|  PE   |
|       |  IPv6 BGP Peer |       |
+-------+                +-------+
IPv4 forwarding           IPv4 forwarding
IPv6 forwarding           IPv6 forwarding
```

Figure 12: Ping and Trace Test Case

6. IPv6-Only PE ALL AFI/SFI Operational Considerations

With a single IPv6 Peer carrying both IPv4 and IPv6 NLRI there are some operational considerations in terms of what changes and what does not change.

What does not change with a single IPv6 transport peer carrying IPv4 NLRI and IPv6 NLRI below:

Routing Policy configuration is still separate for IPv4 and IPv6 configured by capability as previously.

Layer 1, Layer 2 issues such as one-way fiber or fiber cut will impact both IPv4 and IPv6 as previously.

If the interface is in the Admin Down state, the IPv6 peer would go down, and IPv4 NLRI and IPv6 NLRI would be withdrawn as previously.

Changes resulting from a single IPv6 transport peer carrying IPv4 NLRI and IPv6 NLRI below:

Physical interface is no longer dual stacked.

Any change in IPv6 address or DAD state will impact both IPv4 and IPv6 NLRI exchange.

Single BFD session for both IPv4 and IPv6 NLRI fate sharing as the session is now tied to the transport, which now is only IPv6 address family.

Both IPv4 and IPv6 peer now exists under the IPv6 address family configuration.

Fate sharing of IPv4 and IPv6 address family from a logical perspective now carried over a single physical IPv6 peer.

From an operations perspective, prior to elimination of IPv4 peers, an audit is recommended to identify and IPv4 and IPv6 peering incongruencies that may exist and to rectify them. No operational impacts or issues are expected with this change.

With MPLS VPN overlay, per-CE next-hop label allcoation mode where both IPv4 and IPv6 prefixes have the same label in no table lookup pop-n-forward mode should be taken into consideration.

7. Vendor Implementations and Operator Deployments

Vendor implementations are with Cisco, Juniper, Nokia, Arista and Huawei

8. IANA Considerations

There are not any IANA considerations.

9. Security Considerations

The extensions defined in this document allow BGP to propagate reachability information about IPv4 prefixes over an MPLS or SR IPv6-Only core network. As such, no new security issues are raised beyond those that already exist in BGP-4 and the use of MP-BGP for IPv6. Both IPv4 and IPv6 peers exist under the IPv6 address family configuration. The security features of BGP and corresponding security policy defined in the ISP domain are applicable. For the inter-AS distribution of IPv6 routes according to case (a) of Section 4 of this document, no new security issues are raised beyond those that already exist in the use of eBGP for IPv6 [RFC2545].

10. Acknowledgments

Thanks to Kaliraj Vairavakkalai, Linda Dunbar, Aijun Wang, Eduardfor Vasilenko, Joel Harlpern, Michael McBride, Ketan Talaulikar for review comments.

11. Contributors

The following people contributed substantive text to this document:

Mohana Sundari
EMail: mohanas@juniper.net

12. References

12.1. Normative References

[I-D.ietf-bess-ipv6-only-pe-design]
Mishra, G., Mishra, M., Tantsura, J., Madhavi, S., Yang, Q., Simpson, A., and S. Chen, "IPv6-Only PE Design for IPv4-NLRI with IPv6-NH", Work in Progress, Internet-Draft, draft-ietf-bess-ipv6-only-pe-design-00, 20 September 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-ipv6-only-pe-design-00.txt>.

[I-D.nalawade-kapoor-tunnel-safi]
Nalawade, G., "BGP Tunnel SAFI", Work in Progress, Internet-Draft, draft-nalawade-kapoor-tunnel-safi-05, 29 June 2006, <https://www.ietf.org/archive/id/draft-nalawade-kapoor-tunnel-safi-05.txt>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <https://www.rfc-editor.org/info/rfc2545>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <https://www.rfc-editor.org/info/rfc4761>.

[RFC5195] Ould-Brahim, H., Fedyk, D., and Y. Rekhter, "BGP-Based Auto-Discovery for Layer-1 VPNs", RFC 5195, DOI 10.17487/RFC5195, June 2008, <https://www.rfc-editor.org/info/rfc5195>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC5747] Wu, J., Cui, Y., Li, X., Xu, M., and C. Metz, "4over6 Transit Solution Using IP Encapsulation and MP-BGP Extensions", RFC 5747, DOI 10.17487/RFC5747, March 2010, <https://www.rfc-editor.org/info/rfc5747>.

[RFC6037] Rosen, E., Ed., Cai, Y., Ed., and IJ. Wijnands, "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037, DOI 10.17487/RFC6037, October 2010, <https://www.rfc-editor.org/info/rfc6037>.

[RFC7117] Aggarwal, R., Ed., Kamite, Y., Fang, L., Rekhter, Y., and C. Kodeboniya, "Multicast in Virtual Private LAN Service (VPLS)", RFC 7117, DOI 10.17487/RFC7117, February 2014, <https://www.rfc-editor.org/info/rfc7117>.

[RFC7267] Martini, L., Ed., Bocci, M., Ed., and F. Balus, Ed., "Dynamic Placement of Multi-Segment Pseudowires", RFC 7267, DOI 10.17487/RFC7267, June 2014, <https://www.rfc-editor.org/info/rfc7267>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8277] Rosen, E., "Using BGP to Bind MPLS Labels to Address Prefixes", RFC 8277, DOI 10.17487/RFC8277, October 2017, <https://www.rfc-editor.org/info/rfc8277>.

[RFC8955] Loibl, C., Hares, S., Raszuk, R., McPherson, D., and M. Bacher, "Dissemination of Flow Specification Rules", RFC 8955, DOI 10.17487/RFC8955, December 2020, <https://www.rfc-editor.org/info/rfc8955>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[RFC9015] Farrel, A., Drake, J., Rosen, E., Uttaro, J., and L. Jalil, "BGP Control Plane for the Network Service Header in Service Function Chaining", RFC 9015, DOI 10.17487/RFC9015, June 2021, <https://www.rfc-editor.org/info/rfc9015>.

12.2. Informative References

[I-D.ietf-idr-dynamic-cap]
Chen, E. and S. R. Sangli, "Dynamic Capability for BGP-4", Work in Progress, Internet-Draft, draft-ietf-idr-dynamic-cap-16, 21 October 2021, <https://www.ietf.org/archive/id/draft-ietf-idr-dynamic-cap-16.txt>.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, DOI 10.17487/RFC4798, February 2007, <https://www.rfc-editor.org/info/rfc4798>.

[RFC4925] Li, X., Ed., Dawkins, S., Ed., Ward, D., Ed., and A. Durand, Ed., "Softwire Problem Statement", RFC 4925, DOI 10.17487/RFC4925, July 2007, <https://www.rfc-editor.org/info/rfc4925>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/info/rfc5549>.

[RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, DOI 10.17487/RFC5565, June 2009, <https://www.rfc-editor.org/info/rfc5565>.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <https://www.rfc-editor.org/info/rfc6074>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8950] Litkowski, S., Agrawal, S., Ananthamurthy, K., and K. Patel, "Advertising IPv4 Network Layer Reachability Information (NLRI) with an IPv6 Next Hop", RFC 8950, DOI 10.17487/RFC8950, November 2020, <https://www.rfc-editor.org/info/rfc8950>.

Authors' Addresses

Gyan Mishra
Verizon Inc.
Email: gyan.s.mishra@verizon.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS
Email: mankamis@cisco.com

Jeff Tantsura
Microsoft, Inc.
Email: jefftant.ietf@gmail.com

Sudha Madhavi
Juniper Networks, Inc.
Email: smadhavi@juniper.net

Qing Yang
Arista Networks
Email: qyang@arista.com

Adam Simpson
Nokia
Email: adam.1.simpson@nokia.com

Shuanglong Chen
Huawei Technologies
Email: chenshuanglong@huawei.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: 8 May 2024

T. Murakami
K. Patel
Arrcus, Inc.
S. Matsushima
SoftBank
J. Zhang
Juniper Networks
S. Agrawal
Cisco Systems
D. Voyer
Bell Canada
5 November 2023

BGP Extensions for the Mobile User Plane (MUP) SAFI
draft-mpmz-bess-mup-safi-03

Abstract

This document defines a new SAFI known as a BGP Mobile User Plane (BGP-MUP) SAFI to support MUP Extensions and a extended community for BGP. This document also provides BGP signaling and procedures for the new SAFI to convert mobile session information into appropriate IP forwarding information. These extensions can be used by operators between a PE, and a Controller for integrating mobile user plane into BGP MUP network using the IP based routing.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[I-D.mhkk-dmm-srv6mup-architecture] defines the Segment Routing IPv6 Mobile User Plane (SRv6 MUP) architecture for Distributed Mobility Management. As part of the architecture, the document defines a new SRv6 segment type called as a MUP Segment, new routing information that can carried within BGP, and advertised from a PE and a MUP Controller.

This document defines a new SAFI known as a BGP Mobile User Plane (BGP-MUP) SAFI to support MUP Extensions for BGP. This draft also provides BGP signaling and procedures for the new SAFI to convert mobile session information into appropriate IP routing information. These extensions can be used by operators between the PE and a MUP Controller for integrating mobile user plane into Segment Routing network using the IP based routing information. These extensions also works with routing instances accomodationg two new wellknown segment types known as Interwork and Direct [I-D.mhkk-dmm-srv6mup-architecture]. Finally, the BGP encoding and procedures defined in this document that uses SRv6 as the forwarding fabric follows the SRv6 MUP architecture defined in [I-D.mhkk-dmm-srv6mup-architecture]. The BGP extensions to build networks that use forwarding mechanisms other than SRv6 (SRv6 MUP) are outside the scope of this document.

1.1. Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Terminology

MUP: Mobile User Plane

MUP Segment: Representation of mobile user plane segment

PE: Provider Edge node in an SR network

MUP Controller: Controller node for an SR network

UE: User Equipment, as per [TS.23501]

gNodeB: 3GPP-compliant implementation of 5G-NR base station, as per [TS.23501]

UPF: 3GPP-compliant implementation of User Plane Function, as per [TS.23501]

TEID: Tunnel Endpoint Identifier, as per [TS.29281]

3. BGP MUP Extensions

3.1. BGP MUP SAFI

This draft defines a new BGP SAFI known as a BGP-MUP SAFI. The value of this SAFI is to be assigned by IANA from the Subsequent Address Family Identifiers (SAFI) registry.

This document also defines a new BGP NLRI known as the BGP-MUP NLRI. The following is the format of the BGP-MUP NLRI:

```
+-----------------------------------+
|    Architecture Type (1 octet)    |
+-----------------------------------+
|       Route Type (2 octets)       |
+-----------------------------------+
|         Length (1 octet)          |
+-----------------------------------+
|  Route Type specific (variable)   |
+-----------------------------------+
```

The Architecture Type field defines the encoding of the rest of BGP-MUP NLRI for a given Mobile User Plane architecture. This draft defines the following architecture type for BGP-MUP NLRI:

+ 1 - 3gpp-5g;

The Route Type field defines the encoding of the rest of BGP-MUP NLRI (Route Type specific BGP-MUP NLRI) for a given architecture type.

The Length field indicates the length in octets of the Route Type specific field of the BGP-MUP NLRI.

This document defines the following Route Types for BGP-MUP NLRI.

+ 1 - Interwork Segment Discovery route;
+ 2 - Direct Segment Discovery route;
+ 3 - Type 1 Session Transformed (ST) route;
+ 4 - Type 2 Session Transformed (ST) route;

These Route Types are applicable for the 3gpp-5G architecture type as per [I-D.mhkk-dmm-srv6mup-architecture]. Other new architectures can share them if it is applicable as well.

The BGP-MUP NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an Address Family Identifier (AFI) of 1 or 2 and a Subsequent AFI (SAFI) of BGP-MUP. The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the BGP-MUP NLRI (encoded as specified above). The value of the AFI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute that carries the BGP-MUP NLRI determines whether the addresses carried in the routes are IPv4 or IPv6 addresses (AFI 1 indicates IPv4 addresses, AFI 2 indicates IPv6 addresses).

In order for two BGP speakers to exchange BGP-MUP NLRIs, they must use a BGP Capabilities Advertisement to ensure that they both are capable of properly processing such an NLRI. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of 1 or 2 and an SAFI of BGP-MUP.

This document defines 4 Route Types for 3gpp-5G architecture type. Any other Route Types MUST be silently ignored upon a receipt if a BGP speaker supports only 3gpp-5G architecture type. An implementation MAY log an error when such Route Types are ignored. An implementation MAY considered retrieving any discarded Route Types by simply resetting the session or issuing a Route-REFRESH message [RFC2918] if the Route Refresh Capability is successfully negotiated.

The following sections describes the format of the Route Type specific BGP-MUP NLRI for various Route Types defined in this document.

3.1.1. BGP Interwork Segment Discovery route

A BGP Interwork Segment Discovery route Type specific BGP-MUP NLRI consist of the following:

```
+-----------------------------------+
|           RD  (8 octets)          |
+-----------------------------------+
|       Prefix Length (1 octet)     |
+-----------------------------------+
|        Prefix (variable)          |
+-----------------------------------+
```

The Interwork Segment Discovery route Type NLRI consist of RD which is encoded as described in [RFC4364]. It also has an prefix associated with Interwork segment connected locally. For the purpose of BGP route key processing, only the RD, Prefix Length and Prefix are considered to be part of the prefix in the NLRI.

In 3GPP 5G specific case, a prefix used for a N3 interface of the gNodeB connected locally MAY be used as this prefix. The prefix length of one octet indicating length of the prefix. If the AFI is IPv4, then the maximum value of the Prefix Length is 32 bits otherwise it is considered as a malformed NLRI. If the AFI is IPv6, then the maximum value of of the Prefix length is 128 bits otherwise it is considered as a malformed NLRI. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

3.1.2. BGP Direct Segment Discovery route

A BGP Direct Segment Discovery route Type specific BGP-MUP NLRI consist of the following:

```
+-----------------------------------+
|           RD  (8 octets)          |
+-----------------------------------+
|        Address (4 or 16 octets)   |
+-----------------------------------+
```

The Direct Segment Discovery route Type NLRI consist of RD which is encoded as described in [RFC4364]. It also has an Address of originating BGP speaker. For the purpose of BGP route key processing, only the RD and Address are considered to be part of the prefix in the NLRI.

If the AFI is IPv4 then the address length is 4 octets otherwise it is considered as a malformed NLRI. If the AFI is IPv6 then the address length is 16 octets otherwise it is considered as a malformed NLRI. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

3.1.3. BGP Type 1 Session Transformed (ST) Route

A BGP Type 1 ST Route Type specific BGP-MUP NLRI consist of the following:

```
+-----------------------------------+
|           RD  (8 octets)          |
+-----------------------------------+
|      Prefix Length (1 octet)      |
+-----------------------------------+
|         Prefix (variable)         |
+-----------------------------------+
| Architecture specific (variable)  |
+-----------------------------------+
```

The Type 1 ST Route Type NLRI consist of RD which is encoded as described in [RFC4364]. It also has Prefix length of one octet indicating length of the Prefix. For the purpose of BGP route key processing, only the RD, Prefix Length and Prefix are considered to be part of the prefix in the NLRI.

In 3GPP 5G specific case, Prefix is the prefix allocated to a UE. If the AFI is IPv4, then the maximum value of the Prefix Length field is 32. If the AFI is IPv6, then the maximum value of the Prefix Length field is 128. Any other length field is considered a a malformed NLRI. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

The architecture specific encoding values will follow after the variable length Prefix.

3.1.3.1. 3gpp-5g Specific BGP Type 1 ST Route

A BGP 3gpp-5g Type 1 ST Route Type specific BGP-MUP NLRI consist of the following:

```
+-----------------------------------+
|          TEID (4 octets)          |
+-----------------------------------+
|          QFI (1 octet)            |
+-----------------------------------+
| Endpoint Address Length (1 octet) |
+-----------------------------------+
|    Endpoint Address (variable)    |
+-----------------------------------+
|  Source Address Length (1 octet)  |
+-----------------------------------+
|     Source Address (variable)     |
+-----------------------------------+
```

The TEID has a fixed length of 4 octets. The TEID value of 0 is considered as an invalid and a malformed TEID. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

The QFI has a fixed length of 1 octet.

The Endpoint Address Length has a fixed length of 1 octet. Endpoint Address field contains of an IPv4 address, then the value of the Endpoint Address Length field is 32. If the Endpoint Address field contains of an IPv6 Address, then the value of the Endpoint Address Length field is 128. Any other value is considered as an invalid and a malformed Endpoint Address. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

The Source Address Length has a fixed length of 1 octet. Source Address Length is 0 bytes then the Source address is not carried within the NLRI. A BGP speaker MAY have a local configuration for using a Source address. The exact mechanism of local configuration is outside the scope of this document. Source Address field contains of an IPv4 address, then the value of the Source Address Length field is 32. If the Source Address field contains of an IPv6 Address, then the value of the Source Address Length field is 128. Any other value is considered as an invalid and a malformed Endpoint Address. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

The NLRI architecture field MUST be encoded as shown above if a BGP speaker receives 3gpp-5g specific BGP Type 1 ST route. Otherwise the NLRI is considered as a malformed. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

3.1.4. BGP Type 2 Session Transformed (ST) Route

A BGP Type 2 ST Route Type specific BGP-MUP NLRI consist of the following:

```
+-----------------------------------+
|           RD  (8 octets)          |
+-----------------------------------+
|      Endpoint Length (1 octet)    |
+-----------------------------------+
|      Endpoint Address (variable)  |
+-----------------------------------+
| Architecture specific Endpoint    |
|         Identifier (variable)     |
+-----------------------------------+
```

The Type 2 ST Route Type NLRI consist of RD which is encoded as described in [RFC4364]. It also has Endpoint length of one octet indicating length of the Endpoint Address and the Architecture specific Endpoint Identifier as per [I-D.mhkk-dmm-srv6mup-architecture] defines aggregation capability by the Type2 ST Route. If the AFI is IPv4 and the Endpoint Length is longer than 32 then the Architecture specific endpoint Identifier field exists with the IPv4 Endpoint Address. If the AFI is IPv6 and the Endpoint Length is longer than 128 then the Architecture specific endpoint Identifier field exists with then the IPv6 Endpoint Address. For the purpose of BGP route key processing, only the RD, Endpoint Address and Architecture specific Endpoint Identifier are considered to be part of the prefix in the NLRI.

In 3GPP 5G specific case, the Endpoint Address is a N3 interface address of the UPF. If the AFI is IPv4, then the maximum Endpoint length is 64 otherwise it is considered as a malformed NLRI. If the AFI is IPv6, then the maximum Endpoint length is 160 otherwise it is considered as a malformed NLRI. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

The architecture specific encoding values will follow after the variable length Prefix.

3.1.4.1. 3gpp-5g Specific BGP Type 2 ST Route

A BGP 3gpp-5g Type 2 ST Route Type specific BGP-MUP NLRI consist of the following:

```
+-----------------------------------+
|          TEID (0-4 octets)        |
+-----------------------------------+
```

The maximum length of TEID is 4 octets. The TEID value of 0 is considered as an invalid and a malformed TEID. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

The NLRI architecture field MUST be encoded as shown above if a BGP speaker receives 3gpp-5g specific BGP Type 2 ST route. Otherwise the NLRI is considered as a malformed. A BGP speaker MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606]. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

3.2. BGP MUP Extended Community

This document defines a new BGP Extended community known as BGP MUP Extended community as per [I-D.mhkk-dmm-srv6mup-architecture].

This is a BGP MUP specific Extended Community, of an extended type, and is transitive across AS boundaries [RFC4360].

The Type value of this Extended community is set to MUP type, to be assigned by IANA from BGP Extended community transtive registry. The Sub-Type value is set to Direct-Type Segment Identifier type, to be assigned by IANA from BGP Extended community transtive registry. The value field of the community is set to the 6 bytes of configurable segment identifier value.

The usage of this Extended community is described in the Section 3.3.12 and Section 3.3.10

3.3. Operation

BGP speakers acting as a PE, and a MUP Controller MUST establish a BGP session to exchange BGP-MUP NLRIs for both, IPv4 and IPv6 AFIs. Once the session is established successfully, BGP speakers can exchange the Discovery routes as well as Session Transformed routes. This information is specific to a given routing instance. BGP-MUP SAFI is expected to work with routing instances accomodationg MUP segments. In 3GPP 5G specific case, the routing instances are depicted as N3RAN and N6DN instances defined in [I-D.mhkk-dmm-srv6mup-architecture]. The subsequent sections describes procedures of generating and processing of each route types.

3.3.1. Generation of the Interwork Segment Discovery route

The Interwork Segment Discovery route is generated by the PE when a routing instance accommodates an Interwork type MUP Segment, e.g., N3 interfaces or routes on RAN side in 3GPP 5G specific case. It generates route per each N3RAN IP prefix and stores the route in the routing instance of N3RAN. The IP prefix MAY include a gNodeB address which is connecting to the PE. The BGP AFI for BGP MP_REACH_NLRI attribute to carry the Discovery route is decided based on the AFI of the prefix.

When advertising the Interwork Segment Discovery route, a PE MUST attach the export BGP Route Target Extended Community of the associated routing instance.

When advertising the Interwork Segment Discovery route, a PE MUST use the IPv6 address of the PE as the nexthop address in the MP_REACH_NLRI attribute.

The Interwork Segment Discovery route update MUST have a prefix SID attribute which the SID consists of the PE locater followed by a function. In 3GPP 5G specific case, if the BGP AFI is IPv4, the function MUST be GTP4.E [I-D.ietf-dmm-srv6-mobile-uplane], or MUST be GTP6.E [I-D.ietf-dmm-srv6-mobile-uplane] if the BGP AFI is IPv6.

3.3.2. Withdrawal of the Interwork Segment Discovery route

The Interwork Segment Discovery route is withdrawn by the PE when it detects that the MUP Segment no longer present for the N3RAN. The BGP AFI for BGP MP_UNREACH_NLRI attribute to carry the Discovery route is decided based on the AFI of the prefix.

When withdrawing the Interwork Segment Discovery route, a PE MUST attach the export BGP Route Target Extended Community of the associated routing instance.

3.3.3. Processing of the Interwork Segment Discovery routes

A BGP speaker acting as a PE MAY keep the received MUP Interwork Segment Discovery routes advertised from other PEs. The receiving BGP speaker will perform the importing of the received MUP Interwork Segment Discovery routes in the configured routing instance based on the Route Target extended communities. The IP prefixes for the received segments are imported into the configured routing instance table. Thereby the receiving BGP speaker can provide network connectivity between the nodes that exist in the segments. A BGP speaker MAY discard the received Interwork Segment Discovery route if the speaker fails to import the route based on the Route Target

extended communities.

The BGP speaker receiving the Interwork Segment Discovery routes SHOULD ignore the nexthop in the MP_REACH_NLRI attribute. However, the receiving BGP speaker MUST ensure that the value of Address filed in the NLRI is an address of the originator of the locator value in the prefix SID attribute. The originator of the locator value can be resolved from the IPv6 IGP table. If the result of the match is not identical then the receiving BGP speaker MUST consider it as a malformed NLRI and the "Treat-as-withdraw procedure of [RFC7606] is applied. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

When a BGP speaker receives the Interwork Segment Dicovery routes with a MP_REACH_NLRI attribute without a prefix SID attribute, then it MUST be treated as if it contained a malformed prefix SID attribute and the "Treat-as-withdraw procedure of [RFC7606] is applied. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

When a BGP speaker receives an MP_UNREACH_NLRI attribute update message it MUST delete the withdrawn Interwork Segment Discovery route from the routing instance table where it was created.

3.3.4. Generation of the Direct Segment Discovery route

The Direct Segment Discovery route is generated by the PE when a routing instance accommodates a Direct type MUP Segment, e.g., N6 interface or routes on DN side in 3GPP 5G specific case. It generates the Direct Segment Discovery route per each routing instance for the MUP Segment. The address in the BGP-MUP NLRI MUST be a unique PE identifier. The BGP AFI for BGP MP_REACH_NLRI attribute to carry the Direct Segment Discovery route is decided based on the AFI of the PE identifier [I-D.mhkk-dmm-srv6mup-architecture].

When announcing the Direct Segment Discovery route, a PE MUST attach a BGP MUP Extended community of the associated routing instance. The sub-type of the Extended community is Direct-Type Segment Identifier.

When advertising the Direct Segment Discovery route, a PE MUST use the IPv6 address of the PE as the nexthop address in the MP_REACH_NLRI attribute.

The Direct Segment Discovery route update MUST have a prefix SID attribute which the SID consists of the PE locator followed by a function. The function MAY be End.DT4/6 or End.DX4/6.

3.3.5. Withdrawal of the Direct Segment Discovery route

The Direct Segment Discovery route is withdrawn by the PE when it detects that the MUP Segment no longer present for the routing instance. The BGP AFI for BGP MP_UNREACH_NLRI attribute to carry the Discovery route is decided based on the AFI of the PE identifier.

When withdrawing the Direct Segment Discovery route, a BGP speaker MUST attach a BGP MUP Extended community of the associated routing instance.

3.3.6. Processing of the Direct Segment Discovery routes

A BGP speaker acting as a PE MAY keep the received MUP Direct Segment Discovery routes advertised from other PEs. The receiving BGP speaker will perform the importing of the received MUP Direct Segment Discovery routes in the configured routing instance based on the Route Target extended communities. The IP prefixes for the received segments are imported into the configured routing instance table. Thereby the receiving BGP speaker can provide network connectivity between the nodes that exist in the segments. A BGP speaker MAY discard the received Direct Segment Discovery route if the speaker fails to import the route based on the Route Target extended communities.

The BGP speaker receiving the Direct Segment Discovery routes SHOULD ignore the nexthop in the MP_REACH_NLRI attribute. However, the receiving BGP speaker MUST ensure that the received nexthop value in the MP_REACH_NLRI attribute is identical to the originator of the locator value in the prefix SID attribute. The originator of the locator value can be resolved from the IPv6 IGP table. If the result of the match is not identical then the receiving BGP speaker MUST consider it as a malformed NLRI and the "Treat-as-withdraw procedure of [RFC7606] is applied. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

When a BGP speaker receives a MP_REACH_NLRI attribute update message with a Direct Segment Discovery route without a prefix SID attribute, than it MUST be treated as if it contained a malformed prefix SID attribute and the "Treat-as-withdraw procedure of [RFC7606] is applied. A BGP speaker MUST skip such NLRIs and continue processing of rest of the Update message.

When a BGP speaker receives an MP_UNREACH_NLRI attribute update message it MUST delete the withdrawn Direct Segment Discovery route from the routing instance table where it was created.

3.3.7. Generation of the Type 1 ST Route

A BGP speaker acting as a MUP Controller generates Type 1 ST route from corresponding session information through a northbound API as per [I-D.mhkk-dmm-srv6mup-architecture]. The northbound API definition for ST1 route creation is out of scope of this document.

In 3GPP 5G specific case, to compose a Type 1 ST route the controller acquires UE address or prefix, Tunnel Endpoint address of GTP [TS.29281] tunnel, TEID, and QFI for access side as input parameters from the northbound API. Controller also acquires optionally the associated Source Address. The controller decides the RD of the Type 1 ST route based on operator policy.

When advertising the Type 1 ST route, the controller SHOULD attach a Route Target Extended community which the PEs are importing into the routing instance for the corresponding Direct segment.

The MUP Controller MUST set the nexthop of the route to the address of the controller.

The controller MUST announce this route using a AFI of the route and the SAFI of BGP-MUP to all other BGP speakers within the SRv6 domain.

3.3.8. Withdrawing of the Type 1 ST Route

A BGP speaker acting as a MUP Controller withdraws Type 1 ST route based on deletion of corresponding session information through a northbound API as per [I-D.mhkk-dmm-srv6mup-architecture]. The northbound API definition for ST1 route withdraw is out of scope of this document.

In 3GPP 5G specific case, to withdraw a Type 1 ST route the controller acquires the UE address or prefix, Tunnel Endpoint address of GTP [TS.29281] tunnel, TEID, and QFI for access side as input parameters from the northbound API. Controller also acquires optionally Source Address. The controller MUST advertise the withdraws of the Type 1 ST route.

When withdrawing the Type 1 ST route, the controller SHOULD attach the Route Target Extended community which the PEs are importing into the routing instance accomodating the corresponding Direct segment to the Route Target Extended community.

3.3.9. Processing of the Type 1 ST Route

A BGP speaker acting as a PE MAY keep the received MUP Type 1 ST routes advertised from the MUP Controller. The receiving BGP speaker will perform the importing of the received MUP Type 1 ST routes in the configured N6DN routing instance based on the Route Target extended communities. A BGP speaker MAY discard the received Type 1 ST route if the speaker fails to import the route based on the Route Target extended communities.

In case of a BGP speaker receiving a Type 1 ST routes is a PE, the PE SHOULD use the received Tunnel Endpoint Address in this NLRI as a key to lookup the associated Interwork Segment Discovery route and extract the locator and the function in the prefix SID attribute of the Interwork route.

In 3GPP 5G specific case, the PE extracts TEID, QFI, Tunnel Endpoint address and the Source address if it is present from the NLRI. A PE MAY use a locally configured Source address in an event if the Source address is not present in Type 1 ST route. The Source address is used as the source address in the SRv6 Header. The PE SHOULD generate the forwarding SID for GTP4/6.E based on the procedures mentioned in the [I-D.ietf-dmm-srv6-mobile-uplane]. If the PE cannot generate the prefix SID, then it SHOULD mark the received Type 1 ST route as an invalid route. If the PE does not have a Source address for the given route, then the PE MAY hold such an invalid route until it is withdrawn. Similarly, the PE MAY hold such an invalid route until the route as a valid route upon successful generation of prefix SID.

The PE receiving Type 1 ST routes SHOULD ignore the received nexthop in the MP_REACH_NLRI attribute.

The PE receiving Type 1 ST routes in MP_UNREACH_NLRI attribute MUST delete all the routes from the associated routing instance. In an event where the received Type 1 ST route in MP_UNREACH_NLRI attribute does not have a Source address a receiving PE MUST delete all the matching Type 1 ST Routes with different Source addresses.

3.3.10. Generation of the Type 2 ST Route

A BGP speaker acting as a MUP Controller generates Type 2 ST route from corresponding session information through a northbound API, or pre-defined configuration as per [I-D.mhkk-dmm-srv6mup-architecture]. The northbound API definition for ST2 route creation is out of scope of this document.

In 3GPP 5G specific case, to compose a Type 2 ST route the controller acquires the Endpoint consists of Endpoint address of GTP [TS.29281] tunnel and TEID for core side with the effective length of the Endpoint as input parameters. The controller decides the RD of the Type 2 ST route based on operator policy.

When advertising the Type 2 ST route, the controller SHOULD attach a BGP MUP Extended community corresponding to the Direct segment. The sub-type of the Extended community is Direct-Type Segment Identifier. This Segment Identifier is generated from the information received through a northbound API, or a pre-defined configuration as per [I-D.mhkk-dmm-srv6mup-architecture]. The northbound API definition for receving this information is out of scope of this document.

The controller MUST also attach a Route Target Extended community of the routing instances in the PE accomodating the corresponding Interwork segment.

The controller MUST set the nexthop of the route to the address of the MUP Controller.

3.3.11. Withdrawing of the Type 2 ST Route

A BGP speaker acting as a MUP Controller withdraws Type 2 ST route based on deletion of corresponding session information through a northbound API as per [I-D.mhkk-dmm-srv6mup-architecture]. The northbound API definition for ST2 route withdraw is out of scope of this document.

In 3GPP 5G specific case, to withdraw a Type 2 ST route the controller acquires the Endpoint consists of Endpoint address of GTP [TS.29281] tunnel and TEID for core side with the effective length of the Endpoint as input parameters. The controller MUST advertise the withdraws of the Type 2 ST route.

When withdrawing the Type 2 ST route, the controller SHOULD attach the BGP MUP Extended community corresponding to the Direct segment, and the Route Target Extended community which the PEs are importing into the routing instance accomodating the corresponding Interwork segment to the Route Target Extended community.

3.3.12. Processing of the Type 2 ST Route

A BGP speaker acting as a PE MAY keep the received MUP Type 2 ST routes advertised from the MUP Controller. The receiving BGP speaker will perform the importing of the received MUP Type 2 ST routes in the configured N3RAN routing instance based on the Route Target extended communities. A BGP speaker MAY discard the received Type 2 ST route if the speaker fails to import the route based on the Route Target extended communities.

The BGP speaker receiving the Type 2 ST routes SHOULD ignore the received nexthop in the MP_REACH_NLRI attribute.

A PE receiving the Type 2 ST routes in a MP_REACH_NLRI attribute without a BGP MUP Extended community SHOULD consider the route as a malformed route. The PE MUST handle such a malformed NLRI as a "Treat-as-withdraw" [RFC7606].

The PE receiving Type 2 ST route with a BGP MUP Extended Community should extract the received segment identifier from the community. The segment identifier is used to resolve an appropriate Direct segment routing instance.

4. Security Considerations

The mechanisms described in this document could reuse the existing BGP security mechanisms [RFC4271] [RFC4272]. The security model and threats specific to Provider Provisioned VPNs, including L3VPNs, are discussed in [RFC4111]. The method defined in [RFC5925] SHOULD be used where authentication of BGP control packets is needed.

The PEs and MUP Controller SHOULD NOT establish BGP sessions with other BGP speakers in the domains which are not trusted without any explicit configuration or an operator intervention. Usage of procedures defined in [RFC5925] SHOULD be enforced at such boundaries to ensure the proper authentication of BGP control packets.

Furthermore, [RFC5925] will not help to keep the BGP messages private. To protect the BGP messages exchanged between BGP speakers from eavesdrop, establishing BGP sessions over encrypted paths SHOULD be considered.

PEs SHOULD impose an upper bound on number of routes they should store to protect their control plane load. For example, BGP implementations MAY provide a configuration knob to impose an upper bound on Type 1 ST Routes.

5. IANA Considerations

This document defines a new BGP SAFI known as a BGP-MUP SAFI. IANA is requested to assign the value for the new SAFI from the Subsequent Address Family Identifiers (SAFI) registry.

This document defines a new Architecture Type for a BGP-MUP SAFI. IANA is requested to create a new Architecture Type NLRI registry for BGP-MUP SAFI. Furthermore, IANA is also requested to assign values for the following Architecture Types from the newly created BGP-MUP Architecture Type NLRI registry:

+ 1 - 3gpp-5g;

This document defines new NLRIs for a BGP-MUP SAFI. IANA is requested to create a new NLRI registry for BGP-MUP SAFI. Furthermore, IANA is also requested to assign values for the following NLRIs from the newly created BGP-MUP NLRI registry:

+ 1 - Discovery Internetwork route;
+ 2 - Direct Segment Discovery route;
+ 3 - Type 1 Session Transformed (ST) route;
+ 4 - Type 2 Session Transformed (ST) route;

This document defines a new BGP Extended Community called "SRv6 MUP Extended Community". This Community is of an extended type and is transitive. IANA is requested to assign the Type and the Sub-Type value for this Community from the Transitive Extended Community registry.

6. Contributors

In addition to the authors listed on the front page, the following individuals have also made significant contributions to the draft:

Katsuhiro Horiba
SoftBank
Email: katsuhiro.horiba@g.softbank.co.jp

Yuya Kawakami
SoftBank
Email: yuya.kawakami01@g.softbank.co.jp

Derek Yeung
Arrcus
Email: derek@arrcus.com

Kalyani Rajaraman
Email: kalyanir@yahoo.com

7. References

7.1. Normative References

[I-D.mhkk-dmm-srv6mup-architecture]
Matsushima, S., Horiba, K., Khan, A., Kawakami, Y., Murakami, T., Patel, K., Kohno, M., Kamata, T., Camarillo, P., Horn, J., Voyer, D., Zadok, S., Meilik, I., Agrawal, A., and K. Perumal, "Mobile User Plane Architecture using Segment Routing for Distributed Mobility Management", Work in Progress, Internet-Draft, draft-mhkk-dmm-srv6mup-architecture-06, 23 October 2023, <https://datatracker.ietf.org/doc/html/draft-mhkk-dmm-srv6mup-architecture-06>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, DOI 10.17487/RFC2918, September 2000, <https://www.rfc-editor.org/info/rfc2918>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[I-D.ietf-dmm-srv6-mobile-uplane] Matsushima, S., Filsfils, C., Kohno, M., Camarillo, P., and D. Voyer, "Segment Routing IPv6 for Mobile User Plane", Work in Progress, Internet-Draft, draft-ietf-dmm-srv6-mobile-uplane-24, 17 January 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-dmm-srv6-mobile-uplane-24>.

[RFC4111] Fang, L., Ed., "Security Framework for Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4111, DOI 10.17487/RFC4111, July 2005, <https://www.rfc-editor.org/info/rfc4111>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC6811] Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", RFC 6811, DOI 10.17487/RFC6811, January 2013, <https://www.rfc-editor.org/info/rfc6811>.

[RFC8205] Lepinski, M., Ed. and K. Sriram, Ed., "BGPsec Protocol Specification", RFC 8205, DOI 10.17487/RFC8205, September 2017, <https://www.rfc-editor.org/info/rfc8205>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[TS.23501] 3GPP, "System architecture for the 5G System (5GS)", 3GPP TS 23.501 17.2.0, 24 September 2021, <http://www.3gpp.org/ftp/Specs/html-info/23501.htm>.

[TS.29281] 3GPP, "General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U)", 3GPP TS 29.281 16.1.0, September 2020.

Authors' Addresses

Tetsuya Murakami
Arrcus, Inc.
2077 Gateway Place, Suite 400
San Jose, CA 95110
United States of America
Email: tetsuya@arrcus.com

Keyur Patel
Arrcus, Inc.
2077 Gateway Place, Suite 400
San Jose, CA 95110
United States of America
Email: keyur@arrcus.com

Satoru Matsushima
SoftBank
Japan
Email: satoru.matsushima@g.softbank.co.jp

Jeffrey Zhang
Juniper Networks
United States of America
Email: zzhang@juniper.net

Swadesh Agrawal
Cisco Systems
United States of America
Email: swaagraw@cisco.com

Daniel Voyer
Bell Canada
Montreal
Canada
Email: daniel.voyer@bell.ca

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 September 2022

A. Sajassi, Ed.
G. Badoni
P. Warade
S. Pasupula
L. Krattiger
Cisco Systems
J. Drake, Ed.
Juniper
J. Rabadan, Ed.
Nokia
7 March 2022

EVPN Support for L3 Fast Convergence and Aliasing/Backup Path
draft-sajassi-bess-evpn-ip-aliasing-04

Abstract

This document proposes an EVPN extension to allow several of its multihoming functions, fast convergence and aliasing/backup path, to be used in conjunction with inter-subnet forwarding.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

This document proposes an EVPN extension to allow several of its multihoming functions, fast convergence and aliasing/backup path, to be used in conjunction with inter-subnet forwarding. It re-uses the existing EVPN routes, the Ethernet A-D per ES and the Ethernet A-D per EVI routes, which are used for these multihoming functions. In particular, there are three use-cases that could benefit from the use of these multihoming functions:

a. Inter-subnet forwarding for host routes in symmetric IRB [RFC9135].

b. Inter-subnet forwarding for prefix routes in the interface-less IP-VRF-to-IP-VRF model [RFC9136].

c. Inter-subnet forwarding for prefix routes when the ESI is used exclusively as an L3 construct [RFC9136].

1.1. Ethernet Segments for Host Routes in Symmetric IRB

Consider a pair of multi-homing PEs, PE1 and PE2, as illustrated in Figure 1. Let there be a host H1 attached to them. Consider PE3 and a host H3 attached to it.

```
                                +----------------+
                                |     EVPN       |
                             +------+            |
                             | PE1  | +--->      |
                      +------+      | RT-2       |
                      |      |      | IP1     +--+---+
           +---+  | ES1  +------+ ESI1    | PE3  |
      H1+--+CE1+--+         |              |      +-+H3
           +---+  |      +------+         |      |
                      |      | PE2  |         +--+---+
                      +------+      |            |
                             |      |            |
                             +------+            |
                                |                |
                                +----------------+
```

Figure 1: Inter-subnet traffic between Multihoming PEs and Remote PE

With Asymmetric IRB [RFC9135], if H3 sends inter-subnet traffic to H1, routing will happen at PE3. PE3 will be attached to the destination IRB interface and will trigger ARP/ND requests if it does not have an ARP/ND adjacency to H1. A subsequent routing lookup will resolve the destination MAC to H1's MAC address. Furthermore, H1's

MAC will point to an ECMP EVPN destination on PE1 and PE2, either due to host route advertisement from both PE1 and PE2, or due to Ethernet Segment MAC Aliasing as detailed in [RFC7432].

With Symmetric IRB [RFC9135], if H3 sends inter-subnet traffic to H1, a routing lookup will happen at PE3's IP-VRF and this routing lookup will not yield the destination IRB interface and therefore MAC Aliasing is not possible. In order to have per-flow load balancing for H3's routed traffic to H1, an IP ECMP list (to PE1/PE2) needs to be associated to H1's host route in the IP-VRF route-table. If H1 is locally learned only at one of the multi-homing PEs, PE1 or PE2, due to LAG hashing, PE3 will not be able to build an IP ECMP list for the H1 host route.

With the extension described in this document, PE3's IP-VRF becomes Ethernet-Segment-aware and builds an IP ECMP list for H1 based on the advertisement of ES1 along with H1 in a MAC/IP route and the availability of ES1 on PE1 and PE2.

1.2. Inter-subnet Forwarding for Prefix Routes in the Interface-less IP-VRF-to-IP-VRF Model

In the Interface-less IP-VRF-to-IP-VRF model described in [RFC9136] there is no Overlay Index and hence no recursive resolution of the IP Prefix route to either a MAC/IP Advertisement or an Ethernet A-D per ES/EVI route, which means that the fast convergence and aliasing/backup path functions are disabled. The recursive resolution of an IP Prefix route to an Ethernet A-D per ES/EVI route is already described in [RFC9136].

The scenario illustrated in Figure 2 will be used to explain the procedures.

```
                                +----------------+
                                |     EVPN       |
                             +------+            |
                             | PE1  | +--->      |
                       +------+      | RT-5       |
                       |     |       | IP1/32  +--+---+
            +---+  | ES1  +------+ ESI1    | PE3  |
       H1+--+CE1+--+         |               |      +-+H3
            +---+  |      +------+         |      |
                       |     | PE2  |         +--+---+
                       +------+      |            |
                             |      |            |
                             +------+            |
                                |                |
                                +----------------+
```

Figure 2: Inter-subnet example with IP Prefix routes

Consider PE1 and PE2 are multi-homed to CE1 (in an All-Active Ethernet Segment ES1), and PE1, PE2 and PE3 are attached to an IP-VRF of the same tenant. Suppose H1's host route is learned (via ARP or ND snooping) on PE1 only, and PE1 advertises an EVPN IP Prefix route for H1's host route. If H3 sends inter-subnet traffic to H1, a routing lookup on PE3 would normally yield a single next-hop, i.e., PE1.

This document proposes the use of the ESI in the IP Prefix route and the recursive resolution to A-D per ES/EVI routes advertised from PE1 and PE2, so that H1's host route in PE3 can be associated to an IP ECMP list (to PE1/PE2) for aliasing purposes.

1.3. Ethernet Segments for Prefix routes in IP-VRF-to-IP-VRF use-cases

This document also enables fast convergence and aliasing/backup path to be used even when the ESI is used exclusively as an L3 construct, in an Interface-less IP-VRF-to-IP-VRF scenario [RFC9136]. There are two use cases analyzed and supported by this document:

* IP Aliasing for EVPN IP Prefix routes

* Centralized Routing Model

1.3.1. IP Aliasing for EVPN IP Prefix routes

As an example, consider the scenario in Figure 3 in which PE1 and PE2 are multi-homed to CE1. However, and contrary to CE1 in Figure 2, in this case the links between CE1 and PE1/PE2 are used exclusively for L3 protocols and L3 forwarding in different BDs, and a BGP session established between CE1's loopback address and PE1's IRB address.

```
                                          +-----------------------+
                                          |        EVPN           |
                          PE1             |                       |
                         +-------------------+                    |
                         |       IRB1        |                    |
                         |  +---+   +------+ | ------->           |
                +-----------|BD1|---|IPVRF1| | RT-5               |
        eBGP    |        |  +---+   |      | | 50.0/24            | PE3
     +------------------------->10.1 +------+ | ESI1  +----------------+
     |          |        +-------------------+       | +------+       |
    +-----+10.2 |                         |   ^        | |IPVRF1| +---+ |
    | CE1 |-----+    ES1                 |   |        | |      |-|BD3| |
    |     |-----+                         |   +--------| +------+ +---+ |
    +-----+20.2 |         PE2             |        +---|              |  |
    lo1         |        +--------------+----+   |   +------------|---+
    1.1.1.1     |        |       IRB2        |   |                | |
    Prefixes:   |        |  +---+   +------+ |   |                | H4
    50.0/24     +-----------|BD2|---|IPVRF1| |<--+                |
    60.0/24              |  +---+   |      | |                    |
                         |     20.1 +------+ |                    |
                         +-------------------+                    |
                                          |                       |
                                          +-----------------------+
```

Note:
IP addresses expanded by adding 0s
E.g., 50.0 expands to 50.0.0.0

Figure 3: Layer-3 Multihoming PEs

In these use-cases, sometimes the CE supports a single BGP session to one of the PEs (through which it advertises a number of IP Prefixes seating behind itself) and yet, it is desired that remote PEs can build an IP ECMP list or backup IP list including all the PEs multi-homed to the same CE. For example, in Figure 3, CE1 has a single eBGP neighbor, i.e., PE1. Load-balancing for traffic from CE1 to H4 can be accomplished by a default route with next-hops PE1 and PE2, however, load-balancing from H4 to any of the prefixes attached to CE1 would not be possible since only PE1 would advertise EVPN IP Prefix routes for CE1's prefixes. This document provides a solution so that PE3 considers PE2 as a next-hop in the IP ECMP list for CE1's prefixes, even if PE2 did not advertise the IP Prefix routes for those prefixes in the first place.

1.3.2. Centralized Routing Model

Figure 4 illustrates a model in which multiple CEs establish an eBGP PE-CE session with a Centralized PE.

```
                    +-------------------------------+
                    | PE1       EVPN                |
               +----------+                         |
               |  +------+|                         |
               |  |IP-VRF||                         |
       10.1 --------------------+                   |
      +---+    |+--+     ||      |eBGP               |
      |CE1|----||BD|-----+|      |PE-CE              |
      |   |-+  |+--+      |      |50.0/24            | PE3
      +---+ |  +----------+      |NH 10.1      +----------+
   Prefixes:|        |           |             |+------+  |
   50.0/24  |        |           |             ||IP-VRF|  |
   60.0/24  |        | PE2       |   +--------->|    +--+ |
            |  +----------+      |   |         |+----|BD| |
            |  |  +------+|      |   |         |     +--+ |
            |  |  |IP-VRF||      |   |         +----------+
            |  |  |      ||      |   |              |  |
            |  |+--+     ||      |   |RT-5          |  |
          +--||BD|-----+|      |   |50.0/24       | H4
               |+--+      |      |   |ESI1          |
               +----------+      |   |NH PEC        |
                    |            |   |              |
                    |        30.1|   | PEC          |
                    |         +--V---|-+            |
                    |         |+------+|            |
                    +---------||IP-VRF||------------+
                              |+------+|
                              +--------+
     Note:
       IP addresses expanded by adding 0s
       E.g., 50.0 expands to 50.0.0.0
```

Figure 4: Centralized Routing Model

The CEs in this case are usually VNFs (Virtual Network Function entities) or CNFs (Containerized Network Function entities) and by provisioning the same network parameters on all of them, the operation gets significantly simplified. The configuration on the PEs also gets simplified, since the PE-CE eBGP sessions to the CEs are only configured on a centralized PE. In the diagram, CE1 is one of these VNF/CNFs that sets up a multi-hop eBGP session to the centralized PEC. As an example, CE1 advertises prefix 50.0.0.0/24 with Next Hop 10.0.0.1 (to PEC) via the multi-hop eBGP session. PEC then exports the prefix into a RT-5 route, following the Interface-less IP-VRF-to-IP-VRF model [RFC9136], with Next Hop PEC. When H4 sends traffic to an IP address of the subnet 50.0.0.0/24, the traffic will be forwarded to PEC first, and PEC will then forward to PE1 (or PE2). In other words, this model simplifies the configuration and

operation of the CEs, however, it introduces an inefficiency since traffic needs to go through the Centralized PE (PEC) instead of going directly to the PE(s) attached to the destination CE. The IP Aliasing solution specified in this document overcomes this inefficiency and allows traffic from PE3 to be forwarded directly to PE1 or PE2, without going through PEC.

1.4. Terminology and Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

* IRB: Integrated Routing and Bridging

* IRB Interface: Integrated Bridging and Routing Interface. A virtual interface that connects the Bridge Table and the IP-VRF on an NVE.

* BD: Broadcast Domain. An EVI may be comprised of one BD (VLAN-based or VLAN Bundle services) or multiple BDs (VLAN-aware Bundle services).

* Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.

* CE: Customer Edge device, e.g., a host, router, or switch.s

* EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

* MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

* Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

* Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

* IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by any routing protocol, E.g., EVPN, IP-VPN and BGP PE-CE IP address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

* EVPN IP route: An EVPN IP Prefix route or an EVPN MAC/IP Advertisement route.

* LACP: Link Aggregation Control Protocol.

* PE: Provider Edge device.

* Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

* All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

* RT-2: EVPN MAC/IP Advertisement route, as specified in [RFC7432].

* RT-4: EVPN Ethernet Segment route, as specified in [RFC7432].

* RT-5: EVPN IP Prefix route, as specified in [RFC9136].

2. Ethernet Segments for L3 Aliasing/Backup Path and Fast Convergence

The first two use cases described in Section 1 do not require any extensions to the Ethernet Segment definition and both cases support Ethernet Segments as a set of Ethernet links and specified in [RFC7432], or virtual Ethernet Segments as a set of logical links specified in [I-D.ietf-bess-evpn-virtual-eth-segment].

The third use case in Section 1 requires an extension to the way Ethernet Segments are defined and associated. In this case, the Ethernet Segment is a Layer-3 construct characterized as follows:

1. The ES is defined as a set of Layer-3 links to the multi-homed CE and its state MUST be linked to the layer-3 reachability from each multi-homed PE to the CE's loopback address via a non-EVPN route in the PE's IP-VRF.

2. The ESI SHOULD be of type 4 [RFC7432] and set to the router ID of the multi-homed CE.

3. All-active or single-active multi-homing redundancy modes are supported, however, the redundancy mode only affects the procedures in Section 3.

4. PEs attached to the same Layer-3 ES discover each other through the exchange of RT-4 routes (Ethernet Segment routes). DF Election procedures [RFC8584] MAY be used for single-active multi-homing mode.

5. The routes advertised from the multi-homed CE's and installed in the PE's IP-VRF table with the CE's loopback as the next-hop MUST be re-advertised by the PE in EVPN IP Prefix routes with the ESI of the CE. The rest of the EVPN IP Prefix routes fields are set as per the Interface-less model in [RFC9136]. Note that the BGP PE-CE routes advertised by the multi-homed CE are installed in the IP-VRF normally irrespective of the Next Hop being resolved to an EVPN or a non-EVPN route, and they are exported as a RT-5 with the ESI.

In the example depicted in Figure 3, ES1 is defined as the set of layer-3 links that connects PE1 and PE2 to CE1. Its ESI, e.g., ESI-1, is derived as a type 4 ESI using the CE's router ID. ES-1 will be operationally up in the PE as long as CE1's loopback route is installed in the PE's IP-VRF and learned via any routing protocol except for an EVPN route. E.g., an active static route to 1.1.1.1 via next-hop 10.0.0.2 would make the ES operationally up in PE1, and the eBGP routes received from CE1 with next-hop 1.1.1.1 will be re-advertised as RT-5 routes with ESI-1.

In the example illustrated in Figure 4, ES1 is a set of layer-3 links connecting PE1, PE2 and PEC to CE1. ESI-1 is derived as a type 4 ESI using the CE's router ID, as in the previous example. CE1's loopback route (which is associated to ES1) is installed in PE1 and PE2 via non-EVPN route, hence ES1 is operationally up in PE1 and PE2. On PE-C though, CE1's loopback is installed via EVPN IP Prefix route, therefore, as per point 1 in the current section, ES1 is operationally down in PEC. As per point 5, this does not prevent PEC from exporting CE1's prefixes into RT-5 routes with ESI-1. However, since ES-1 is operationally down in PEC, no IP A-D per EVI routes (Section 3) and no IP A-D per ES routes Section 4 for ESI-1 will be advertised from PEC, preventing PEC from attracting traffic destined to CE1.

3. IP Aliasing and Backup Path

In order to address the use-cases described in Section 1, above, this document proposes that:

1. A PE that is attached to a given ES will advertise a set of one or more Ethernet A-D per ES routes for that ES. Each is termed an 'IP A-D per ES' route and is tagged with the route targets (RTs) for one or more of the IP-VRFs defined on it for that ES; the complete set of IP A-D per ES routes contains the RTs for all of the IP-VRFs defined on it for that ES.

   A remote PE imports an IP A-D per ES route into the IP-VRFs corresponding to the RTs with which the route is tagged. When the complete set of IP A-D per ES routes has been processed, a remote PE will have imported an IP A-D per ES route into each of the IP-VRFs defined on it for that ES; this enables fast convergence for each of these IP-VRFs.

2. A PE advertises for this ES, an Ethernet A-D Per EVI route for each of the IP-VRFs defined on it. Each is termed an 'IP A-D per EVI' route and is tagged with the RT for a given IP-VRF, and conveys a label that identifies that IP-VRF.

   A remote PE imports an IP A-D per EVI route into the IP-VRF corresponding to the RT with which the route is tagged. The label contained in the route enables aliasing/backup path for the routes in that IP-VRF.

To address the third use-case described in Section 1, where the links between a CE and its multihomed PEs are used exclusively for L3 protocols and L3 forwarding, a PE uses the procedures described in 1) and 2), above.

The processing of the IP A-D per ES and the IP A-D per EVI routes is as defined in [RFC7432] and [RFC8365] except that the fast convergence and aliasing/backup path functions apply to the routes contained in an IP-VRF. In particular, a remote PE that receives an EVPN MAC/IP Advertisement route or an IP Prefix route with a non-reserved ESI and the RT of a particular IP-VRF SHOULD consider it reachable by every PE that has advertised an IP A-D per ES and IP A-D per EVI route for that ESI and IP-VRF.

3.1. Constructing the IP A-D per EVI Route

The construction of the IP A-D per EVI route is the same as that of the Ethernet A-D per EVI route, as described in [RFC7432], with the following exceptions:

* The Route-Distinguisher is for the corresponding IP-VRF.

* The Ethernet Tag should be set to 0.

* The route SHOULD carry the Route Target of the corresponding IP-VRF.

* The route MUST carry the MPLS label, VNI (VXLAN Network Identifier [RFC8365]) or Segment Routing IPv6 SID (Segment Identifier [I-D.ietf-bess-srv6-services]) that identifies the corresponding IP-VRF.

* The route MUST carry the PE's MAC Extended Community if the encapsulation used between the PEs for inter-subnet forwarding is an Ethernet NVO tunnel [RFC9136].

* The route SHOULD carry the EVPN Layer 2 Extended Community [I-D.ietf-bess-rfc7432bis]. For all-active multihoming, all PEs attached to the specified ES will advertise P=1. For backup path, the Primary PE will advertise P=1 and the Backup PE will advertise P=0, B=1.

  - The Primary PE SHOULD be a PE with a routing adjacency to the attached CE.

  - The Primary PE MAY be determined by policy or MAY be elected by a DF Election as in [RFC8584] as described in Section 2.

4. Fast Convergence for Routed Traffic

Host or Prefix reachability is learned via the BGP-EVPN control plane over the MPLS/NVO network. EVPN IP routes for a given ES are advertised by one or more of the PEs attached to that ES. When one of these PEs fails, a remote PE needs to quickly invalidate the EVPN IP routes received from it.

To accomplish this, EVPN defined the fast convergence function specified in [RFC7432]. This document extends fast convergence to inter-subnet forwarding by having each PE advertise a set of one or more IP A-D per ES routes for each locally attached Ethernet segment (refer to Section 4.1 below for details on how these routes are constructed). A PE may need to advertise more than one IP A-D per ES route for a given ES because the ES may be in a multiplicity of IP-VRFs and the Route Targets for all of these IP-VRFs may not fit into a single route. Advertising a set of IP A-D per ES routes for the ES allows each route to contain a subset of the complete set of Route Targets. Each IP A-D per ES route is differentiated from the other routes in the set by a different Route Distinguisher (RD).

Upon failure in connectivity to the attached ES, the PE withdraws the corresponding set of IP A-D per ES routes. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for

all IP addresses associated with the Ethernet Segment in question, across IP-VRFs. If no other PE has advertised an IP A-D per ES route for the same Ethernet Segment, then the PE that received the withdrawal simply invalidates the IP entries for that segment. Otherwise, the PE updates its next-hop adjacencies accordingly.

These routes should be processed with higher priority than EVPN IP route withdrawals upon failure. Similar priority processing is needed even on the intermediate Route Reflectors.

4.1. Constructing IP A-D per Ethernet Segment Route

This section describes the procedures used to construct the IP A-D per ES route, which is used for fast convergence (as discussed in Section 4). The usage/construction of this route remains similar to that described in section 8.2.1. of [RFC7432] with a few notable exceptions as explained in following sections.

4.1.1. IP A-D per ES Route Targets

Each IP A-D per ES route MUST carry one or more Route Targets. The set of IP A-D per ES routes MUST carry the entire set of IP-VRF Route Targets for all the IP-VRFs defined on that ES.

4.2. Avoiding convergence issues by synchronizing IP prefixes

Consider a pair of multi-homing PEs, PE1 and PE2. Let there be a host H1 attached to them. Consider PE3 and a host H3 attached to it.

If the host H1 is learned on both the PEs, the ECMP path list is formed on PE3 pointing to (PE1/PE2). Traffic from H3 to H1 is not impacted even if one of the PEs fails as the path list gets corrected upon receiving the withdrawal of the fast convergence route(s) (IP A-D per ES routes).

In a case where H1 is locally learned only on PE1 due to LAG hashing or a single routing protocol adjacency to PE1, at PE3, H1 has ECMP path list (PE1/PE2) using Aliasing as described in this document. Traffic from H3 can reach H1 via either PE1 or PE2.

PE2 should install local forwarding state for EVPN IP routes advertised by other PEs attached to the same ES (i.e., PE1) but not advertise them as local routes. When the traffic from H3 reaches PE2, PE2 will be able forward the traffic to H1 without any convergence delay (caused by triggering ARP/ND to H1 or to the next-hop to reach H1). The synchronization of the EVPN IP routes across all PEs of the same Ethernet Segment is important to solve convergence issues.

4.3. Handling Silent Host MAC/IP route for IP Aliasing

Consider the example of Figure 1 for IP aliasing. If PE1 fails, PE3 will receive the withdrawal of the fast convergence route(s) and update the ECMP list for H1 to be just PE2. When the EVPN IP route for H1 is also withdrawn, neither PE2 nor PE3 will have a route to H1, and traffic from H3 to H1 is blackholed until PE2 learns H1 and advertises an EVPN IP route for it.

This blackholing can be much worse if the H1 behaves like a silent host. IP address of H1 will not be re-learned on PE2 till H1 ARP/ND messages or some traffic triggers ARP/ND for H1.

PE2 can detect the failure of PE1's reachability in different ways:

a. When PE1 fails, the next hop tracking to PE1 in the underlay routing protocols can help detect the failure.

b. Upon the failure of its link to CE1, PE1 will withdraw its IP A-D route(s) and PE2 can use this as a trigger to detect failure.

Thus to avoid blackholing, when PE2 detects loss of reachability to PE1, it should trigger ARP/ND requests for all remote IP prefixes received from PE1 across all affected IP-VRFs. This will force host H1 to reply to the solicited ARP/ND messages from PE2 and refresh both MAC and IP for the corresponding host in its tables.

Even in core failure scenario on PE1, PE1 must withdraw all its local layer-2 connectivity, as Layer-2 traffic should not be received by PE1. So when ARP/ND is triggered from PE2 the replies from host H1 can only be received by PE2. Thus H1 will be learned as local route and also advertised from PE2.

It is recommended to have a staggered or delayed deletion of the EVPN IP routes from PE1, so that ARP/ND refresh can happen on PE2 before the deletion.

4.4. MAC Aging

In the same example as in Section 4.3, PE1 would do ARP/ND refresh for H1 before it ages out. During this process, H1 can age out genuinely or due to the ARP/ND reply landing on PE2. PE1 must withdraw the local entry from BGP when H1 entry ages out. PE1 deletes the entry from the local forwarding only when there are no remote synced entries.

5. Determining Reachability to Unicast IP Addresses

5.1. Local Learning

The procedures for local learning do not change from [RFC7432] or [RFC9136].

5.2. Remote Learning

The procedures for remote learning do not change from [RFC7432] or [RFC9136].

5.3. Constructing the EVPN IP Routes

The procedures for constructing MAC/IP Address or IP Prefix Advertisements do not change from [RFC7432] or [RFC9136].

5.3.1. Route Resolution

If the ESI field is set to reserved values of 0 or MAX-ESI, the EVPN IP route resolution MUST be based on the EVPN IP route alone.

If the ESI field is set to a non-reserved ESI, the EVPN IP route resolution MUST happen only when both the EVPN IP route and the associated set of IP A-D per ES routes have been received. To illustrate this with an example, consider a pair of multi-homed PEs, PE1 and PE2, connected to an all-active Ethernet Segment. A given host with IP address H1 is learned by PE1 but not by PE2. When the EVPN IP route from PE1 and a set of IP A-D per ES and IP A-D per EVI routes from PE1 and PE2 are received, then (1) PE3 can forward traffic destined to H1 to both PE1 and PE2.

If after (1) PE1 withdraws the IP A-D per ES route, then PE3 will forward the traffic to PE2 only.

If after (1) PE2 withdraws the IP A-D per ES route, then PE3 will forward the traffic to PE1 only.

If after (1) PE1 withdraws the EVPN IP route, then PE3 will do delayed deletion of H1, as described in Section 4.3.

If after (1) PE2 advertised the EVPN IP route, but PE1 withdraws it, PE3 will continue forwarding to both PE1 and PE2 as long as it has the IP A-D per ES and the IP A-D per EVI route from both.

6. Forwarding Unicast Packets

Refer to Section 5 in [RFC9135] and [RFC9136].

7. Load Balancing of Unicast Packets

The procedures for load balancing of Unicast Packets do not change from [RFC7432]

8. IP Aliasing and Unequal ECMP for IP Prefix Routes

[I-D.ietf-bess-evpn-unequal-lb] specifies the use of the EVPN Link bandwidth extended community to achieve weighted load balancing to an ES or Virtual ES for unicast traffic. The procedures in [I-D.ietf-bess-evpn-unequal-lb] MAY be used along with the procedures described in this document for any of the three cases described in Section 1, with the following considerations:

* The ES weight is signaled by the multi-homed PEs in the IP A-D per ES routes.

* The remote ingress PE learning an EVPN IP Route to prefix/host P that is associated to a weighted load balancing ES, will follow the procedures in [I-D.ietf-bess-evpn-unequal-lb] to influence the load balancing for traffic to P.

* [I-D.ietf-bess-evpn-unequal-lb] also allows the use of the EVPN Link Bandwidth Extended Community along with RT-5s. If the ingress PE learns a prefix P via a non-reserved ESI RT-5 route with a weight (for which IP A-D per ES routes also signal a weight) and a zero ESI RT-5 that includes a weight, the ingress PE will consider all the PEs attached to the ES as a single PE when normalizing weights.

  As an example, consider PE1 and PE2 are attached to ES-1 and PE1 advertises an RT-5 for prefix P with ESI-1 (and EVPN Link Bandwidth of 1). Consider PE3 advertises an RT-5 for P with ESI=0 and EVPN Link Bandwidth of 2. If PE1 and PE2 advertise an EVPN Link Bandwidth of 1 and 2, respectively, in the IP A-D per ES routes for ES-1, an ingress PE4 SHOULD assign a normalized weight of 1 to ES-1 and a normalized weight of 2 to PE3. When PE4 sprays the flows to P, it will send twice as many flows to PE3. For the flows sent to ES-1, the individual PE EVPN Link Bandwidths advertised in the IP A-D per ES routes will be considered.

9. Security Considerations

The mechanisms in this document use EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable. This document uses MPLS and IP-based tunnel technologies to support data plane transport. Security considerations described in [RFC7432] and in [RFC8365] are equally applicable.

10. IANA Considerations

No IANA considerations.

11. Contributors

12. Acknowledgments

13. References

13.1. Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/info/rfc9135>.

[RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <https://www.rfc-editor.org/info/rfc9136>.

[I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", Work in Progress, Internet-Draft, draft-ietf-bess-rfc7432bis-03, 28 February 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-rfc7432bis-03.txt>.

13.2. Informative References

[I-D.ietf-bess-evpn-virtual-eth-segment] Sajassi, A., Brissette, P., Schell, R., Drake, J. E., and J. Rabadan, "EVPN Virtual Ethernet Segment", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-virtual-eth-segment-07, 6 July 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-virtual-eth-segment-07.txt>.

[I-D.ietf-bess-evpn-unequal-lb] Malhotra, N., Sajassi, A., Rabadan, J., Drake, J., Lingala, A., and S. Thoria, "Weighted Multi-Path Procedures for EVPN Multi-Homing", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-unequal-lb-15, 17 November 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-unequal-lb-15.txt>.

[I-D.ietf-bess-srv6-services] Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay Services", Work in Progress, Internet-Draft, draft-ietf-bess-srv6-services-12, 5 March 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-srv6-services-12.txt>.

Authors' Addresses

A. Sajassi (editor)
Cisco Systems
Email: sajassi@cisco.com

G. Badoni
Cisco Systems
Email: gbadoni@cisco.com

P. Warade
Cisco Systems
Email: pwarade@cisco.com

S. Pasupula
Cisco Systems
Email: surpasup@cisco.com

L. Krattiger
Cisco Systems
Email: lkrattig@cisco.com

J. Drake (editor)
Juniper
Email: jdrake@juniper.net

J. Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: jorge.rabadan@nokia.com

BESS WG
Internet-Draft
Intended status: Informational
Expires: 31 July 2024

S. Dikshit
Aruba, HPE
V. Joshi
Oracle India Pvt Ltd
S. Shankar
Aruba, HPE

28 January 2024

Defreezing Optimization post EVPN Mac Dampening
draft-saum-bess-dampening-backoff-08

Abstract

MAC move handling in EVPN deployments is discussed in detail in [RFC7432]. There are few optimizations which can be done in existing way of handling the mac duplication. This document describes few of the potential techniques to do so. This document is of informational type based on comments in the ietf meeting.

Status of This Memo

Copyright Notice

Table of Contents

1. Important Terms

MDAS: Mac Dampening Attribute Set:

MDT: Mac Dampening Timer

MDC: Mac Dampening Count

MFT: MAC Freezing Timer

Mac Dampening: Process of stalling the mobility of MAC as define in [RFC7432].

VTEP: Virtual Tunnel End Point or Vxlan Tunnel End Point

DT: Dampened Time: Actual time taken to dampen the contentious MAC

2. Introduction

The host mobility solution described in [RFC7432] elaborates on few use-cases related to dual mac discovery which leads to dampening logic coming into play. The host move handling logic addresses the problem of frequent mac-moves and culminates by freezing the MACs against further moves. If there is no mellowing down of the issue, then it leads to unending cycle of mac dampening and freezing. Hence this problem needs organic measures for arriving at MAC freezing point, sooner than later.

The events that can lead to never ending duplication are as follows:

(a) Misconfiguration of hosts with identical configuration, in the same bridge-domain, across ESIs and across NVEs.

(b) Looping of traffic due to layer 2 loops created in the bridge domain in the tenant network behind the NVEs.

2.1. Misconfiguration of Hosts

Consider the following figure wherein two hosts, Host-1 and Host-2, are misconfigured with same mac-address MAC-1. These hosts are placed behind two different Ethernet segments, ES12 and ES3 respectively and hooked to the same bridge-domain (BD-1). PE1, PE2 and PE3 will get into a never ending loop of learning the MAC-1 locally and also from the remote Vtep. Thus entering into a control-plane BGP-EVPN cycle of bumping up the sequence number in the MACMobility Extended Community till the maximum MAC move count is hit with the stipulated time. The MAC published to other Vteps like PE4 also changes accordingly based on the latest update with highest sequence number.

```
                                +------+
                                |Host-3|
                                +------+
                                   |
                            PE4    |
                           +-----+-----+
             +-------------|  +-----+  |-------------+
             |             |  | BD-1|  |             |
             |             +-----------+             |
             |                                       |
             |                 EVPN                  |
             |                                       |
             | PE1             PE2               PE3 |
             |                                       |
         +-----------+     +-----------+     +-----------+
         |  | BD-1|  |     |  | BD-1|  |     |  | BD-1|  |
         |  +-----+  |-----|  +-----+  |-----|  +-----+  |
         +-----------+     +-----------+     +-----------+
                \     ES12      /                 / ES3
                 \             /                 /
                  \           /                 /
                   +-------------+           +-------------+
                   |Host-1, MAC-1|           |Host-2, MAC-1|
                   +-------------+           +-------------+
```

Figure 1: Figure 1: Misconfiguration of Hosts

LEGEND:
PE1, PE2, PE3: Vxlan/overlay gateways
HOST-1, HOST-2: Hosts behind PEs
MAC_1 : MAC address which is duplicated across hosts HOST-1 and HOST-2
ES12: Ethernet segment between PE1 and PE2 for BD-1
ES3: Ethernet segment attached to PE3 for BD-1
BD-1: Bridge Domain 1

2.2. Loopy Traffic in Tenant Network

Consider the following case of a loopy tenant network, leading to MAC duplicity in the network. Lets say, Host-1 generates a BUM traffic like GARP (Gratuitous ARP) and sends it over the VLAN which is part of BD-1 and mapped to a configured EVI on the PEs. PE1 sprays the BUM over the EVPN fabric tying it with the mapped EVI. The BUM packet arrives at PE1 (assuming it’s the elected DF) over the EVPN fabric. PE1 sprays the traffic towards the directly attached tenant network attached, tagging it with Vlan that maps to to the bridge domain, BD-1, which inturn maps to the MAC-VRF pointed to by the EVI. If the layer-2 network on tenant side is loopy due to STP network not

converging or STP not configured at all, or for some other unknown reasoni (not under the purview of this document); then the BUM traffic may loop back to PE1, thus creating a duplicate MAC learning for MAC-1. Till the tenant network is curtailed or put to order via admin intervention or otherwize, continuous MAC moves for MAC-1 can be observed between PEs attached to ethernet segment ES12 (PE1) and ES3 (PE2).

```
                            PE4  |
                           +-----+-----+
            +--------------|  +-----+  |--------------+
            |              |  | BD-1|  |              |
            |              +-----------+              |
            |                                         |
            |                  EVPN                   |
            |                                         |
            | PE1              PE2               PE3  |
            |
        +-----------+       +-----------+       +-----------+
        |  | BD-1|  |       |  | BD-1|  |       |  | BD-1|  |
        |  +-----+  |-------|  +-----+  |-------|  +-----+  |
        +-----------+       +-----------+       +-----------+
         |   ES12   |                              /
         \/         /\                             /
         |          |                            / ES3
       [Tenant-Network]                          /
         |          |                           /
         I->-loop->-I                      +-------------+
                                           |Host-1, MAC-1|
                                           +-------------+
```

Figure 2: Figure 2: Loopy traffic in Tenant Network

LEGEND:
PE1, PE2, PE3: Vxlan/overlay gateways
HOST-1: Hosts behind PE3
MAC_1 : MAC address of Host-1
ES12: Ethernet segment between PE1 and PE2 for BD-1
ES3: Ethernet segment attached to PE3 for BD-1
BD-1: Bridge Domain 1

3. Requirements

3.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

When used in lowercase, these words convey their typical use in common language, and they are not to be interpreted as described in [RFC2119].

4. Problem Description

The mac dampening procedure mentioned in [RFC7432], suggests that a Overlay Tunnel Endpoint that detects the mac mobility event upon local learning, should start a 'M' seconds timer and track the MAC for 'N' moves before the timer expires. Hence forth, concluding that it is a MAC Duplication issue and freezing the MAC while also raising the alarm, for the admin to take corrective action. It is observed in few vendor implementations, that involves defreezing the MAC in deterministic time (configurable or derived) after freezing it, with a positive assumption that admin shall take corrective action meanwhile. Else, the subsequent unfreeze shall end up in the same cycle of MAC Duplication detection and freezing of the MAC. In case of lazy, none or inaccurate intervention by the admin, this can potentially result in ia prolong state of network disarray:

(1) Unnecessary and periodic control-plane protocol churn

(2) Exchange of control plane states which are transient and inaccurate

(3) Reachability to end device remains in the realms of ambiguity for prolonged duration

(4) Traffic destined to the Duplicate MAC case, panning across fabrics, sites or across geographies, ends up hogging the precious WAN bandwidth.

Potential solutions are discussed subsequent sections.

5. Solution(s)

The potential solutions are as follows:

5.1. Mac Freeze

The eventual solution is to FREEZE the MAC forever till admin does the clearing of the MAC. The unfreeze and clearing actions are not organic in nature and can be accompanied by unwarranted impact like clearing of other MACs in the bridge-domain. The way out may be resetting the layer-2 port and thus impacting all tenant bridge-domains hosted on the port. This solution, hence, does not always solves or mitigate the situation, or, it may create a situation from which the eventual bail-out is expensive and not restricted to the

impacted Host.

5.2. Backing Off MAC Mobility Timer and Count

The best-bet to organically mellow down the never ending MAC-mobility (indicating Duplicate MAC), is to freeze the MAC temporarily, for lets say, the same time as MAC Dampening Timer(MDT). Lets term this timer as MAC Freeze Timer(MFT). MFT is the time span for which the contentious MAC is frozen, i.e., no further control plane and data flow is allowed for this MAC. The duplicity/un-ending-mobility is expected to be addressed by the admin. In case the problem is not addressed within MAC Freeze Timer, the MAC duplicity is again identified based on the MAC mobility count within the MAC Dampening timer. The best way forward MAY be:

(1) to get to the duplicity conclusion faster than the earlier iteration

(2) and freeze the MAC for a longer duration than earlier iteration

(3) , With the assumption that the problem shall be resolved in that time frame.

The MAC Dampening Attribute Set (MDAS), comprises of following three parameters:

(1) MAC Dampening Timer (MDT): Defined in [RFC7432]

(2) MAC Dampening Count (MDC): Defined in [RFC7432]

(3) MAC Freeze Timer (MFT): Time for which the MAC is frozen after MAC duplicity is detected

For example, let the first iteration of MDAS_iter_1 {MDT=180 seconds, MDC=5, MFT=180 seconds}. The default values of MDT and MDC are picket from [RFC7432], while lets define the default value of MFT same as MDT. In case admin fails to intervene, the MAC is unfrozen after MFT expires.

For second iteration of the MDAS for the problem-MAC, i.e. MDAS_iter_2 = function (MDAS_iter_1). The MDT and MDC values in second iteration are derived by backing off the MDT and MCD values by a pre-defined delta, i.e.

(1) MDAS_iter_2 (MDT) = MDAS_iter_1 (MDT) decrement_timer_delta

(2) MDAS_iter_2 (MDC) = MDAS_iter_1 (MDC) decrement_count_delta

Thus reducing the time and moves to conclude on duplicity of the MAC. The values of decrement_timer_delta and decrement_count_delta can be configured or derived on a case to case basis. [TBD: Elaborate on the case]. The next step is to freeze the MAC for some more time as compared to the previous iteration set of MDAS, thus increasing the probability of the admin, correcting the issue:

(1) MDAS_iter_2 (MFT) = MDAS_iter_1 (MFT) + increment_timer_delta

(2) The value of increment_timer_delta is also configurable in nature.

5.2.1. MDAS Derivation

The following formulae generalizes the derivation of MDAS attributes in the Nth iteration of Duplicate MAC detection on a PE:

(1) MDAS_iter_(N) (MDT) = (MDAS_iter_(N-1) (MDT)) - decrement_mdt_delta

(2) MDAS_iter_(N) (MDC) = (MDAS_iter_(N-1) (MDC)) decrement_mdc_delta

(3) MDAS_iter_(N) (MFT) = MDAS_iter_(N-1) (MFT) + increment_mft_delta

Where in, the following values for 1st iteration can be define as follows:

MDAS_iter_1 (MDT) = 180 seconds

MDAS_iter_1 (MDC) = 5

MDAS_iter_1 (MFT) = 180 seconds. Many implementations keep the MDT and MFT values as same.

The derivation of MDAS perimeters can be exponential in nature. The delta values can be exponentially increased or decreased after certain iterations, thus triggering a exponential backing off the delta values.

5.2.1.1. MDAS Boundry Values

The new MDT value SHOULD not be less than the time taken to Dampen the MAC movement in last set of MDAS iteration. On the same lines, the new MDC count SHOULD not go below ’2’, as count below 2, the MAC Dampening procedure does not gets triggered.

5.2.2. Delta Values Calculation

Following bullets give a overview of potential ways the delta values, i.e. decrement_mdt_delta, increment_mdc_delta and decrement_mft_delta:

(a) Delta values should be such that they SHOULD not infringe the time or count taken to reach Dampening state in the last set

(b) Delta values are static all through the sets

(c) Delta variable gets incremented/decremented based on the reduction in time (proportionally) to achieve the 'Dampened state' in the last 'MDAS set' as compared to the time to reach the 'Dampened state' in the MDAS set previous to the last one. For the same, the time taken to reach the Dampened State should be cached so that comparisons can be made in subsequent sets. In case, it is the first 'MDAS Set', the delta values MAY be either default or configured ones. For the second 'MDAS set', the value MAY be cross-checked against the Dampened time for the first set.

(d) Delta values are always inherited from admin configuration.

As mentioned in the Section 5.2.1 , the derivation of new delta values can done by exponentially backing them off in subsequent MDAS set(s).

5.3. Backing Off Example

This section describes the example of MDAS calculation with respect to the use-case defined in Section 2.1. Though it's equally applicable to the case described in Section 2.2. This example explains the logic in perspective of PE1. Let's say PE1 learns the MAC-1 locally and publishes it over EVPN control plane before PE2 does the same. PE1 publishes it over control plane before PE2 learns it locally (ignoring the case where both learn in tandem and publish it over control plane). Subsequently, PE2 learns it and publishes it over control plane with sequence number 1. PE1 starts the dampening logic by incrementing the local count by 1 and starting the dampening timer. If this jiggle goes on for 5 counts at PE1, MAC Dampening logic described in [RFC7432]. shall freeze the MAC. PE1 SHOULD cache the time it took to dampen the MAC. Let's say it's 30 seconds.

Assuming admin does not takes any action, before MAC freeze timer expires and PE1 defreezes the MAC, it will start moving again. PE1 shall reduce the MDT value by decrement_mdt_delta = 30 seconds to 150 seconds. The MDC counts are reduced by decrement_mdc_count = 1 to 4

and the MFT is incremented by increment_mft_delta = 20 seconds to 170 seconds. Thus PE1 shall wait for 150 seconds for concluding the dampening logic and tracks the MAC for 4 moves. Once dampening is hit, MAC is rendered as frozen for 170 seconds for admin to take action thus giving some more time for admin to take action.

The whole intention is to gradually move towards a permanent freeze of the MAC if no admin does not do the needful in the stipulated time frame.

6. Backward Compatibility

The backward comptability is a no-op for MDAS derivation and recalculation, as MAC Dampening logic is very local to the Vtep. Even if the remote Vtep does not conforms to the logic presented in this literature, it will still work towards the dampening the frequent mac-mobility with the same parameters of MDT and MDS. The instant freezing or temporary freezing of the dampened MAC is implementation dependent and should not impact or get impacted by the MDAS derivations presented in this document.

7. Security Considerations

This document inherits all the security considerations discussed in [RFC7432].

8. IANA Considerations

TBD as of now.

9. Acknowledgements

The authors of this draft would like to thank Jorge Rabadan, Sergey Fomin and Luc Andre Burdet for their valuable comments.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://www.rfc-editor.org/rfc/rfc2119.txt>.

10.2. Informative References

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for

Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014, <http://www.rfc-editor.org/rfc/rfc7348.txt>.

[RFC7432] Sajassi, A., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015, <http://www.rfc-editor.org/rfc/rfc7432.txt>.

[RFC9014] Rabadan, J., Sathappan, S., Henderickx, W., Sajassi, A., and W. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, May 2021, <http://www.rfc-editor.org/rfc/rfc9014.txt>.

Authors' Addresses

Saumya Dikshit
Aruba Networks, HPE
Email: saumya.dikshit@hpe.com

Vinayak Joshi
Oracle India Pvt Ltd
Email: vinayak.j.joshi@oracle.com

Swathi Shankar
Aruba Networks, HPE
Email: swathi.shankar@hpe.com

BESS WG
Internet-Draft
Intended status: Standards Track
Expires: 31 July 2024

S. Dikshit
Aruba, HPE
V. Joshi
Oracle India Pvt Ltd

28 January 2024

All PEs as DF
draft-saumvinayak-bess-all-df-bum-07

Abstract

The Designated forwarder concept is leveraged to prevent looping of BUM traffic into tenant network sourced across NVO fabric for multihoming deployments. [RFC7432] defines a preliminary approach to select the DF for an ES,VLAN or ES,Vlan Group, panning across multiple NVE's. [RFC8584] makes the election logic more robust and fine grained by inculcating fair election of DF handling most of the prevalent use-cases. This document presents a deployment problem and a corresponding solution which cannot be easily resolve by rules mentioned in [RFC7432] and [RFC8584]. It involves redundant firewall deployment on disparate overlay sites connected over WAN. The requirement is to allow reachability, ONLY, to the local firewall, unless there is an outage. In case of outage the reachability can be extended to remote site's firewall over WAN.

Status of This Memo

Copyright Notice

Table of Contents

1. Important Terms

DF: Designated Forwarder as defined in [RFC7432].

VTEP: Virtual Tunnel End Point or Vxlan Tunnel End Point

2. Introduction

The Designated forwarder concept is leveraged to prevent looping of BUM traffic into tenant network sourced across NVO fabric for multihoming deployments. [RFC7432] defines a preliminary approach to select the DF for an ES,VLAN or ES,Vlan Group, panning across multiple NVE's. [RFC8584] makes the election logic more robust and fine grained by inculcating fair election of DF handling most of the prevalent use-cases. This document presents a deployment problem and a corresponding solution which cannot be easily resolve by rules mentioned in [RFC7432] and [RFC8584]. It involves redundant firewall

deployment on disparate overlay sites connected over WAN. The requirement is to allow reachability, ONLY, to the local firewall, unless there is an outage. In case of outage the reachability can be extended to remote site's firewall over WAN.

3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

When used in lowercase, these words convey their typical use in common language, and they are not to be interpreted as described in [RFC2119].

4. Problem Description

It's a typical deployment case of Firewall devices, also configured as default gateway for NVO fabric. The default gateways inturn redirects traffic to firewalls over a shared vlan. This example, for simplicity, assumes the former case wherein firewall is also configured as default gateway for all VLANs on the site (SITE-1 and SITE-2).

All PEs(Vtep1 and Vtep2 in below example) in the diagram are attached to same ES and both intend to act act as DF for the broadcast domain (BD-1) for their respective sites. As already mentioned, this is a typical case of firewall-gateways (active/active) across fabrics (sites). The preferred firewall-gateway is the one local to the site. All ARP broadcast request generated for the gateway are directed to the local firewall and NOT to the remote one.

Whereas, upon failure of the local firewall, all packets orginating from the affected site, including broadcast packets like ARP requests, need to be redirected (over WAN, via DCI/VPN) towards the remote site firewall. The firewall-device is connected to it's first-hop vtep over the same bridge-domain and same ESI across all sites.

All in all, it's an emulated multi-homing scenario. This is a scenario of firewall devices hosting same(IP and MAC) credentials.

4.1. Problem Example

The following details out the problem further. There are two sites, SITE-1 and SITE-2 in the below diagram. Traffic (including BUM) generated by Host1 (in SITE-1) (for a bridge-domain) should run through site-local firewall instance (firewall_1) preferably and should not be leaked to the remote sites.

Only in case of local-outage, the traffic should be send across over WAN to the remote firewall (firewall_2). Same should apply to traffic generated by Host2 (in SITE-2), wherein, it should ONLY run through the local firewall (firewall_2), unless there is local-firewall. In that case should go over the WAN towards remote sites firewall, firewall_1.

Vtep1 and/or Vtep2 learn the firewall MAC (MAC_F) as a local host learning and also from the remote vteps, Vtep2 and Vtep1, respectively. But since both the learnings are over the same ESI, it should not lead to MAC move. Cometh the local firewall failure, Vtep1 or Vtep2 should start redirecting the traffic to remote SITE firewall, firewall_2 and firewall_1, respectively. Any ARP request (BUM traffic) for firewall credentials landing at either Vtep1 or Vtep2 from the remote fabric, should then be flooded to network or LAN towards the locally connected firewall.

```
         SITE-1          |          SITE-2
-----------------------------------------------------
      Host1                              Host2
       \                                   /
        \                                 /
Vtep_host1                               Vtep_host2
         |                               |
         |       [ EVPN-fabric ]         |
         |                               |
     Vtep1============WAN==============Vtep2
        /                                  \
       /                                    \
Firewall _1                            Firewall_2
  (MAC_F)                               (MAC_F)
```

Figure 1: Figure 1: Active-Active Firewall Across Sites

5. Solution(s)

The control plane part of the solution can be leveraged from the 'DF Election Extended Community' described in [RFC8584]. Since the requirement is to ensure all the PEs attached to ESI, forward the BUM traffic arriving from hosts connected to local NVO fabric towards the Attachment circuits (ACs), that are configured over the ES for a BD (broadcast or bridge domain) mapped to Vlan or bundle of Vlans. As explained in the above section, that this is a case, where PEs are in disparate networks and the ACs behind them are not connected to common physical device, even though they are part of the same ES. The diagram gives an overview of the network or deployment in contention.

This document proposes a new mode of DF-election called 'ALL-PEs-DF', where-in, all the pariticipating PEs, intend to play DF role for subset of vlan(s) enabled on an ESI. This requires "DF Election Extended Community" to carry this information with the ES route to indicate it to remote PEs. This ensures all PEs receiving BUM traffic over NVO fabric destined to ESI, BD, SHOULD flood it on the associated ES on the access/tenant side. A PE device MAY be explicitly configured to choose the ALL-PEs-DF mode.

5.1. Sending All PEs are DF mode

The All-PEs-DF mode is used as follows:

(1) PEs configured to use ALL-PEs-DF mode SHOULD set "DF Alg" algorithm field in 'DF Election Extended Community' to appropriate value.

(2) This document proposes value TBD for All-PEs-DF mode, as values '0', '1' and '2' are already reserved for usage.

(3) This algorithm is agnostic to the values carried in 'Bitmap' but does not discounts any use-case(s) in future which may need extra information carried in 'Bitmap' along with All-PEs-DF mode.

5.2. Receive All PEs are DF mode

When a PE receives the ES routes from all the other PEs, for the ES in question, carrying the ALL-PEs-DF mode set, in 'DF Election Extended Community', it SHOULD, check to see if all the advertisements have the Extended Community with 'All-DF-mode' set as 'DF Algo'. If yes, then PE SHOULD ignore the 'Bitmap' and 'Rsvd' field in the extended community. As also mentioned in [RFC8584] , even if, a single advertisement for Route Type 4 is received without

the locally configured DF Alg and capability, the default DF election algorithm MUST be used as mandated in [RFC7432].

5.3. Example of algorithm

The BGP-EVPN control plane extension, as mentioned in this document, helps in resolving the problem decsribed in Section 4. If PEs, Vtep1 and Vtep2 are configured to use ALL-PEs-DF mode, then any BUM traffic from respective local hosts Host1/Host2 connected to the EVPN fabric, SHOULD get redirected towards the AC for the ESI,Vlan to which the firewall_1/firewall_2 (respectively) is attached. For example the arp-request for the Firewall IP will be honored by the Firewall_1 behind the Vtep1 which receives the ARP-request. Whereas, when Vtep2 receives the arp-request it will be honored by Firewall_2. Vtep1 and Vtep2 will publish the arp-request in their respective ACs attached to the firewall on which Vlan,ESI is configured and enabled

6. Interoperability with other Algos

Since All-DF-algo is special mode and not exactly an algorithm, which requires the participation of all PEs for an ESI, VLAN. Hence, even if one PE publishes an algo which is NOT "All-DF-mode", other PEs SHOULD revert back to default algorithm. The reason being that, if there are PE1, PE2, PE3 and PE4 in contention. PE1 and PE2 publishes DF Algo ’ALL-PEs-DF’, PE3 publishes ’0’ and PE4 publishes ’1’. Once this mismatch is perceived, all PEs SHOULD try and converge towards the default mode. An admin intervention may be required to achieve the same or to converge on any other supported ’DF Algo’.

7. Backward Compatibility

As prescribed in [RFC8584], PEs not supporting (hence not publishing) ’ALL-PEs-DF’, SHOULD ignore the processing of the ’DF Election Extended Community’ and SHOULD indulge in DF-election using the default aglorithm mentioned in [RFC7432]. The PEs configured with this new alogrithm (hence publishing it), if receive Route Type 4 without ’DF Election Extended Community’, SHOULD also revert back to default algorithm. If PEs receive Route Type 4 with another algorithm published in ’DF Election Extended Community’, then it should follow procedures prescribed in Section 6.

8. Impact on Local Bias

There is no impact on the local-bias handling, as the PE receiving the BUM from access side over {ESI, VLAN} and relays it to other PEs that published {ESI, VLAN} in Route Type 4; the receiving side PEs will not relay it to EVPN fabric nor will they redirect it to same ESI configured with same VLAN on the access/tenant side.

9. Security Considerations

This document inherits all the security considerations discussed in [RFC7432] and [RFC8584].

10. IANA Considerations

IANA considerations yet to be concluded as the value of mode proposed here is still under discussion.

11. Acknowledgements

The authors want to thank Jorge Rabadan and Luc Andre for their valuable comment. They also advised on other potential solution while helping in paraphrasing the problem statement.

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://www.rfc-editor.org/rfc/rfc2119.txt>.

12.2. Informative References

[RFC7348] Mahalingam, M., "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014, <http://www.rfc-editor.org/rfc/rfc7348.txt>.

[RFC7432] Sajassi, A., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015, <http://www.rfc-editor.org/rfc/rfc7432.txt>.

[RFC8584] Rabadan, J., "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, April 2019, <http://www.rfc-editor.org/rfc/rfc8584.txt>.

[RFC9014] Rabadan, J., "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, May 2021, <http://www.rfc-editor.org/rfc/rfc9014.txt>.

Authors' Addresses

Saumya Dikshit
Aruba Networks, HPE
Mahadevpura

Bangalore 560 048
Karnataka
India
Email: saumya.dikshit@hpe.com

Vinayak Joshi
Oracle India Pvt Ltd
Bangalore
Karnataka
India
Email: vinayak.j.joshi@oracle.com

BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: 8 September 2022

J. Rabadan, Ed.
S. Sathappan
M. Gautam
Nokia
P. Brissette
Cisco Systems
W. Lin
Juniper
7 March 2022

Domain Path (D-PATH) for Ethernet VPN (EVPN) Interconnect Networks
draft-sr-bess-evpn-dpath-01

Abstract

The BGP Domain PATH (D-PATH) attribute is defined for Inter-Subnet Forwarding (ISF) BGP Sub-Address Families that advertise IP prefixes. When used along with EVPN IP Prefix routes or IP-VPN routes, it identifies the domain(s) through which the routes have passed and that information can be used by the receiver BGP speakers to detect routing loops or influence the BGP best path selection. This document extends the use of D-PATH so that it can also be used along with other EVPN route types.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The BGP Domain PATH (D-PATH) attribute [I-D.ietf-bess-evpn-ipvpn-interworking] is defined for Inter-Subnet Forwarding (ISF) BGP Sub-Address Families that advertise IP prefixes. When used along with EVPN IP Prefix routes or IP-VPN routes, it identifies the domain(s) through which the routes have passed and that information can be used by the receiver BGP speakers to detect routing loops or influence the BGP best path selection. This document extends the use of D-PATH so that it can also be used along with other EVPN route types.

The D-PATH attribute can be used to prevent control plane loops for EVPN routes, or to provide full path visibility of all the EVPN Interconnect Gateways through which a route has gone and modify the best path selection based on it. Some use cases in which D-PATH can be used along with (non-IP Prefix) EVPN routes follow, but the use cases are not limited to the ones described in this section.

1.1. D-PATH to Prevent Loops for EVPN Routes

Figure 1 illustrates an EVPN Interconnect case where EVPN MAC/IP Advertisement routes can be looped indefinitely. The three Gateways (GW1, GW2 and GW3) and PE1 in the diagram are attached to the same EVPN Broadcast Domain (BD1). However, BD1 is extended throughout three different domains that are interconnected by the Gateways, which follow [RFC9014] procedures. Suppose a host with MAC address M1 is learned on GW1 and GW1 advertises an EVPN MAC/IP Advertisement route for M1 into Domain-1 and Domain-2. When the route gets imported by GW2 and GW3 and later exported into Domain-3, GW2 and GW3 may redistribute each other's route for M1 back into Domain-1 and Domain-2, respectively, creating a loop. D-PATH can be used by the Gateways when redistributing the route between Domains, to identify the Domains through which the route for M1 has gone. When GW1 receives an EVPN MAC/IP Advertisement route for M1 that contains a D-PATH with a domain-id locally assigned, GW1 identifies the route as "looped".

```
          +----------------+ GW2
          |   EVPN        +-------+
          |   Domain-1    | +---+ |
          |               | |BD1| |----------------+
          |               | +---+ |                |
     GW1  |               +-------+                |    PE1
   +-------+               |     |    EVPN       +-------+
   | +---+ |----------------+     |    Domain-3   | +---+ |
   | |BD1| |                      |               | |BD1| |
   | +---+ |----------------+     |               | +---+ |
   +---|---+                | GW3 |               +---|---+
       |   |   EVPN        +-------+                |   |
   M1--+   |   Domain-2    | +---+ |                |   +--M2
           |               | |BD1| |----------------+
           |               | +---+ |
           |               +-------+
          +----------------+
```

Figure 1: Loops for EVPN routes

Similar examples are possible with EVPN VPWS services on the Gateways and PEs, where loop prevention for the redistributed A-D per EVI routes is needed. D-PATH provides the end to end path visibility that is required to prevent the loop.

1.2. Add Path Visibility and Influence Best Path Selection for EVPN Routes

Figure 2 illustrates another [RFC9014] EVPN Interconnect case where, in addition to using D-PATH to prevent EVPN MAC/IP Advertisement route loops when redistributing routes between domains, the D-PATH attribute can also influence the best path selection for the routes. For example, if all the Gateways in the diagram are attached to the same BD1, an EVPN MAC/IP Advertisement route for MAC address M1 advertised by GW1 is advertised into Domain-1 and Domain-4. Two routes for M1 will arrive at GW3 with different Route Distinguishers and BGP Next Hops. If D-PATH is used by all the Gateways, the two routes arriving at GW3 will have a different sequence of domain-ids in the D-PATH attribute. GW3 can use the length of the D-PATH as a way of influencing the selection (i.e., the shortest D-PATH route is selected). D-PATH improves the path visibility of the route since it provides information about all the Domains through which the route has passed.

```
          +----------+ GW11  +----------+  GW2  +----------+
          | EVPN     +-------+ EVPN     +-------+ EVPN     |
          | Domain-1 | +---+ | Domain-2 | +---+ | Domain-3 |
          |          | |BD1| |          | |BD1| |          |
          |          | +---+ |          | +---+ |          |
    GW1   |          +-------+          +-------+          |  GW3
  +-------+         |       |          |       |         +-------+
  | +---+ |---------+       +----------+       +---------| +---+ |
  | |BD1| |                                              | |BD1| |
  | +---+ |---------+       +----------------------------| +---+ |
  +---|---+         | GW12  |                            +---|---+
      |   | EVPN     +-------+      EVPN                  |   |
  M1--+   | Domain-4 | +---+ |      Domain-5               |   +--M2
          |          | |BD1| |                            |
          |          | +---+ |                            |
          |          +-------+                            |
          +----------+       +----------------------------+
```

Figure 2: Influence Best Path Selection for EVPN routes

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Terminology

This section summarizes the terminology that is used throughout the rest of the document.

* AC: Attachment Circuit or logical interface associated to a given BT. To determine the AC on which a packet arrived, the PE examines the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).

* BD and BT: a Broadcast Domain and Bridge Table, as defined in [I-D.ietf-bess-rfc7432bis]. A BD is a group of PEs attached to the same EVPN layer-2 multipoint service. A BT is the instantiation of a Broadcast Domain in a PE. When there is a single Broadcast Domain in a given EVI, the MAC-VRF in each PE contains a single BT. When there are multiple BTs within the same MAC-VRF, each BT is associated to a different Ethernet Tag. The EVPN routes specific to a BT indicate which Ethernet Tag the route corresponds to.

* ES and ESI: Ethernet Segment and Ethernet Segment Identifier, as defined in [I-D.ietf-bess-rfc7432bis].

* EVPN Domain: two PEs are in the same EVPN Domain if they are attached to the same service and the packets between them do not require a data path lookup of the inner frame (e.g., in the BT of a MAC-VRF) in any intermediate router. An EVPN Domain Gateway PE is always configured with multiple Domain identifiers (EVPN Domain-ID) in the MAC-VRF or VPWS that connects those EVPN Domains, each EVPN Domain-ID representing an EVPN Domain.

  Example: Figure 3 illustrates an example where PE1 and PE2 belong to different EVPN Domains since packets between them (for flows between hosts with MAC addresses M1 and M2) require a MAC lookup in two of the gateways that are connecting the three EVPN Domains. E.g., if frames from M1 to M2 go through PE1, GW1, GW3 and PE2, a MAC lookup is performed at GW1 and GW3.

```
                              GW1------------GW3
                            +------+       +------+
                +-------------| BD1  |       | BD1  |-------------+
               PE1          +------+       +------+            PE2
             +------+          |              |             +------+
         M1-| BD1  |   EVPN    |     EVPN     |     EVPN    |  BD1 |-M2
             +------+         GW2            GW4            +---+--+
                |            +------+       +------+             |
                +-------------| BD1  |       | BD1  |-------------+
                            +------+       +------+
                               +--------------+
                EVPN Domain 1     EVPN Domain 2  EVPN Domain 3
               <---------------> <------------> <---------------->
```

Figure 3: EVPN Domain Interconnect Example

* EVPN Domain Gateway: a PE where a service (BD or VPWS instance) is instantiated and is attached to two or more EVPN Domains. An example of EVPN Domain Gateway is a PE following the procedures in section 4.4 or section 4.6 of [RFC9014]. In the example in Figure 3, GW1 and GW2 connect EVPN Domains 1 and 2, whereas GW3 and GW4 connect EVPN Domains 2 and 3. GW1 and GW2 import the MAC/IP Advertisement route for M1 coming from the EVPN Domain 1 into the MAC-VRF for BD1, and redistribute it into EVPN Domain 2. Likewise, GW3 and GW4 import the route into their MAC-VRF and re-advertise it into EVPN Domain 3.

* MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [I-D.ietf-bess-rfc7432bis]. It is also the instantiation of an EVI (EVPN Instance) in a PE.

4. Use of Domain Path Attribute (D-PATH) with EVPN routes

This document extends the use of the D-PATH attribute specified in [I-D.ietf-bess-evpn-ipvpn-interworking] so that D-PATH can be advertised and processed along with the following EVPN route types:

* EVPN MAC/IP Advertisement routes that are not used for Inter-Subnet Forwarding (ISF). Note that if the EVPN MAC/IP Advertisement route is used for Inter-Subnet Forwarding as in [RFC9135], the procedures for the D-PATH advertisement and processing are described in [I-D.ietf-bess-evpn-ipvpn-interworking].

* EVPN A-D per EVI routes that are used for EVPN VPWS [RFC8214]. Advertised A-D per EVI routes not used for EVPN VPWS SHOULD NOT contain D-PATH.

* EVPN Inclusive Multicast Ethernet Tag routes [I-D.ietf-bess-rfc7432bis].

As discussed, the use of D-PATH with EVPN IP Prefix routes is specified in [I-D.ietf-bess-evpn-ipvpn-interworking]. When used along with EVPN routes other than IP Prefix routes, the D-PATH attribute is characterized as follows:

1. D-PATH is composed of a sequence of Domain segments following the format specified in [I-D.ietf-bess-evpn-ipvpn-interworking] where each Domain is represented as <DOMAIN-ID:ISF_SAFI_TYPE>. In this specification, DOMAIN-ID is an EVPN Domain identifier configured in an EVPN Domain Gateway and ISF_SAFI_TYPE is set to either 70 (EVPN) or 0 (local route). To simplify the explanation, this document represents the domains for EVPN routes as <Domain-ID:TYPE>.

2. D-PATH identifies the sequence of EVPN Domains the route has gone through, with the last <Domain-ID:TYPE> entry (rightmost) identifying the first PE or EVPN Domain Gateway that added the D-PATH attribute.

3. For non-ISF MAC/IP Advertisement routes or A-D per EVI routes, D-PATH SHOULD be added/modified by a EVPN Domain Gateway that redistributes the route between EVPN Domains and MAY be added by a PE or EVPN Domain Gateway that originates the route, as follows:

   a. An EVPN Domain Gateway that connects two EVPN Domains "X" and "Y", and receives a route on a EVPN Domain identified by a Domain-ID "X" SHOULD append a domain <X:EVPN> to the existing (or newly added) D-PATH attribute when redistributing the route to EVPN Domain "Y". The route is redistributed if it is first imported in a MAC-VRF (or VPWS instance), the MAC (or Ethernet Tag) is active, and policy allows the re-export of the route to a BGP neighbor.

   b. An EVPN Domain Gateway MAY also add the D-PATH attribute for locally learned MACs or MAC/IP pairs. In this case, the domain added would be <A:0>, where "A" is the Domain-ID configured on the Gateway’s MAC-VRF that is specific to local routes (MAC/IP learned via local AC), and "0" is the TYPE of the EVPN Domain and indicates that the route is locally originated and not redistributed after receiving it from a BGP-EVPN neighbor. The EVPN Domain-ID for local routes MAY be shared by all the EVPN Domain Gateways of the same redundancy group for local routes, or each EVPN Domain Gateway of the redundancy group can use its own Domain-ID.

c. A PE that is connected to a single EVPN Domain (therefore the PE is not an EVPN Domain Gateway) MAY add the D-PATH with a domain <B:0>, where "B" is the Domain-ID configured on the PE's MAC-VRF (or VPWS) for locally learned MAC/IPs (or Ethernet Tag IDs for VPWS). "0" is the TYPE that indicates the route is not re-advertised, but originated in the PE.

4. For EVPN Inclusive Multicast Ethernet Tag routes, an EVPN Domain Gateway must not redistribute routes between Domains as specified in [RFC9014]. An EVPN Domain Gateway originates an EVPN Inclusive Multicast Ethernet Tag route per Domain to which the Gateway is attached, so that BUM traffic can be attracted from one Domain to the rest. Therefore, only the above point 3.b. applies to EVPN Domain Gateways. That is, an EVPN Domain Gateway MAY add a <A:0> D-PATH attribute for the Inclusive Multicast Ethernet Tag routes generated for the EVPN Domains, where "A" is the configured local Domain-ID, and "0" is the TYPE that indicates the route is locally originated and not redistributed across EVPN Domains. When two EVPN Domain Gateways of the same redundancy group connect two EVPN Domains "X" and "Y" and D-PATH is used for EVPN Inclusive Multicast Ethernet Tag routes, it is RECOMMENDED to add the D-PATH attribute with the same local Domain-ID and only on "X" or "Y" but not on both Domains.

5. On received EVPN routes, D-PATH is processed and used for loop detection (Section 4.4) as well as to influence the best path selection (Section 4.1, Section 4.2 and Section 4.3).

4.1. D-PATH and Best Path Selection for MAC/IP Advertisement routes

When two (or more) MAC/IP Advertisement routes with the same route key (and same or different RDs) are received, a best path selection algorithm is used to select and install only one route. This section summarizes the best path selection for MAC/IP Advertisement routes, which extends the rules in [I-D.ietf-bess-rfc7432bis] by including D-PATH in the tie-breaking algorithm. While the algorithm may be implemented in different ways, the selection result SHOULD be the same as the result of the rules that follow.

The tie-breaking algorithm begins by considering all EVPN MAC/IP Advertisement routes equally preferable routes to the same destination, and then selects routes to be removed from consideration. The process terminates as soon as only one route remains in consideration.

1. When the Default Gateway extended community is present in some of the routes, the MAC/IP Advertisement routes without the Default Gateway indication are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis].

2. Then the routes with the Static bit set in the MAC Mobility extended community are preferred, and the routes without the Static bit set are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis]. Note that this rule does not apply to routes with the Default Gateway extended community, since these routes SHALL NOT convey the MAC Mobility extended community [I-D.ietf-bess-rfc7432bis]. Therefore if two or more routes with the Default Gateway extended community remain in consideration, the selection process skips this step.

3. Then the routes with the highest MAC Mobility Sequence number are preferred, hence the routes that are not tied for having the highest Sequence number are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis]. If two or more routes with the Default Gateway extended community remain in consideration, the selection process skips this step (for the same reason as in step 2).

4. Then routes with the highest Local Preference are preferred, hence routes that are not tied for having the highest Local Preference are removed from consideration, as defined in [RFC4271].

5. Then routes with the shortest D-PATH are preferred, hence routes not tied for the shortest D-PATH are removed. Routes without D-PATH are considered zero-length D-PATH.

6. Then routes with the numerically lowest left-most Domain-ID are preferred, hence routes not tied for the numerically lowest left-most Domain-ID are removed from consideration.

7. If the steps above do not produce a single route, the rest of the rules after the highest Local Preference in [RFC4271] apply after step 6.

The above selection criteria is followed irrespective of the ESI value in the routes. EVPN Multi-Homing procedures for Aliasing or Backup paths in [I-D.ietf-bess-rfc7432bis] are applied to the selected MAC/IP Advertisement route.

4.2. D-PATH and Best Path Selection for Ethernet A-D per EVI routes

When two (or more) EVPN A-D per EVI routes with the same route key (and same or different RDs) are received for a VPWS, a best path selection algorithm is used. The selection algorithm follows the same steps as in Section 4.1 except for steps 1-3 which do not apply since the Default Gateway and MAC Mobility extended community are irrelevant to the EVPN A-D per EVI routes.

The above selection is followed for A-D per EVI routes with ESI=0. For non-zero ESI routes, the EVPN Multi-Homing procedures in [RFC8214] for Aliasing and Backup path are followed to select the routes (P and B flags are considered for the selection of the routes when sending traffic to a remote Ethernet Segment). If the mentioned Multi-Homing procedures do not produce a single route to each of the remote PEs attached to the same ES, steps 4-7 in Section 4.1 are followed.

4.3. D-PATH and Best Path Selection for Inclusive Multicast Ethernet Tag routes

When two (or more) EVPN Inclusive Multicast Ethernet Tag routes with the same route key (and same or different RDs) are received for a MAC-VRF, a best path selection algorithm is used. The selection algorithm follows the same steps as in Section 4.1 except for steps 1-3 which do not apply.

4.4. Loop Detection

An EVPN route received by a PE with a D-PATH attribute that contains one or more of its locally associated Domain-IDs for the MAC-VRF or VPWS instance is considered to be a looped route. A looped route MUST NOT be redistributed to a different domain and SHOULD be flagged as "looped".

EVPN A-D per EVI looped routes and Inclusive Multicast Ethernet Tag looped routes MUST NOT be installed, where "install" means "create forwarding state" in this document. An EVPN MAC/IP Advertisement looped route MAY be installed if selected as the best route.

For instance, in the example of Figure 3, assuming PE1 advertises M1's MAC/IP and does not add the D-PATH attribute, the EVPN Domain Gateway GW1 receives two MAC/IP Advertisement routes for M1's MAC/IP:

* A MAC/IP Advertisement route with next-hop PE1 and no D-PATH.

* A MAC/IP Advertisement route with next-hop GW2 and D-PATH={length=1,<6500:1:EVPN>}, assuming that the Domain-ID for EVPN Domain 1 is 6500:1.

In this case, EVPN Domain Gateway GW1 flags the MAC/IP Advertisement route with D-PATH as "looped", and does not install the MAC in the BT of the MAC-VRF and does not redistribute the route back to EVPN Domain 1 (since the route includes one of GW1's Domain-IDs). In case the MAC/IP Advertisement route with next-hop PE1 is withdrawn, GW1 may install the route with next-hop GW2 and D-PATH <6500:1:EVPN>; this may help speed up convergence in case of failures.

4.5. Error Handling

The error handling for the D-PATH attribute is described in [I-D.ietf-bess-evpn-ipvpn-interworking]. This document extends the use of D-PATH to non-ISF EVPN routes.

5. Use-Case Examples

This section illustrates the use of D-PATH in EVPN routes with examples.

Figure 4 and Figure 5 illustrate an integrated interconnect solution for an EVPN BD, as described in section 4.4 and section 4.6 of [RFC9014]. GW1 and GW2 are EVPN Domain Gateways connecting two EVPN Domains identified by D-PATH domain {1:1:EVPN} and {1:2:EVPN}, respectively. Received Ethernet A-D routes, ES routes, and Inclusive Multicast routes from the routers in one EVPN Domain are consumed and processed by GW1 and GW2, but not redistributed to the other EVPN Domain. However, MAC/IP Advertisement routes received by GW1 and GW2 in one EVPN Domain are processed and, if installed, redistributed into the other EVPN Domain.

```
           +----EVPN Domain-1---+      +----EVPN Domain-2--+
           |     1:1:EVPN       | GW1  |    1:2:EVPN       |
           |                  +---------+                 |
           |                  | +-----+ |                 |
           |                  | | BD1 | X <-+             |
          PE1                 | +-----+ |   |            PE2
      +---------+             +---------+   |          +---------+
      | +-----+ |              |      |     |          | +-----+ |
   M1-----| BD1 | |              |      |     |          | | BD1 |-----M2
      | +-----+ |  ------->    |      |     |          | +-----+ |
      +---------+ (RT2)M1/IP1   |      |     |          +---------+
           |                  +---------+   |             |
           |                  | +-----+ |   |(RT2)M1/IP1  |
           |                  | | BD1 | | --+ <1:1:EVPN>  |
           |                  | +-----+ |                 |
           |                  +---------+                 |
           |                    | GW2 |                   |
           +---------------------+     +-------------------+
```

Figure 4: EVPN Interconnect

Consider the example of Figure 4, where PE1 advertises a MAC/IP Advertisement route for M1/IP1. The route is processed and installed by GW1 and GW2 in BD1, and both redistribute the routes into the EVPN Domain-2. By using D-PATH in GW1 and GW2, when the route is received on PE2, PE2 has the visibility of the EVPN Domains through which the route has gone, and can also use the D-PATH for best path selection. In addition, GW1 and GW2 can compare the D-PATH of the incoming routes with their local list of EVPN Domain-IDs, and detect looped routes if any of the local EVPN Domain-IDs matches a domain in the received D-PATH. This procedure prevents the redistribution of the route back into EVPN Domain-1. For example, when GW1 receives the MAC/IP Advertisement route for M1/IP1 with D-PATH <1:1:EVPN>, GW1 identifies the route as looped and it does not redistribute it back to Domain-1. The M1/IP1 route with Next Hop PE1 is installed. If M1/IP1 with Next Hop PE1 is withdrawn, GW1 MAY install the route M1/IP1 with Next Hop GW2, as specified in Section 4.4.

The example of Figure 5 illustrates how GW1 and GW2 can also have local ACs in BD1 and learn local MAC (or MAC/IP) addresses from devices connected to the ACs.

```
           +----EVPN Domain-1---+      +----EVPN Domain-2--+
           |     1:1:EVPN       | GW1  |    1:2:EVPN       |
           |                  +---------+                  |
           |                  | +-----+ |                  |
           |             +-->X| | BD1 | |X<--+             |
          PE1            |    | +-----+ |    |            PE2
       +---------+       |    +---------+    |         +---------+
       | +-----+ |       |     |      |      |         | +-----+ |
   M1-----| BD1 | |       |     |      |      +--->     | | BD1 |-----M2
       | +-----+ |       |     |      |      |         | +-----+ |
       +---------+       |     |      |      |         +---------+
           |             |     | GW2  |      |             |
           |          <---+--  +---------+ (RT2)M3/IP3      |
           |        (RT2)M3/IP3 | +-----+ |  {1:3:0}         |
           |         {1:3:0}    | | BD1 | |    |             |
           |                    | +-----+ | ---+             |
           |                  +----|----+                    |
           |                     |  |  |                    |
           +---------------------+  |  +--------------------+
                                    +
                                    M3
```

Figure 5: EVPN Interconnect local AC

Assuming GW2 learns M3/IP3 via local AC, GW2 advertises a MAC/IP Advertisement route for M3/IP3 into each of the EVPN Domains that GW2 is connected to. As described in Section 4, GW2 can advertise these two MAC/IP Advertisement routes with a configured EVPN Domain-ID for local MAC/IPs routes that can be shared with GW1. Consider this EVPN Domain-ID is 1:3 and it is configured on both, GW1 and GW2. When GW2 advertises the route into each EVPN Domain, it adds the D-PATH attribute with a domain {1:3:0}. These routes are flagged by GW1 as "looped" since 1:3 is configured as a local EVPN Domain-ID in GW1. In addition, PE1 and PE2 receive the routes with the D-PATH and they have the visibility of the origin of the routes, in this case local EVPN Domain Gateway routes. This information can be used to influence the best path selection in case of multiple routes for M3/IP3 are received on PE1 or PE2 for BD1.

As an alternative solution to configuring the same EVPN Domain-ID for local routes on both EVPN Domain Gateways, GW2 can be configured with EVPN Domain-ID 1:3 for local routes, and GW1 can use a different EVPN Domain-ID, e.g., 1:4. In this case, GW2 advertises the route for M3/IP3 into each EVPN Domain as before, but now GW1 does not flag the route as "looped" since 1:3 is not on the list of GW1's local EVPN Domain-IDs. GW1 receives the routes from both EVPN Domains, and GW1 selects the route from e.g., EVPN Domain-1. GW1 then installs the route in its BT and redistributes the route into EVPN Domain-2 with

D-PATH {1:1:EVPN, 1:3:0}. When PE2 receives two routes for M3/IP3, one from GW2 with D-PATH {1:3:0} and another from GW1 with D-PATH {1:1:EVPN, 1:3:0}, PE2 uses best path selection and choose to send its traffic to GW2. Also GW2 receives the route for M3/IP3 from GW1 and mark it as "looped" since that route conveys its own EVPN Domain-IDs 1:1 and 1:3.

In a nutshell, the use of D-PATH in MAC/IP Advertisement routes helps prevent loops and influences the best path selection so that PEs choose the shortest paths to the destination PEs.

6. Security Considerations

Most of the considerations included in [I-D.ietf-bess-evpn-ipvpn-interworking] apply to this document.

7. IANA Considerations

None.

8. Acknowledgments

9. Contributors

10. References

10.1. Normative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[I-D.ietf-bess-evpn-ipvpn-interworking]
Rabadan, J., Sajassi, A., Rosen, E., Drake, J., Lin, W., Uttaro, J., and A. Simpson, "EVPN Interworking with IPVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-ipvpn-interworking-06, 22 September 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-ipvpn-interworking-06.txt>.

[I-D.ietf-bess-rfc7432bis]
Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", Work in Progress, Internet-Draft, draft-ietf-bess-rfc7432bis-03, 28 February 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-rfc7432bis-03.txt>.

[RFC9014] Rabadan, J., Ed., Sathappan, S., Henderickx, W., Sajassi, A., and J. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, DOI 10.17487/RFC9014, May 2021, <https://www.rfc-editor.org/info/rfc9014>.

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

10.2. Informative References

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/info/rfc9135>.

Authors' Addresses

J. Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: jorge.rabadan@nokia.com

S. Sathappan
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: senthil.sathappan@nokia.com

M. Gautam
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: mallika.gautam@nokia.com

P. Brissette
Cisco Systems
Canada
Email: pbrisset@cisco.com

W. Lin
Juniper
United States of America
Email: wlin@juniper.net

BESS
Internet-Draft
Intended status: Standards Track
Expires: 16 March 2024

P. Thubert, Ed.

A. Przygienda
Juniper Networks, Inc

J. Tantsura
Microsoft

13 September 2023

Secure EVPN MAC Signaling
draft-thubert-bess-secure-evpn-mac-signaling-04

Abstract

This specification adds attributes to EVPN to carry IPv6 address metadata learned from RFC 8505 and RFC 8928 so as to maintain a synchronized copy of the 6LoWPAN ND registrar at each EVPN router and perform locally a unicast IPv6 ND service for address lookup and duplicate address detection.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

"Registration Extensions for IPv6 over 6LoWPAN Neighbor Discovery" [RFC8505] (ND) provides a zeroconf routing-agnostic Host-to-Router Link-Local interface for Stateful Address Autoconfiguration. "Address-Protected Neighbor Discovery for Low-Power and Lossy Networks" [RFC8928] (AP-ND) adds a zeroconf anti-theft protection that protects the ownership of the autoconfigured address with autoconfigured proof of ownership called a Registration Ownership Verifier (ROVR).

[RFC8505] enables the host to claim an IPv6 address and obtain reachability services for that address. It is already used to inject host routes in RPL [RFC9010] and RIFT "Routing in Fat Trees" [RIFT], and to maintain a proxy-ND state in a backbone router [RFC8929]; this specification extends its applicability to the case of Ethernet Virtual Private Network (EVPN).

[RFC8505] specifies a unicast address registration mechanism that enables the host called a 6LowPAN Node (6LN) to install a ND binding state in the 6LowPAN Router (6LR) that can serve as Neighbor Cache Entry (NCE), though it is not operated as a cache. The protocol provides the means to reject the registration in case of address duplication. It also enables to discriminate mobility from multihoming. [RFC8928] adds the capability to verify the ownership of the address and prevent an attacker from stealing and/or impersonating an address.

[RFC8505] defines the 6LoWPAN Border Router (6LBR) as an abstract address registrar that provides authoritative service for Address Registration and duplicate detection. The 6LBR stores address metadata that is obtained during the Address Registration, including an owner ID and a sequence counter. As part of the process of a new Address Registration, the 6LR queries the 6LBR for existing metadata related to the address being registered. This enables in particular to detect a duplication and reject the registration. This specification extends the 6LBR abstract data model to store the Link Layer Address (LLA) of the Registering Node. This enables the 6LBR to perform locally, and using unicast communication, the IPv6 ND services of address lookup and duplicate address detection.

The [RFC8505] address registrar can be centralized, but it can also be distributed and maintained synchronized using a routing protocol. This specification adds attributes to EVPN to carry the IPv6 address metadata learned from [RFC8505] so as to maintain a synchronized copy of the 6LBR abstract data at each EVPN router.

2. Terminology

2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2.2. Glossary

This document uses the following acronyms:

6CIO Capability Indication Option [RFC7400]
6LN: 6LoWPAN Node (the Host) [RFC6775]
6LR: 6LoWPAN router (the router) [RFC6775]
6LBR: 6LoWPAN Border router [RFC6775]
AMC: Address Mapping Confirmation [UNICAST-LOOKUP]
AMR: Address Mapping Request [UNICAST-LOOKUP]
ARO Address Registration Option [RFC6775]
CIPO: Crypto-ID Parameters Option
DAD: Duplicate Address Detection [RFC4862]
ICMPv6: Internet Control Message Protocol for IPv6
DAC Duplicate Address Confirmation [RFC6775]
DAR Duplicate Address Request [RFC6775]
EDAC Extended Duplicate Address Confirmation [RFC8505]
EDAR Extended Duplicate Address Request [RFC8505]
EARO: Extended Address Registration Option [RFC8505]
EVPN: Ethernet VPN [RFC7432]
LLA: Link-Layer Address (the MAC address on Ethernet)
LLN Low-Power and Lossy Network [RFC6550]
NA: Neighbor Advertisement [RFC4861]
NCE: Neighbor Cache Entry [RFC4861]
ND: Neighbor Discovery [RFC4861]
NDPSO: Neighbor Discovery Protocol Signature Option
NS: Neighbor Solicitation [RFC4861]
RA: Router Advertisement [RFC4861]
ROVR: Registration Ownership Verifier [RFC8505]
TID: Transaction ID (a sequence counter in the EARO) [RFC8505]
SLAAC: Stateless Address Autoconfiguration [RFC4862]
SLLAO: Source Link-Layer Address Option [RFC4861]
TLLAO: Target Link-Layer Address Option [RFC4861]
ROVR MAC: MAC obtained from a host meeting requirements in Section 5
Validated ROVR MAC: ROVR MAC validated by procedures specified in [RFC8928]
ROVR Node: EVPN node capable of advertising ROVR MACs
non-ROVR Node: EVPN node not supporting extensions defined in this

document.

VPN: Virtual Private Network

2.3. References

This document uses the terms Clos fabric and Fat Tree interchangeably, to refer to a folded spine-and-leaf topology as defined in the terminology section of "RIFT: Routing in Fat Trees" [RIFT].

The term "leaf" represents the access switch that connects the servers to the Fat Tree. The leaf is typically a Top-of-Rack (ToR) switch.

This specification uses the terms 6LN, 6LR and 6LBR to refer specifically to nodes that implement the said roles in [RFC8505] and does not expect other functionality such as 6LoWPAN Header Compression:

* In the context of this document, the 6LN is a server that advertises an address mapping using [RFC8505], and optionally protects its ownership with [RFC8928].

* The 6LR and 6LBR function are collapsed at the leaf and its state is synchronized with that of the EVPN functional support using an internal interface that is out of scope. That interface could be "pull" meaning that the 6LBR fetches the EVPN information when it needs it, or "push", meaning that any information that EVPN distributes is immediately fed in all the 6LBRs in all the leaves. Note that this is pure control plane and is not subject to abbreviating optimization as the FIB may be.

In this document, readers will encounter terms and concepts that are discussed in the following documents:

EVPN: "BGP MPLS-Based Ethernet VPN" [RFC7432] and "Network Virtualization Overlay Solution" [RFC8365],

Classical IPv6 ND: "Neighbor Discovery for IP version 6" [RFC4861] and "IPv6 Stateless Address Autoconfiguration" [RFC4862],

6LoWPAN ND: Neighbor Discovery Optimization for Low-Power and Lossy Networks [RFC6775], "Registration Extensions for 6LoWPAN Neighbor Discovery" [RFC8505], "Address Protected Neighbor Discovery for Low-power and Lossy Networks" [RFC8928], and "IPv6 Backbone Router" [RFC8929].

3. 6LoWPAN Neighbor Discovery

6LoWPAN Neighbor Discovery defines a stateful address autoconfiguration mechanism for IPv6. 6LoWPAN ND enables to divorce the L3 abstractions for link and subnet from the characteristics of the L2 link and broadcast domain. It is applicable beyond its original field of IoT to any environment where the broadcast nature of the underlaying network should not be exploited, e.g., in the case of a wireless link where broadcast uses an excessive amount of spectrum, and a distributed cloud, where it may span too widely.

In contrast to Stateless Address Autoconfiguration (SLAAC) [RFC4862] which relies on broadcast for duplicate address detection (DAD) and address lookup, 6LoWPAN ND installs and maintains a state in the neighbors for the duration of their interaction. Though it is also called a Neighbor Cache Entry (BCE) in [RFC6775], and in contrast with the the BCE in SLAAC, that state is not a cache that can be casually flushed and rebuilt. It must be installed proactively and refreshed periodically to maintain the connectivity and enable unicast-only operations.

This section goes through the 6LoWPAN ND network abstractions and mechanisms that this specification leverages, as a non-normative reference to the reader. The relevant normative text is to be found in [RFC6775], [RFC8505], and [RFC8928].

3.1. IPv6 Interface, Link, and Subnet

The typical abstraction for an IP Link with 6LoWPAN ND is a logical point-to-point (P2P) link between a node (a host or a router) and a router, regardless of the physical medium between the node and the router, which may or may not be shared with other nodes.

A Subnet is deployed over a mesh of nodes connected with those logical P2P Links, where routers form a connected dominating set as represented in Figure 1; the resulting aggregate is called a multilink subnet (MLSN). An MLSN may be only partially meshed, and the underlaying network is not expected to provide a multicast or a broadcast service across the subnet.

```
                  +------+                     +------+
                  | Host |----------+    +-----| Host |
                  +------+          |    |     +------+
                      |           +--------+
                      |   +-------| Router |
                      |   |       +--------+
                   +--------+        |   |       +------+
                   | Router |        |   +------ | Host |
                   +--------+        |           +------+
                      |   |   +--------+
                      |   +---| Router |
                      |       +--------+
                    +------+     |  |     +------+
                    | Host |-----+  +-----| Host |
                    +------+                +------+
```

Figure 1: P2P Links in a Multilink Subnet

Consequently, the subnet model is not-on-link, meaning that the any-to-any connectivity across the subnet is ensured through L3 operations (routing or proxy) as opposed to transitive (any-to-any) reachability from L2. It also means that hosts do not lookup other nodes using IPv6 Neighbor Discovery but forward all their traffic via their connected routers. Which in turn means that only routers need to be discovered, which is done by sending Router Advertisement (RA) messages to all directly reachable nodes in the subnet, e.g., using a radio broadcast.

As illustrated in Figure 2, an IP interface bundles multiple sub-interfaces to connect the IP links between this node and peers in the same subnet, which is known as a point-to-multipoint (P2MP) interface.

```
+---------------------
| P2MP Interface
| --------------
|  MAC Address
|  IPv6 Link-Local address(es)
|  IPv6 global addresses
|
|   +------------------------
|   | P2P sub-interface
|   | -----------------
|   |  Peer MAC Address
|   |  Peer IP addresses
|   +------------------------
|
|   +------------------------
|   | P2P sub-interface
|   | -----------------
|   |  Peer MAC Address
|   |  Peer IP addresses
|   +------------------------
|
|   +------------------------
|   | P2P sub-interface
|   | -----------------
|   |  Peer MAC Address
|   |  Peer IP addresses
|   +------------------------
|
|    ....
|
+---------------
```

Figure 2: P2MP Interface

In the case of a 6LoWPAN radio, the IP Interface may be physical, and the P2P IP links are virtual based on discovered neighbor routers; the same model can apply when the node is connected via a switch to one or more routers.

In the case of a multihomed NIC card in a datacenter, the NIC is connected to several Top-of-Rack (ToR) switches acting as leaves in the fabric, over as many Ethernet physical interfaces. If the NIC provides a L2 virtual switch, then the stack can apply the same model as above, modeling the virtual port to the virtual switch as a P2MP interface.

On the other hand, if the NIC provides a virtual router, then Ethernet ports are L3 ports and the physical link to the leaf is modeled as P2P. To form the P2MP interface, the router bundles (aggregates) the physical interfaces as the sub-interfaces of a single logical P2MP Link, as shown in Figure 3.

```
+---------------------
| NIC Aggregate interface
| -----------------------
|  virtual MAC Address
|  IPv6 global addresses
|
|   +------------------------
|   |  Ethernet Port 1
|   | -----------------
|   |  Physical MAC Address
|   |  IPv6 Link-Local address(es)
|   |  Leaf MAC Address
|   |  Leaf IP addresses
|   +------------------------
|
|   +------------------------
|   | Ethernet Port 2
|   | -----------------
|   |  Physical MAC Address
|   |  IPv6 Link-Local address(es)
|   |  Leaf MAC Address
|   |  Leaf IP addresses
|   +------------------------
|
|    ....
|
+---------------
```

Figure 3: Logical P2MP Interface

To conserve the same model, it makes sense to configure the same (virtual) MAC address on all the physical interfaces, and use it for the purpose of IPv6 ND. In that case, the same MAC address is exposed as Link-layer Address (LLA) to both leaves for the NIC IP addresses, and the IPv6 address still appears as unicast. Note that the Link-Local addresses used to register the global IPv6 addresses to the leaf may be different but that does not affect the exposed mapping.

When that is not possible, then the same IP address is advertised with the physical MAC address of each port as the LLA over that port. In that case, the global IPv6 address appears as anycast, and SHOULD be advertised as such, more in Section 3.5.

3.2. RFC 6775 Address Registration

The classical "IPv6 Neighbor Discovery (IPv6 ND) Protocol" [RFC4861] [RFC4862] was defined for serial links and transit media such as Ethernet. It is a reactive protocol that relies heavily on multicast operations for Address Discovery (aka Lookup) and Duplicate Address Detection (DAD).

"Neighbor Discovery Optimizations for 6LoWPAN networks" [RFC6775] adapts IPv6 ND for operations over energy-constrained LLNs. The main functions of [RFC6775] are to proactively establish the Neighbor Cache Entry (NCE) in the 6LR and to prevent address duplication. To that effect, [RFC6775] introduces a new unicast Address Registration mechanism that contributes to reducing the use of multicast messages compared to the classical IPv6 ND protocol.

[RFC6775] defines a new Address Registration Option (ARO) that is carried in the unicast Neighbor Solicitation (NS) and Neighbor Advertisement (NA) messages between the 6LoWPAN Node (6LN) and the 6LoWPAN router (6LR). It also defines the Duplicate Address Request (DAR) and Duplicate Address Confirmation (DAC) messages between the 6LR and the 6LBR. In a Low-Power and Lossy Network (LLN), the 6LBR is the central repository of all the Registered Addresses in its domain and the authoritative source of truth for uniqueness and ownership.

3.3. RFC 8505 Extended Address Registration

"Registration Extensions for 6LoWPAN Neighbor Discovery" [RFC8505] updates RFC 6775 into a generic Address Registration mechanism that can be used to access services such as routing and ND proxy. To that effect, [RFC8505] defines the Extended Address Registration Option (EARO), shown in Figure 4:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     Length    |    Status     |    Opaque     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Rsvd | I |R|T|     TID       |     Registration Lifetime     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
...             Registration Ownership Verifier                 ...
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: EARO Option Format

3.3.1. R Flag

[RFC8505] introduces the R Flag in the EARO. The Registering Node sets the R Flag to indicate whether the 6LR should ensure reachability for the Registered Address. If the R Flag is set to 0, then the Registering Node handles the reachability of the Registered Address by other means. In an EVPN network, this means that either it is a RAN that injects the route by itself or that it uses another EVPN router for reachability services.

This document specifies how the R Flag is used in the context of EVPN. An EVPN Host that implements the 6LN functionality from [RFC8505] requires reachability services for an IPv6 address if and only if it sets the R Flag in the NS(EARO) used to register the address to a 6LR acting as an EVPN border router. Upon receiving the NS(EARO), the EVPN router generates a BGP advertisement for the Registered Address if and only if the R flag is set to 1.

[RFC9010] specifies that the ’R’ flags is set in the responded NA messages if and only if the route was installed. This specification echoes that behavior.

3.3.2. TID, "I" Field and Opaque Fields

When the T Flag is set to 1, the EARO includes a sequence counter called Transaction ID (TID), that is needed to format the MAC Mobility Extended Community. This is the reason why the support of [RFC8505] by the Host, as opposed to only [RFC6775], is a prerequisite for this specification); this requirement is fully explained in Section 5.1. The EARO also transports an Opaque field and an associated "I" field that describes what the Opaque field transports and how to use it.

This document specifies the use of the "I" field and the Opaque field by a Host.

3.3.3. Status

The values of the EARO status are maintained by IANA in the Address Registration Option Status Values subregistry [IANA-EARO-STATUS] of the Internet Control Message Protocol version 6 (ICMPv6) Parameters registry.

[RFC6775] and [RFC8505] defined the original values whereas [RFC9010] reduced range to 64 values and reformatted the octet field to enable to transport an external error, e.g., coming form a routing protocol.

This specification uses the format expressed in [RFC9010]. The value of 0 denotes an unqualified success, 1 indicates an address duplication, 3 a TID value that is outdated, and 4 is used in an asynchronous NA to indicate that 6LN should remove that address and possibly form new ones.

3.3.4. Route Ownership Verifier

Section 5.3 of [RFC8505] introduces the Registration Ownership Verifier (ROVR) field of variable length from 64 to 256 bits. The ROVR is a replacement of the EUI-64 in the ARO [RFC6775] that was used to identify uniquely an Address Registration with the Link-Layer address of the owner but provided no protection against spoofing.

"Address Protected Neighbor Discovery for Low-power and Lossy Networks" [RFC8928] leverages the ROVR field as a cryptographic proof of ownership to prevent a rogue third party from registering an address that is already owned. The use of ROVR field enables the 6LR to block traffic that is not sourced at an owned address.

This specification does not address how the protection by [RFC8928] could be extended for use in EVPN. On the other hand, it adds the ROVR to the BGP advertisement to share the state with the other routers via the Reflector (see Section 6.1), which means that the routers that are aware of the Host route are also aware of the ROVR associated to the Target Address, whether it is cryptographic and should be verified.

3.4. RFC 8505 Extended DAR/DAC

[RFC8505] updates the DAR/DAC messages to EDAR/EDAC messages to carry the ROVR field. The EDAR/EDAC exchange takes place between the 6LR to which the node registers an address, and the abstract 6LBR that stores the reference value for the ROVR and the TID associated to that address. It is triggered by an NS(EARO) message from a 6LN to the 6LR, to create, refresh, compare and delete the corresponding state in the 6LBR.

In the status returned with the EDAC message, the 6LBR indicates if the registration is accepted, should be challenged, or is duplicate. The status of 0 (success) indicates that the address is either new or that the current registration matches, and in particular that the ROVR at the 6LBR and the one in the EDAR message are identical.

```
6LN                6LR                 6LBR
 |                  |                    |
 |   IP Link        |    subnetwork      |
 |                  |                    |
 |  NS(EARO)        |                    |
 |----------------->|                    |
 |                  |                    |
 |                  |    EDAR            |
 |                  |------------------->|
 |                  |                    |
 |                  |      EDAC(status)  |
 |                  |<-------------------|
 |                  |                    |
 |       NA(EARO)   |                    |
 |<-----------------|                    |
 |                  |                    |
```

Figure 5: EDAR/EDAC flow

The EDAR/EDAC exchange is protected by the retry mechanism specified in Section 8.2.6 of [RFC6775], though in a data center, a duration significantly shorter than the default value of the Retransmission Timer [RFC4861] of 1 second may be sufficient to cover the round-trip delay between the 6R and the 6LBR.

With this specification, the 6LBR is distributed across the leaves, and all the leaves where an address is currently registered maintain a full 6LBR state for the address, aka local state in the following text. The specification leverages the EDAR/EDAC exchange to ensure that a leaf (acting as a 6LR) that needs to create a 6LBR state for a

new registration has the same value for the ROVR as any 6LBR already serving that address on another leaf. At the same time, the specification avoids placing full ROVR information in BGP so 1) it is not observable by a potential attacker and 2) the new attributes remain reasonably small.

3.5. Anycast and Multicast Addresses

"IPv6 Neighbor Discovery Multicast Address Registration" [I-D.ietf-6lo-multicast-registration] extends [RFC8505] to enable the address registration of IPv6 anycast and multicast addresses. From the host perspective, the registration is very similar to that of unicast addresses, but for a flag in the EARO that signals that the address is multicast or anycast.

This procedure can be used as a replacement to "Multicast Listener Discovery Version 2 (MLDv2) for IPv6" [RFC3810] for source-independant multicast operation. As for unicast, the method saves the need for the host to listen to pollings from the router, and allows the host to sleep for periods of time.

3.6. Prefix Registration

"IPv6 Neighbor Discovery Prefix Registration" [I-D.ietf-6lo-prefix-registration] provides another extension of [RFC8505] to enable the registration of prefixes. The registration is agnostic to the method that the Registering Node used to obtain the prefix (DHCP-PD, IPAM, manual configuration, or autoconfiguration).

As for addresses, the Registering Node may request that the router ensures reachability for the Registered Prefix, typically by redistributing the prefix in a routing protocol. The registration is also agnostic to the routing protocol that is being used.

The Registered Prefix may be owned by the Registering Node, for instance if the prefix is used for a tethered stub or iwithin the node to address applications and / or virtual machines. In that case, the expectation for the routing fabric is to forward packets with a destination address that matches the prefix towards the Registering Node.

Alternatively, the Registered Prefix may be signaled as topologically correct at the Registering Node. This means that the Registering Node can relay packets that are sourced in the Registered Prefix to the outside, and the packets will be not be filtered by the application of [BCP38]. This is used to announce a Provicer Assigned (PA) address range and ensure that packets exit the network via the

ISP that provided the prefix. In that case, the expectation for the routing fabric is to forward packets destined to the outside and with with a source address that matches the prefix towards the Registering Node.

3.7. RFC 7400 Capability Indication Option

"6LoWPAN-GHC: Generic Header Compression for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)" [RFC7400] defines the 6LoWPAN Capability Indication Option (6CIO) that enables a node to expose its capabilities in router Advertisement (RA) messages.

[RFC8505] defines a number of bits in the 6CIO, in particular:

L: Node is a 6LR.
E: Node is an IPv6 ND Registrar -- i.e., it supports registrations based on EARO.
P: Node is a Routing Registrar, -- i.e., an IPv6 ND Registrar that also provides reachability services for the Registered Address.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |   Length = 1  |     Reserved      |D|L|B|P|E|G|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Reserved                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 6: 6CIO flags

A 6LR that provides reachability services for a Host in an EVPN network as specified in this document includes a 6CIO in its RA messages and set the L, P and E flags to 1 as prescribed by [RFC8505].

4. Extending 6LoWPAN ND

4.1. Use of the R flag in NA

This document extends [RFC8928] and [RFC8505] as follows

This document also updates the behavior of a 6LR acting as EVPN router and of a 6LN acting as Host in the 6LoWPAN ND Address Registration as follows:

* The use of the R Flag is extended to the NA(EARO) to confirm whether the route was installed.

4.2. Distributing the 6LBR

This specification enables to distribute the 6LBR at the edge of the EVPN network and collapse the 6LBR function with that of the EVPN support. In that model, the EVPN to 6LBR interaction becomes an internal interface, where each side informs the other in case of new information concerning an IP to Link-Layer Address (LLA) mapping. Since this is an internal interface, this specification makes no assumption on whether the 6LBR stores its own representation of the full EVPN state, which means that the EVPN support informs the 6LBR in case of any change on the EVPN side (this is called the push model, see Figure 13), or if the 6LBR queries the EVPN support when it does not have a mapping to satisfy a request (pull model, see Figure 12).

This specification leverages [RFC8929] that augments the abstract data model of the 6LBR to store the LLA associated with the registered address. Based on that additional state, the 6LBR in a leaf can communicate the mapping to the collocated EVPN function and respond to unicast address mapping lookups from the server side.

In an environment where the server ranges from a classical host to a more complex platform that runs a collection of virtual hosts interconnected by a virtual switch, but where the host-to-leaf interface remains at layer 2, the 6LR and the 6LBR functions can be collapsed in the leaf. The 6LR to 6LBR interaction also becomes an internal interface, and there is no need for EDAR/EDAC messages.

In that case, the MAC address associated to the Registered Address is indicated in the Target Link-Layer Address Option (TLLAO) in the NS message used for the registration, as shown in Figure 7. In the case of a pull model, if the 6LBR does not have a local state for the mapping, it queries the EVPN support to obtain the EVPN state if any. If a mapping is known then the 6LR/6LBR evaluates the registration for address duplication and other possible issues per [RFC8505]. Else (this is for a new mapping), if the registration is accepted, then the 6LBR notifies the EVPN support to inject a route type 2 in the fabric.

```
              Server              Leaf
         +--------------+   +-------------------+
         |              |   |                   |
          6LN                6LR/6LBR       EVPN
           |                |              |
           | <vSwitch> Ethernet | <call I/F> |
           |                |              |
           |  NS(EARO)      |              |
           |------------------->|          |
           |                |              | ^
           |                |   query      | |
           |                |----------->| | if pull
           |                |  response    | | model
           |                |<-----------| |
           |                |              | v
           |            Evaluation (DAD)   |
           |                |              |
           |                |new mapping |
           |                | indication |
           |                |----------->|
           |                |              | Inject/maintain
           |         store a mapping state | EVPN route type 2
           |                |              | ------------------>
           |       NA(EARO)     |          | [via BGP signaling]
           |<-------------------|          |
           |                |              |
```

Figure 7: Direct Registration

In another type of deployment, the 6LR may be a virtual router in the server whereas the 6LBR runs in the leaf node. To address that case, the EDAR/EDAC may be used to communicate as shown in figure 5 of [RFC8505]. This draft leverages the capability to insert IPv6 ND options in the EDAR and EDAC messages introduced in [RFC8929] to place a TLLAO that carries the MAC address associated to the Registered address in the EDAR and EDAC messages as shown in Figure 8:

```
            Server                        Leaf
      +----------------+            +----------------+
      |                |            |                |
       6LN          6LR             6LBR        EVPN
        |             |               |             |
        | <vSwitch>   |  Ethernet     | <call I/F>  |
        |             |               |             |
        |  NS(EARO)   |               |             |
        |----------->|               |             |
        |             |  EDAR(TLLAO)  |             |
        |             |-------------->|             |
        |             |               |             | ^
        |             |               |   query     | |
        |             |               |----------->| | if pull
        |             |               |  response   | | model
        |             |               |<-----------| |
        |             |                             | v
        |             |    Evaluation (DAD)         |
        |             |                             |
        |             |               |new mapping  |
        |             |               | indication  |
        |             |               |----------->|
        |             |                             | Inject/maintain
        |             |   store a mapping state     | EVPN route type 2
        |             |                             | ------------------>
        |             |  EDAC(TLLAO)  |             | [via BGP signaling]
        |             |<--------------|             |
        |  NA(EARO)   |               |             |
        |<-----------|               |             |
        |             |               |             |
```

Figure 8: leveraging EDAR

[RFC8505] updates the DAR/DAC messages into the Extended DAR/DAC to carry the ROVR field. With this specification, the abstract 6LBR is distributed in all the Leaf nodes and synchronized with EVPN. When a server successfully registers an address to a leaf, the 6LR on that leaf becomes 6LBR for that address. It stores the full state for that address including the ROVR and the TID. When the address registration moves to another leaf, an EDAR/EDAC flow between the 6LR in the new leaf and the 6LBR in the old leaf confirms that the ROVR in the NS(EARO) received at the new leaf is correct, in which case the 6LR in the new leaf becomes 6LBR.

When the address is already registered to the local leaf, the EDAR/EDAC exchange is either local between a virtual router in the server and the leaf, or internal to the leaf between a collapsed 6LR and

6LBR. Based on its local state, the 6LBR in the leaf checks whether the proposed address/route is new and legit, and can reject it otherwise.

It results that duplicate addresses and address impersonation attacks can be filtered at the level of IPv6 ND by the 6LBR before the information reaches EVPN.

4.3. Unicast Address Lookup with the 6LBR

A classical IPv6 ND stack in the server that treats the subnet prefix as on-link (more in section 4.6.2. of [RFC4861]), will resolve an unknown LLA mapping with a multicast NS(lookup) message addressed to the solicited node multicast address (SNMA) associated with the destination address being resolved. The RECOMMENDED operation in that case is for the 6LBR that has a mapping state to forward the packet as a unicast MAC to the LLA that is stored for the IPv6 address as expected by [RFC6085]. The actual owner of the address can then answer unicast with a NA message, setting the override (O) bit to 1, as shown in Figure 9.

```
Local             Local                    Remote
Server            Leaf                     Server
+----+          +--------+                 +----+
|    |          |        |                 |    |
 6LN             6LR/6LBR                   6LN
  |                |                         |
  |    Ethernet    |                         |
  |                |     [via EVPN    ]      |
  |    multicast   |     [Data Tunnels]      |
  |    NS(lookup)  |                         |
  |---------------->|                        |
  |                |    forward unicast      |
  |                |     NS(lookup)          |
  |                |----------------------->|
  |                |                         |
  |                |         NA(O)           |
  |                |<-----------------------|
  |      NA(O)     |                         |
  |<----------------|                        |
  |                |                         |
```

Figure 9: Forwarding legacy NS (Lookup)

Section 3.1. of [RFC8929] adds the capability to insert IPv6 ND options in the EDAR and EDAC messages. This enables the 6LBR to store the link-layer address associated with the Registered Address and to serve as a mapping server. [UNICAST-LOOKUP] leverages that state to define a new unicast address lookup operation, extending the EDAR and EDAC messages as the Address Mapping Request (AMR) and Confirmation (AMC) with a different Code Prefix [RFC8505].

In that model, the router advertises the subnet prefix as not on-link by setting the L flag to 0 in the Prefix Information Option (PIO), more in section 4.6.2. of [RFC4861]. The expected behavior is that the host that communicates with a peer in the same subnet refrains from resolving the address mapping and passes the packets directly to the router.

In the case where the router is a virtual 6LR running in the server, and the source and destination are in the same subnet served by EVPN, the router then resolves the address mapping on behalf of the host. To that effect, the router sends a unicast AMR message to the 6LBR. The message contains the SLLAO of the router to which the 6LBR will reply. If the binding is found, the 6LBR replies with an AMC message that contains the TLLOA with the requested MAC address, as shown in Figure 10.

```
                 Local                  Local             Remote
                 Server                  Leaf             Server
           +----------------+       +---------+          +----+
           |                |       |         |          |    |
            6LN          6LR         6LBR/EVPN            6LN
             |            |               |                |
             |            |               |  [via EVPN    ]|
              <vSwitch>    Ethernet       [Data Tunnels]  |
             |            |               |                |
             |            |               |                |
              RA(PIO)     |               |                |
             not onlink   |               |                |
             <-----------|               |                |
             |            |               |                |
              Packet      |               |                |
             ----------->|               |                |
             |            |               |                |
             |            AMR (SLLAO)     |                |
             |            ------------->|                |
             |            |               |                |
             |            AMC (TLLAO)     |                |
             |            <-------------|                |
             |            |               |                |
             |    NCE in STALE state      |                |
             |            |               |                |
             |            |            Packet              |
             |            ------------------------------->|
             |            |               |                |
             |    NCE in DELAY state      |                |
             |            |               |                |
```

Figure 10: Unicast Lookup from the virtual Host

If it is not found, [UNICAST-LOOKUP] provides the capability to indicate immediately that the mapping is not known with a "not found" status in the AMC, as opposed to waiting for an NS(lookup) and retries to time out per [RFC4861].

In a fully stateful subnet where all nodes register all their addresses with [RFC8505], this means that the looked up address is not present in the network; in that case the packet is dropped and an ICMP error type 1 "Destination Unreachable" code 3 "Address unreachable" [RFC4443] is returned as shown in Figure 11.

```
                      Local                 Local
                      Server                 Leaf
                 +----------------+       +---------+
                 |                |       |         |
                  6LN          6LR         6LBR/EVPN
                   |            |               |
                   |            |               |
                   | <vSwitch>    Ethernet      |
                   |            |               |
                   |            |               |
                   |   RA(PIO   |               |
                   |not on-link)|               |
                   |<-----------|               |
                   |            |               |
                   |  Packet    |               |
                   |----------->|               |
                   |            |               |
                   |            | AMR (SLLAO)   |
                   |            |------------->|
                   |            |               |
                   |            | AMC(status=   |
                   |            | "Not Found")  |
                   |            |<-------------|
                   |ICMPv6 Error|               |
                   | "Address   |               |
                   |Unreachable"|               |
                   |<-----------|               |
                   |         Drop Packet        |
                   |            |               |
```

Figure 11: Unicast Lookup failure

Note that the figures above make no assumption on the pull vs. push model. In the case of pull model, the 6LBR queries the EVPN support when it does not have the mapping information to satisfy a request. Figure 12 illustrates a successful pull model lookup flow, when the route type 2 for the mapping is already known on the EVPN side.

```
            Server                        Leaf
      +----------------+            +----------------+
      |                |            |                |
       6LN          6LR              6LBR        EVPN
        |            |                |            |
        | <vSwitch>  |   Ethernet     | <call I/F> |
        |            |                |            |
        |            |                |            |
        |            |                |            | Receive EVPN
        |            |                |            | route type 2 for
        |            |                |            | remote address A'
        |            |                |            | [via BGP signaling]
        |            |                |            |<-----------------
        |            |                |     store a mapping state
        |            |                |            |
        |Packet for A'                |            |
        |----------->|                |            |
        |            |AMR(lookup A')  |            |
        |            |-------------->|            |
        |            |                |Query addr A'
        |            |                |----------->|
        |            |                |            |
        |            |                | return LLA |
        |            |                |<-----------|
        |            |                |            |
        |            |AMC(A''s TLLA)  |            |
        |            |<--------------|            |
        |            |                |            |
        |            |            Packet for A'   |
        |            |            [via EVPN    ]  |
        |            |            [Data Tunnels]  |
        |            |------------------------------------>
        |            |                |            |
```

Figure 12: Pull model

In the case of push model, the EVPN support synchronizes its state upon a route type 2 with the 6LBR, and the 6LBR maintains an abstract data structure for all information known to EVPN. This way, the 6LBR already has the mapping information to satisfy any request for an existing mapping and it can answer right away. Figure 13 illustrates a successful push model lookup flow, when the 6LBR is already in possession of the mapping.

```
            Server                        Leaf
       +----------------+            +----------------+
       |                |            |                |
        6LN          6LR             6LBR        EVPN
         |             |               |             |
         | <vSwitch>   |  Ethernet     | <call I/F>  |
         |             |               |             |
         |             |               |             |
         |             |               |             |
         |             |               |             | Receive EVPN
         |             |               |             | route type 2 for
         |             |               |             | remote address A'
         |             |               |             | [via BGP signaling]
         |             |               |             |<-----------------
         |             |               |     store a mapping state
         |             |               |             |
         |             |               | indicate LLA|
         |             |               |<-----------|
         |             |     store a mapping state   |
         |             |               |             |
         |Packet to A' |               |             |
         |----------->|               |             |
         |             |AMR(lookup A')|             |
         |             |------------->|             |
         |             |               |             |
         |             |AMC(A's TLLA)  |             |
         |             |<-------------|             |
         |             |               |             |
         |             |            Packet to A'     |
         |             |            [via EVPN    ]   |
         |             |            [Data Tunnels]   |
         |             |----------------------------------->
         |             |               |             |
```

Figure 13: Push model

In a mixed environment, a lookup failure (the mapping is not found though the address is present in the network) may be caused by a legacy node that was node discovered (aka a silent node). In that case, it is an administrative decision for the 6LR to broadcast an NS(lookup) or to return an error as shown in Figure 11.

5. Requirements on the EVPN-Unaware Host

This document describes how EVPN routing can be extended to reach a Host. This section specifies the minimal EVPN-independent functionality that the Host needs to implement to obtain routing services for its addresses.

5.1. Support of 6LoWPAN ND

A host sees a prefix as not on-link (e.g., it learned that prefix in a PIO in a RA with the L flag not set) should not attempt to resolve an address within that prefix using a multicast NS(lookup). Instead, it must pass its packets to a router, preferably one that advertises that prefix in a PIO; it must register the address that it uses as source to that router to enable source address validation using [RFC8505]. It is recommended that the Host also implements [RFC8928] to prove its ownership of its addresses.

The Host is expected to request routing services from a router only if that router originates RA messages with a 6CIO that has the L, P, and E flags all set to 1 as discussed in Section 3.7, unless configured to do so. To obtain routing services for one of its addresses, the host must register the address to a router that advertises the prefix, setting the "R" and "T" flags in the EARO to 1 as discussed in Section 3.3.1 and Section 3.3.2, respectively.

This document echoes the behavior specified in [RFC9010] whereby, when the R Flag set to 1 in a NS(EARO) is not echoed in the NA(EARO), the host must understand that the route injection failed, and if the R flag is reset later in an asynchronous NA(EARO), the host must understand that routing service has failed.

The host may attach to multiple 6LRs and is expected to prefer those that provide routing services. The abstract model for this is a P2MP interface that wraps together as many P2P IP Links the host has adjacencies to 6LRs over that interface. The IPv6 address and the subnet are associated to that interface. The interface may be virtual and it may bundle multiple physical Ethernet interfaces that connect to the individual 6LRs over point to point wires, possibly via a software switch. It can also be associated to one physical interface to an external switch, either way the PI Links can be associated to sub-interface of the interface.

The Host needs to register to all the 6LRs from which it desires routing services. The multiple Address Registrations to several 6LRs should be performed in a rapid sequence, using the same EARO for the same Address. Gaps between the Address Registrations will invalidate some of the routes till the Address Registration finally shows on

those routes. The routers recognize the same (ROVR, TID) as the signal of a multihomed address and maintain all the routes. In the case of EVPN, the Ethernet Segment must also be the same. The flow for a successful multihomed registration is illustrated in Figure 17.

[RFC8505] introduces error Status values in the NA(EARO) which can be received synchronously upon an NS(EARO) or asynchronously. The Host needs to support both cases and refrain from using the address when the Status value indicates a rejection.

This specification can be used to register Anycast and Multicast IPv6 addresses as discussed in Section 3.5 and replace MLDv2. To benefit from that capability, both the host and the 6LR MUST support the "A" and "M" flags that indicate Anycast and Multicast Addresses respectively. Those flags are defined in [I-D.ietf-6lo-multicast-registration] for use in the EARO and in the EDAR and EDAC messages.

6. Enhancements to EVPN

This section addresses the necessary changes to EVPN formats and behavior to support address registration security per [RFC8928] and mobility per [RFC8505] while retaining interoperability with traditional nodes. Basically the MAC Mobility Extended Community [RFC7432] and the ARP/ND Extended Community [RFC9047] are extended to advertise the sequencing and ownership validation information received in the EARO. With 6LRs injecting not only MACs via packet sources and TLLAO options but also ROVR into MAC Mobility and ARP/ND Extended Community, their semantics must be extended. Specifically following issues have to be addressed:

* The ROVR extends the semantics of the type-2 MAC advertisement via changes in ARP/ND and MAC Mobility Extended Community in the sense that the MAC must be aligned with the ROVR and under normal circumstances only the validity of ROVR guarantees that the type-2 MAC can be allocated to the requester. A MAC validated by ROVR should take precedence over MAC addresses allocated without using it given it presents a much more trustworthy topological information (it will be called ROVR MAC in further text). EVPN nodes not supporting extensions introduced by this document will need to be led to believe that a ROVR MAC is to be preferred over any advertisement they see as long a ROVR MAC route is present. The primary key of NRLI is still the (IP, MAC, Ethernet Tag ID) tuple as defined in [RFC7432], Section 7.2 and 7.7. This implies that the same MAC (and consequently ROVR MAC) can be assigned multiple IP addresses with differeent ROVRs and those represent independent NLRIs.

* The TID field in the EARO is smaller than the mobility sequence number in [RFC7432]. To allow a ROVR MAC mobility to "win" over legacy MACs in every circumstance, signaling must be introduced that enables to distinguish a TID-generated sequence number from a legacy sequence number.

* [RFC8505] supports IP multihoming, but does not differentiate multihoming from anycast or MAC address rotation. If an anycast IP address is registered with a different ROVR it will be rejected as duplicate. If it is registered with a different TID, the older sequence will be withdrawn. So the basic expectation with [RFC8505] is that the advertisement of an anycast address is coordinated, with the same keypair known to all parties, and the same value of the TID used by all nodes (and possibly never increasing), in other words, with no concept of mobility.

* [I-D.ietf-6lo-multicast-registration] adds new flags in the EARO to signal that an address is anycast or multicast, respectively. This specification injects that information in the ARP/ND extended community using matching flags as follows:

  - This specification uses the "O" flag in the ARP/ND extended community to signal that the IP address is anycast, and requires the local 6LBR to ignore the duplication if the same IP address is registered locally, and then to inject the NLRI with the "O" flag set on the ARP/ND Extended Community as well.

  - This specification adds the new "M" flag in the ARP/ND extended community to signal that the IP address is multicast, and requires the local 6LBR to ignore the duplication if the same IP address is registered locally, and then to inject the NLRI with the "M" flag set on the ARP/ND Extended Community as well.

* [RFC8928] needs the full ROVR to validate the address ownership, but the full ROVR can be too large to advertise through BGP. When an IP address is advertised through EVPN, it is REQUIRED that the EVPN Next Hop represents the address of the 6LBR of the leaf where the address was registered as well. This way, if the address is registered later on a second leaf, the 6LR in second leaf can leverage an out-of-band, i.e. via EVPN traffic carrying tunnels, EDAR/EDAC exchange with that 6LBR to validate that the ROVR in the registration is indeed the same. When that is the case, it can continue with the registration procedure and if successful, become a 6LBR for that IP address, either as a mobility event or as a multihomed registration.

* [RFC8928] expects nodes to autoconfigure the keypair that is used to form the ROVR, in which case the IPv6 address can be locally autoconfigured with no central coordination; in that case, the ROVR protects the address ownership and allows the fabric to enforce first-come first-serve and source address validation (SAVI). But it is also possible to provision the ROVR in the 6LBR in advance and later configure the keypair in the node, e.g., in the case of a trusted server. To enable that capability in EVPN, this specification adds a flag (U) to signal that the 6LBR that injects the address in EVPN does not provide reachability to the address. When that flag is set, the value of the TID is ignored in the mobility computation, the mapping to the MAC address is ignored, and the route to the IP address is not injected in the RIB on ROVR nodes. Non-ROVR nodes will consider the node a "honey-pot". Once the address is registered by a 6LN in the network and the according validation with the node advertising the U-bit version of the route is performed, the owner will inject the route without the U-bit. A node advertising the NLRI with U-bit in its ARP/ND Extended Community MUST withdraw the U-bit route once it sees a validated NLRI without the U-bit and it MAY reinject the route with the U-bit once all routes without the U-bit are withdrawn to protect the address again.

6.1. Updated ARP/ND Extended Community

The ARP/ND Extended Community defined in [RFC9047] is a transitive EVPN Extended Community (Type field value of 0x06) with a Sub-Type of 0x08. It is advertised along with EVPN MAC/IP Advertisement routes that carry an IPv4 or IPv6 address. This the ARP/ND Extended Community to transport fields from the EARO is natural since the EARO is an extension to IPv6 ND.

EVPN signaling is not used to carry the full ROVR since without challenge per [RFC8928] they do not represent any difference over using the IP/MAC combination. A Hash of the ROVR is still passed in the ARP/ND Extended Community to immediately detect a duplication, with 2 different values of the hash for the same address. The full ROVR is verified upon a movement or a multi-homed advertisement using an EDAR/EDAC exchange with the 6LBR that advertised the address first.

Additionally, backwards compatibility could not be preserved given comparing routes based on ROVR would present a change in primary key of NLRIs which non-ROVR routers do not implement. An indication from a ROVR node that a MAC has been validated by proof of ownership is enough to convey the necessary information.

ROVR nodes MUST set the "H" flag in the ARP/ND Extended Community to indicate that the advertisement carries a TID and a ROVR Hash in case the host followed the according procedures.

ROVR nodes MUST set the "V" flag if the address assignment passed proof of ownership per [RFC8928]. Such "validated" ROVR MAC addresses will be preferred by ROVR nodes over non-validated ROVR MACs.

In case a ROVR node configures the address as "sticky" (since the sticky bit semantics have been changed to the point a ROVR cannot tell whether address is really sticky unless advertised as such by non-ROVR node) a new "X" flag called "super sticky" is introduced.

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type=0x06     | Sub-Type=0x08 |    Flags (2 octet)            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      TID      |                 ROVR Hash                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

   Flags field:
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|U|M|V|H|I| |O|R|   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 14: Updated ARP/ND Extended Community

The "I","O", and "R" flags are defined in [RFC9047]. The following new fields are defined:

U: 1-bit flag. "Unreachable", indicating that the IP address is not reachable via that EVPN next hop, but is advertised for the purpose of protecting the value of the ROVR until a first 6LBR that can reach the address becomes available.

M: 1-bit flag. "Multicast", indicating that the IP address duplication should be ignored. When this bit is set, TID should be ignored in comparison of EVPN advertisements, i.e. all ROVR MACs at same level of validation MUST be considered having same TID.

V: 1-bit flag. "ROVR Validated" indicates that the MAC passed proof of ownership per [RFC8928]. Presence of this bit implies the "R" bit being set regardless of its value.

H: 1-bit flag. "ROVR Capable" indicates that the advertisement is

originated after processing signaling from host meeting the requirements in Section 5, and that the advertised MAC address is a ROVR MAC. This also indicates that the TID and ROVR Hash fields are populated based on information from the most recent EARO [RFC8505] from the host.

TID: 8-bits structure. The TID [RFC8505] from the Registering Node.

ROVR Hash: 16-bits hash of ROVR [RFC8505] from the Registering Node. The Hash is built by XOR'ing ROVR bytes in network order into the least significant byte and rotating the two bytes result after every byte by one bit to the left.

6.2. Updated Mobility Extended Community

This specification extends the MAC Mobility Extended Community to transport the TID instead of increasing the normal sequence number. The TID is placed in the high bits of the sequence number field to "override" any normal MAC advertisement (further considerations will be provided in Section 6.3). this allows to design a solution that, while backwards compatible, allows to introduce ROVR MAC as "more trusted" entities. Figure 15 presents the according extensions that will however necessitate some further explanation.

To introduce a "precedence" of ROVR MACs over normal EVPN MACs ROVR MACs are advertised to look like "sticky" MACs for non-ROVR nodes. As defined in the glossary, for simplicity reasons such nodes will be called non-ROVR nodes vs. ROVR nodes. The "sticky" bit will force non-ROVR nodes to disregard the sequence number and accept any IP/MAC route provided.

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type=0x06     | Sub-Type=0x00 | Flags = 0 |X|S|  Reserved = 0 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|T| Flags = 0   |       TID     |     Reserved = 0              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 15: Updating the MAC Mobility Extended Community for ROVR MACs

This specification updates the Sequence Number field defined in [RFC7432]. The field is split in 3 parts, one 8-bit flags field, the TID, and a reserver 2-byte field. The unspecified flags and the reserved fields MUST be set to 0 by the sender and ignored by the receiver.

The "S" flag is defined in [RFC7432]. The following new fields are defined:

X: 1-bit flag. "Super Sticky" indicates that the ROVR MAC is sticky and should follow procedures of sticky per [RFC7432].

T: 1-bit flag. "TID Present", MUST be set to 1 when this specification is used. This ensures that the TID always wins vs. the sequence counter defined in [RFC7432]

TID: The TID copied from the most recent EARO [RFC8505] from the Registering Node.

6.3. Extended ROVR MAC Procedures

In case a non-ROVR node advertises a sticky MAC by setting the "S" bit and a ROVR node sees an ROVR address registration for the same MAC it MUST follow procedures per [RFC7432].

In case a non-ROVR node advertises a sequence number larger than the one generated by TID on a ROVR node, the ROVR node SHOULD advertise a Sequence Number consisting of all bits being set to force a "roll-over" on all nodes and then fall back to advertising the TID generated sequence number again. In case a non-ROVR node persists in increasing the sequence number after that it is indication of violation of [RFC7432] on its part.

A ROVR node advertising a ROVR MAC that has not been validated and receiving same type-2 route that has been validated MUST immediately withdraw its advertisement.

A ROVR node advertising a ROVR MAC and receiving an equivalent ROVR MAC from other node with a higher TID MUST immediately withdraw its advertisement. This will allow the non-ROVR nodes to correctly interpret the sequence as MAC move despite ignoring the sequence number due to presence of "S" bit.

A ROVR node that receives a ROVR MAC with "super sticky" indication and seeing the MAC locally MUST follow analogous procedures to [RFC7432].

Multi-homing a MAC on mix of ROVR and non-ROVR nodes will lead to operational notifications since per [RFC7432] the non-ROVR node will interpret the situation as a sticky MAC that has shown up on its local interface unless an implementation is somewhat clever and understands that the presence of the same ESI on all the routes indicates that this situation does not represent a sticky MAC being moved.

7. Protocol Operations

Following section illustrates several situations and resulting signaling in EVPN from the point of view of a ROVR node.

Figure 16 illustrates the registration flow of a new address protected by [RFC8928]. The ROVR in the EARO is a Crypto-ID that derives from a public address through hashing with some other terms. The router challenges the host with a status of 5 (validation requested).

The host performs the NS again, passing the parameters that enable to build the Crypto-ID in a Crypto-ID Parameters Option (CIPO), and signing that set of parameters together with a pair of Nonce values, one from each side, in a resulting Neighbor Discovery Protocol Signature Option (NDPDO). The 6LR first verifies that the Crypto-ID can be rebuilt based on the public key, then verifies that the signature in the NDPSO was effectively performed with the associated public key. When that is the case, the registration flow can continue, else the registration is rejected with a status of 10 (Validation Failed) in the NA(EARO).

With this specification, the 6LBR communicates internally with the collocated EVPN router to inject the route in EVPN. Since the [RFC8928] validation was performed, the V flag is set. Once this is done, the local 6LBR installs a local state associated to the NCE and becomes owner of the registration, whereas the remote leaves optionally install a remote state for the address with the indication of the 6LBR that owns the registration. The local 6LBR MUST be signalled as EVPN Next Hop for the route.

```
       Local            Local              Route           Remote
       Server           Leaf             Reflector         Leaf
      +-------+       +---------+        +-------+        +---------+
      |       |       |         |        |       |        |         |
         6LN           6LBR/EVPN            BGP            EVPN/6LBR
          |                |                 |                |
          |    Ethernet    |       [via BGP signaling]        |
          |                |                 |                |
          |                |                 |                |
          |  NS(EARO(      |                 |                |
          |       R set))  |                 |                |
          |--------------->|                 |                |
          |   ROVR in EARO is cryptographic  |                |
          |                |                 |                |
          |     NA(EARO(   |                 |                |
          |   R not set,   |                 |                |
          |   status = 5), |                 |                |
          |        Nonce1) |                 |                |
          |<---------------|                 |                |
          |                |                 |                |
          |   NS(EARO(     |                 |                |
          |        R set), |                 |                |
          |      CIPO,     |                 |                |
          |      Nonce2,   |                 |                |
          |      NDPSO)    |                 |                |
          |--------------->|                 |                |
          |        Proof verified            |                |
          |     no state for that address    |                |
          |      install local state         |                |
          |                |                 |                |
          |                |  ROVR MAC Route A'               |
          |                |---------------->|                |
          |        route injected            |                |
          |                |                 |                |
          |     NA(EARO(   |                 |                |
          |   R set,       |                 |                |
          |   status = 0)) |                 |                |
          |<---------------|                 |                |
          |                |                 | ROVR MAC Route A'
          |                |                 |--------------->|
          |                |                 |   install remote state
          |                |                 |                |
```

Figure 16: Host Registration

Figure 17 presents the same flow but for a multihomed address; here and in the following flows, the proof of ownership section is not shown, but its use is RECOMMENDED. The interesting piece is that

when the node registers to the second 6LBR, that second 6LBR find that there is a first 6LBR that already own the registration. Using and EDAR / EDAC flow, the second 6LBR validates that the ROVR and TID are identical, in which case it accepts the registration and becomes another 6LBR owner of the registration. The result is that the 2 6LBRs are synchronized and any of the 2 can now be used, e.g., if the address is registered a third time.

```
      Local            Local          Local
      Server           Leaf 1         Leaf 2           Reflector
    +-------+       +---------+    +---------+       +---------+
    |       |       |         |    |         |       |         |
       6LN           6LBR/EVPN      6LBR/EVPN             |
        |            |              |                     |
        |    Ethernet|              |                     |
        |            |              |       [via BGP signaling]
        |            |              |                     |
        |   NS(EARO) |              |                     |
        |--------------->|          |                     |
        |      install local state  |                     |
        |            |              |                     |
        |            |------ROVR MAC Route A'   ----->|
        |   NA(EARO) |              |                     |
        |<---------------|          |                     |
        |            |              |<----------------|
        |            |    install remote state         |
        |            |              |                     |
        | NS(same address, ES and EARO) |                |
        |------------------------------>|                |
        |            |              |                     |
        |            |        Same ES and ROVR Hash       |
        |            |   EDAR       |                     |
        |            |<-------------|                     |
        |            |              |                     |
        |       ROVR match, TID OK  |                     |
        |            |              |                     |
        |            |EDAC(status=0)|                     |
        |            |------------->|                     |
        |            |    install local state             |
        |            |              |                     |
        |            |       ROVR MAC Route A'            |
        |            |              |---------------->|
        |            |<------------------------------|
        |        install remote state |                  |
        |            |              |                     |
        |     NA(EARO(status = 0))  |                     |
        |<------------------------------|                |
        |            |              |                     |
```

Figure 17: Multihoming

The registration is associated with a lifetime, and it must be renewed with an incremented TID. The new TID is propagated in EVPN as illustrated in Figure 18.

```
    Local            Local              Route           Remote
    Server           Leaf             Reflector         Leaf
   +-------+       +---------+        +-------+        +---------+
   |       |       |         |        |       |        |         |
      6LN           6LBR/EVPN            BGP            EVPN/6LBR
       |               |                  |                |
       |    Ethernet   |       [via BGP signaling]         |
       |               |                  |                |
       |               |                  |                |
       |  NS(EARO(     |                  |                |
       |    TID+1))    |                  |                |
       |--------------->|                  |                |
       |   renew lifetime locally         |                |
       |               |                  |                |
       |               |  ROVR MAC Route A'(TID+1)         |
       |               |---------------->|                |
       |   NA(EARO     |                  |--------------->|
       |   status = 0)  |                  |     update remote state
       |<---------------|                  |                |
       |               |                  |                |
```

Figure 18: Host Registration Renewal

Figure 19 illustrates the case where a second host registers the same address, creating a potential address duplication situation. in most cases, the ROVR hash will be different, and the local 6LBR can reject the registration is a status of 1 (duplicate) right away.

```
      Local            Local              Route           Remote
      Server           Leaf             Reflector         Leaf
     +-------+       +---------+        +-------+        +---------+
     |       |       |         |        |       |        |         |
        6LN           6LBR/EVPN            BGP            EVPN/6LBR
         |                |                 |                |
         |    Ethernet    |                 |                |
         |                |        [via BGP signaling]       |
         |                |                 |                |
         |                |                 |         Local state for A'
         |                |                 |                |
         |                |          ROVR MAC Route A'       |
         |                |              ROVR Hash G         |
         |                |                 |<---------------|
         |                |<----------------|                |
         |      Create remote state for A'  |                |
         |                |                 |                |
         |   NS(EARO)     |                 |                |
         |--------------->|                 |                |
         |         compute ROVR Hash F      |                |
         |                |                 |                |
         |   NA(EARO(     |                 |                |
         |   status = 1)) |                 |                |
         |<---------------|                 |                |
         |                |                 |                |
         |                |                 |                |
```

Figure 19: Duplicate Addresses

Figure 20 illustrates the case of an address duplication situation where by chance, the ROVR hashes are the same. In that case, the local 6LR checks with the 6LBR that owns the registration using an EDAR/EDAC message exchange. As opposed to the ROVR hash, the full ROVRs do not collide and the registration is also rejected.

```
        Local            Local              Route           Remote
        Server           Leaf             Reflector         Leaf
       +-------+       +---------+        +-------+        +---------+
       |       |       |         |        |       |        |         |
          6LN           6LBR EVPN            BGP            EVPN/6LBR
           |              |   |               |                |   |
           |    Ethernet  |   |               |                |   |
           |              |   |        [via BGP signaling]     |   |
           |              |   |               |                |   |
           |              |   |               |     Local state for A'
           |              |   |               |                |   |
           |              |   |        ROVR MAC Route A'       |   |
           |              |   |            ROVR Hash G         |   |
           |              |   |               |<--------------|   |
           |              |   |<--------------|                |   |
           |      Create remote state for A'  |                |   |
           |              |   |                                |   |
           |              |                                        |
           |              |       [[out of band signaling]]        |
           |   NS(EARO)   |                                        |
           |------------->|                                        |
           |         compute ROVR Hash G                           |
           |              |                                        |
           |              |      EDAR to EVPN Next Hop             |
           |              |--------------------------------------->|
           |              |                                        |
           |              |              EDAC (status = 1)         |
           |              |<---------------------------------------|
           |              |                                        |
           |  NA(EARO(    |                                        |
           |  status = 1))|                                        |
           |<-------------|                                        |
           |              |                                        |
```

Figure 20: Duplicate Addresses, ROVR Hash Collision

Figure 21 shows a rare case where the registration has already moved elsewhere with an incremented TID when the local registration is received after being delayed in the network. In that case, the registration is rejected with a status of 3 (moved).

```
        Local            Local              Route           Remote
        Server           Leaf             Reflector         Leaf
       +-------+       +---------+        +-------+        +---------+
       |       |       |         |        |       |        |         |
          6LN           6LBR/EVPN            BGP            EVPN/6LBR
           |                |                  |                |
           |    Ethernet    |                  |                |
           |                |        [via BGP signaling]        |
           |                |                  |                |
           |                |          ROVR MAC Route A'        |
           |                |         ESI X', Hash F, TID Z     |
           |                |                  |<---------------|
           |                |<----------------|                |
           |         create remote start A'    |                |
           |       ESI X', ROVR Hash F, TID Z  |                |
           |                |                  |                |
           |   NS(EARO(     |                  |                |
           |     TID=Z-1))  |                  |                |
           |--------------->|                  |                |
           |           computes as             |                |
           |           ROVR Hash F             |                |
           |           ESI X'', TID Z-1        |                |
           |   NA(EARO)     |                  |                |
           |   status = 3)) |                  |                |
           |<---------------|                  |                |
           |                |                  |                |
```

Figure 21: Address Already Moved

Address move differs from multi-homing by the ESI being different as visualized by Figure 22. In case of different ESI BGP signalling happens immediately, in case of multi-homing we can reasonably expect for the signalling to catch up on the other leg with a new, higher TID. However, since ESI matches TID doesn't matter strictly speaking and the new remote state can be installed as is. However, if 6LN is not refreshing it registration we can expect elapsed lifetime to create scenario Figure 25 over time.

```
      Local            Local              Route           Remote
      Server           Leaf             Reflector         Leaf
     +-------+       +---------+        +-------+        +---------+
     |       |       |         |        |       |        |         |
        6LN           6LBR/EVPN            BGP            EVPN/6LBR
         |                |                |                |
         |    Ethernet    |                |                |
         |                |        [via BGP signaling]      |
         |   NS(EARO)     |                |                |
         |--------------->|                |                |
         |        Create local state       |                |
         |                |        ROVR MAC Route A'        |
         |                |     ESI X', ROVR Hash F, TID Z  |
         |                |---------------->|               |
         |                |                 |-------------->|
         |                |                 |     Create remote state

   Same ES (multihomed):
         |                |                 |               |<--
         |                |                 | New local state A' created
         |                |                 |               |
         |                |          ROVR MAC Route A'      |
         |                |         ESI X', Hash F, TID Z+1 |
         |                |                 |<--------------|
         |                |<----------------|               |
         |        Create remote state       |               |
         |       Keep local expect renew    |               |
         |                |                 |               |

   Different ES (movement):
         |                |                 |               |<--
         |                |                 | New local state A' created
         |                |                 |               |
         |                |          ROVR MAC Route A'      |
         |                |    ESI X'', ROVR Hash F, TID Z+1 |
         |                |                 |<--------------|
         |                |<----------------|               |
         |        Create remote state       |               |
         |                |                 |               |
         |   NA(EARO(     |                 |               |
         |   status = 4)) |                 |               |
         |<---------------|                 |               |
         |                | Withdraw ROVR MAC Route A'      |
         |                |---------------->|               |
         |                |                 |-------------->|
         |     remove local state           |               |
         |                |                 |               |
```

Figure 22: Address Move

The host that registered the address may cancel the registration at any time, e.g., if the address is removed fmor its own interface. This is done by registering with a lifetime if 0 as shown in Figure 23. The Leaf may respond with a status of 0 to indicate success, but a status of 4 (removed) is preferred for this situation.

```
     Local            Local              Route           Remote
     Server           Leaf             Reflector         Leaf
    +-------+       +---------+        +-------+        +---------+
    |       |       |         |        |       |        |         |
       6LN           6LBR/EVPN            BGP            EVPN/6LBR
        |                |                |                |
        |    Ethernet    |                |                |
        |                |        [via BGP signaling]      |
        |                |                |                |
        | NS(EARO,       |                |                |
        |    lifetime=0) |                |                |
        |--------------->|                |                |
        |                |                |                |
        |                | Withdraw ROVR MAC Route A'      |
        |                |---------------->|                |
        |                |                |--------------->|
        |   NA(EARO(     |                |                |
        |   status = 4)) |                |                |
        |<---------------|                |                |
        |                |                |                |
        |     remove local state          |                |
        |                |                |                |
```

Figure 23: Address Removal

The host that registered the address may withdraw the route but maintain the NCE, e.g., in the case where it is multihomed but does not want to use one interface for the traffic back as this time. This is done by registering with the R flag set to 0 as shown in Figure 24.

```
        Local            Local              Route           Remote
       Server           Leaf             Reflector         Leaf
      +-------+       +---------+        +-------+        +---------+
      |       |       |         |        |       |        |         |
         6LN           6LBR/EVPN            BGP            EVPN/6LBR
          |                 |                 |                 |
          |    Ethernet     |                 |                 |
          |                 |        [via BGP signaling]        |
          |                 |                 |                 |
          | NS(EARO,        |                 |                 |
          |    R reset)     |                 |                 |
          |---------------->|                 |                 |
          |                 |                 |                 |
          |                 | Withdraw ROVR MAC Route A'        |
          |                 |---------------->|                 |
          |                 |                 |--------------->|
          |   NA(EARO(      |                 |                 |
          |   R reset,      |                 |                 |
          |   status = 0))  |                 |                 |
          <----------------|                 |                 |
          |                 |                 |                 |
          |     retain only NCE               |                 |
          |                 |                 |                 |
```

Figure 24: Route Type 2 Removal

When the lifetime elapses, the 6LBR requires the collocated EVPN router to withdraw the route.

```
         Local            Local              Route           Remote
        Server           Leaf             Reflector         Leaf
       +-------+       +---------+        +-------+        +---------+
       |       |       |         |        |       |        |         |
          6LN           6LBR/EVPN            BGP            EVPN/6LBR
           |                |                  |                |
           |    Ethernet    |                  |                |
           |                |        [via BGP signaling]        |
           |                |                  |                |
           |     lifetime Expired              |                |
           |                |                  |                |
           |                | Withdraw ROVR MAC Route A'        |
           |                |---------------->|                |
           |                |                  |--------------->|
           |   NA(EARO(     |                  |                |
           |   status = 4)) |                  |                |
           |<---------------|                  |                |
           |                |                  |                |
           |     remove local state            |                |
           |                |                  |                |
```

Figure 25: Lifetime Elapse

8. Security Considerations

TBD

9. IANA Considerations

9.1. MAC Mobility Extended Community Flags

This document creates the "MAC Mobility Extended Community Flags" registry based on one flag initially defined in [RFC9047] and more flags defined in Section 6.2 as shown in Table 1:

Flag Position	Name	Reference
0..5	Unassigned	N/A
6 (Suggested)	Super-sticky (X)	THIS RFC
7	Sticky (S)	RFC 7432

Table 1: MAC Mobility Extended Community Flags

9.2. ARP/ND Extended Community Flags

This document modifies the "ARP/ND Extended Community Flags" registry initially created in Section 5 of [RFC9047] . Section 6.1 suggests to assign new entries in the Registry as indicated in Table 2:

Flag Position	Name	Reference
0 (Suggested)	Unreachable(U)	THIS RFC
1 (Suggested)	Multicast (M)	THIS RFC
2 (Suggested)	ROVR Validated (V)	THIS RFC
3 (Suggested)	ROVR Capable (H)	THIS RFC

Table 2: New ARP/ND Extended Community Flags

10. Acknowledgments

The authors wish to thank you for reading that far. We acknowledge and express gratitude to Wen Lin, Stephane Litkowski, Eric Levy-Abegnoli, Lukas Krattiger, Jerome Tollet, and Ali Sajassi, for their help and support.

11. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/info/rfc4861>.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, DOI 10.17487/RFC4862, September 2007, <https://www.rfc-editor.org/info/rfc4862>.

[RFC6775] Shelby, Z., Ed., Chakrabarti, S., Nordmark, E., and C. Bormann, "Neighbor Discovery Optimization for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)", RFC 6775, DOI 10.17487/RFC6775, November 2012, <https://www.rfc-editor.org/info/rfc6775>.

[RFC6085] Gundavelli, S., Townsley, M., Troan, O., and W. Dec, "Address Mapping of IPv6 Multicast Packets on Ethernet", RFC 6085, DOI 10.17487/RFC6085, January 2011, <https://www.rfc-editor.org/info/rfc6085>.

[RFC7400] Bormann, C., "6LoWPAN-GHC: Generic Header Compression for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)", RFC 7400, DOI 10.17487/RFC7400, November 2014, <https://www.rfc-editor.org/info/rfc7400>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8505] Thubert, P., Ed., Nordmark, E., Chakrabarti, S., and C. Perkins, "Registration Extensions for IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Neighbor Discovery", RFC 8505, DOI 10.17487/RFC8505, November 2018, <https://www.rfc-editor.org/info/rfc8505>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC8928] Thubert, P., Ed., Sarikaya, B., Sethi, M., and R. Struik, "Address-Protected Neighbor Discovery for Low-Power and Lossy Networks", RFC 8928, DOI 10.17487/RFC8928, November 2020, <https://www.rfc-editor.org/info/rfc8928>.

[RFC9047] Rabadan, J., Ed., Sathappan, S., Nagaraj, K., and W. Lin, "Propagation of ARP/ND Flags in an Ethernet Virtual Private Network (EVPN)", RFC 9047, DOI 10.17487/RFC9047, June 2021, <https://www.rfc-editor.org/info/rfc9047>.

[UNICAST-LOOKUP] Thubert, P. and E. Levy-Abegnoli, "IPv6 Neighbor Discovery Unicast Lookup", Work in Progress, Internet-Draft, draft-thubert-6lo-unicast-lookup-02, 8 November 2021, <https://datatracker.ietf.org/doc/html/draft-thubert-6lo-unicast-lookup-02>.

[I-D.ietf-6lo-multicast-registration] Thubert, P., "IPv6 Neighbor Discovery Multicast and Anycast Address Listener Subscription", Work in Progress, Internet-Draft, draft-ietf-6lo-multicast-registration-15, 30 May 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-6lo-multicast-registration-15>.

[I-D.ietf-6lo-prefix-registration] Thubert, P., "IPv6 Neighbor Discovery Prefix Registration", Work in Progress, Internet-Draft, draft-ietf-6lo-prefix-registration-01, 14 August 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-6lo-prefix-registration-01>.

12. Informative References

[BCP38] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, DOI 10.17487/RFC2827, May 2000, <https://www.rfc-editor.org/info/rfc2827>.

[RFC6550] Winter, T., Ed., Thubert, P., Ed., Brandt, A., Hui, J., Kelsey, R., Levis, P., Pister, K., Struik, R., Vasseur, JP., and R. Alexander, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks", RFC 6550, DOI 10.17487/RFC6550, March 2012, <https://www.rfc-editor.org/info/rfc6550>.

[RFC8929] Thubert, P., Ed., Perkins, C.E., and E. Levy-Abegnoli, "IPv6 Backbone Router", RFC 8929, DOI 10.17487/RFC8929, November 2020, <https://www.rfc-editor.org/info/rfc8929>.

[RFC9010] Thubert, P., Ed. and M. Richardson, "Routing for RPL (Routing Protocol for Low-Power and Lossy Networks) Leaves", RFC 9010, DOI 10.17487/RFC9010, April 2021, <https://www.rfc-editor.org/info/rfc9010>.

[RIFT] Przygienda, T., Sharma, A., Thubert, P., Rijsman, B., Afanasiev, D., and J. Head, "RIFT: Routing in Fat Trees", Work in Progress, Internet-Draft, draft-ietf-rift-rift-18, 10 July 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-rift-rift-18>.

[IANA-EARO-STATUS] IANA, "Address Registration Option Status Values", <https://www.iana.org/assignments/icmpv6-parameters/icmpv6-parameters.xhtml#address-registration>.

Authors' Addresses

Pascal Thubert (editor)
06330 Roquefort-les-Pins
France
Email: pascal.thubert@gmail.com

Tony Przygienda
Juniper Networks, Inc
Switzerland
Email: prz@juniper.net

Jeff Tantsura
Microsoft
Email: jefftant.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 24 May 2024

F. Duan
S. Chen
Huawei Technologies
Y. Liu
China Mobile
H. Wang
Ruijie Networks Co., Ltd.

21 November 2023

Multicast VPN Upstream Designated Forwarder Selection
draft-wang-bess-mvpn-upstream-df-selection-08

Abstract

This document defines Multicast Virtual Private Network (VPN) extensions and procedures of designated forwarder election performed between ingress PEs, which is different from the one described in [RFC9026] in which the upstream designated forwarder determined by using the "Standby PE Community" carried in the C-Multicast routes. Based on the DF election, the failure detection point discovery mechanism between DF and standby DF is extended in MVPN procedures to achieve fast failover by using BFD session to track the status of detection point. To realize a stable "warm root standby", this document obsolete the P-Tunnel status determining procedure for downstream PEs in regular MVPN by introducing a RPF Set Checking mechanism as an instead.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

MVPN [RFC6513] and [RFC6514] defines the MVPN architecture and MVPN protocol specification which include the basic procedures for selecting the Upstream Multicast Hop. Further [RFC9026] defines some extensions to select the primary and standby upstream PE for a VPN multicast flow on downstream PEs. After selecting the Upstream Multicast Hop, the downstream PEs send MVPN C-Multicast routes to both primary and standby Upstream PE. Upon receiving the MVPN join routes, the upstream / ingress PEs can either perform "hot root standby" or "warm root standby". For the "hot root standby" mechanism, all the ingress PEs, regardless of the primary or standby role, forward (C-S,C-G) flow to other PEs through a P-tunnel, forcing the egress PEs to discard all but one. In this way, the failover can be conducted by leaf PE within extremely short duration when the failure of upstream link or device is detected. However, this will cause the steady traffic redundancy throughout the backbone network. In the scenario where bandwidth waste issue is concerned, such as enterprise networks crossing provider networks, the "warm root standby" mechanism is expected to be a better solution. However, there are some problems when deploying the "warm root standby" mechanism described in [RFC9026].

a. Upon the failure of primary ingress PE, the leaf PE needs to send the new C-multicast route towards the standby ingress PE without carrying the Standby PE BGP Community according to [RFC9026]. Leaf PE needs to update all relevant C-multicast routes and sends them to the standby ingress PE. For example, if there are 1000 (C-S,C-G)s, 1000 C-multicast routes will be updated and resent so that the standby PE can finally forward traffic. The failover time can hardly reach the same level of "hot root standby" mechanism.

b. There is no endogenous mechanism for standby ingress PEs to discover and detect the failure of primary ingress PEs, resulting in the uncertainty in deployment and implementation. If the standby ingress PE can directly detect the failure of the primary ingress PE, it can take over the role of designated forwarder and send the traffic immediately.

c. In [RFC9026], the standby ingress PE is determined by using "Standby PE Community" carried in the C-Multicast routes. The premise of this mechanism is that all leaf PEs choose the same primary and standby ingress PEs, which may not be met due to transient unicast routing inconsistencies, the inconsistencies of

P-Tunnel status determined by each leaf PE or lack of support of the Standby PE community on leaf PE, causing that the "warm root standby" mechanism is not stable and returns to "hot root standby" mode because the standby ingress PE also sends multicast traffic to backbone when the condition is not satisfied.

d. When the primary and standby designated forwarders are selected based on IP addresses of root PEs, the primary and standby roles are fixed for each multicast flow. Ingress PEs cannot perform load balancing for different multicast traffic. Hashing algorithm in [RFC6513] utilized source and group addresses and allows load balancing for different (C-S,C-G)s. However, the specific procedure of selecting a standby PE was not specified.

The hot root standby is good at fast failover. The warm root standby has advantages of saving the bandwidth. In order to have both advantages of hot root standby and warm root standby, this document defines a new MVPN procedure of designated forwarder election performed between ingress PEs. Based on the DF election, the failure detection point discovery mechanism between DF and standby DF is extended to achieve fast failover by using a BFD session to track the status of detection point. To realize a stable "warm root standby", this document obsoletes the P-Tunnel status determining procedure for downstream PEs in regular MVPN by introducing a RPF Set Checking mechanism as an instead.

2. Terminology

The terminology used in this document is the terminology defined in[RFC6513], [RFC6514] and [RFC9026].

For convenience of description, the abbreviations used in this document is listed below.

DF: Designated Forwarder

IDF: Ingress Designated Forwarder

UMH: Upstream Multicast Hop

P-tunnel: Provider-Tunnel

VPN: Virtual Private Network

MVPN: Multicast VPN

GTM: Global Table Multicast

RD: Route Distinguisher

NLRI: Network Layer Reachability Information

BFD: Bidirectional Forwarding Detection

MD: My Discriminator

VRI: VRF Route Import Extended Community

RR: Route Reflector

SFS: Selective Forwarder Selection

PTA: PMSI Tunnel Attribute

A new term is introduced below.

RPF Set Checking: RPF Set is a set of valid upstream interfaces that can accept multicast traffic. RPF Set checking allows multicast traffic to be received from backup P-Tunnel quickly when failure occurs.

3. Scenario

3.1. Passive IDF Negotiation Mode

```
                  +-----------------------+
                  |                       |
             / Root PE1               Leaf PE1 ----- R1
           /      |                       |
      S1 --- CE   |    Provider Backbone  |
           \      |                       |
            \  Root PE2               Leaf PE2 ----- R2
                  |                       |
                  +-----------------------+
```

Figure 1: Passive IDF Negotiation Mode

In this scenario, the interfaces multihoming CE to provider's root PEs are bundled together and working in a eth-trunk mode, and a multichassis protocol is running between the multi-homed root PEs to coordinate with the CE to perform single active or all active data sending mode between CE and root PEs. Regardless either of the two sending mode is chosen, CE received multicast data from S1 only selects one interface to forward traffic, thus the root PE homed by

the selected interface is responsible for sending the corresponding multicast traffic to leaf PEs. The multi-homed root PEs do not really run an IDF negotiation procedure between themselves but accept the IDF role passively. Therefore, we call this scenario using Passive IDF Negotiation Mode in this document.

3.2. Active IDF Negotiation Mode

```
                                +------------------------+
          +----------------+    |                        |
          |                | -- Root PE1               Leaf PE1 ----- R1
          |                |    |                        |
   S1 --- | Client Network |    |    Provider Backbone   |
          |                |    |                        |
          |                | -- Root PE2               Leaf PE2 ----- R2
          +----------------+    |                        |
                                +------------------------+
```

Figure 2: Active IDF Negotiation Mode

In this scenario, the "Client Network" is a layer 3 network area containing one or more CE routers. If only one CE router is included in the "Client Network" , the main difference between this circumstance and above is that the interfaces multihoming CE to root PEs are not bundled and each of them is an individual layer 3 interface. The IP subnet of the multihoming interfaces can be in either same or different, each of the multi-homed root PEs can receive one copy of the specific multicast stream (S1, G) received through the "Client Network". For the "warm root standby" mechanism, only one root PE (Called IDF in this document) can send the received multicast traffic to leaf PEs through provider’s backbone. Thus the IDF must be selected among the multi-homed root PEs by themselves. So, in this document, we call this scenario using Active IDF Negotiation Mode.

4. Specification

4.1. IDF Negotiation Community

This community is carried in the UMH routes and used by the multi-homed root PEs to notify each other to perform IDF election. Leaf PEs can also check whether the UMH route is containing this community to perform checking according to the RPF Set Checklist. The value of this community will be allocated by IANA for each negotiation mode individually from the "Border Gateway Protocol (BGP) Well-known Communities" registry using the First Come First Served registration policy.

4.2. BFD Discriminator Attribute

This attribute is carried in UMH routes and its format reuses the one defined in [RFC9026] with the "BFD Mode" field redefined as a unicast BFD session type, of which the value is recommended to be 2 and will be allocated by IANA according to the registration policy. The source IP optional TLV in this document is mandatory and used to discover the failure detection point of the IDF.

5. Procedure

5.1. Signaling

In this section, the procedure is under the condition that the value of the RDs of multi-homed root PEs for a same MVPN are distinct, which means that the VPN route originated by each multi-homed PE can be received by the others and leaf PEs can also perform SFS reliably.

5.1.1. Originating VPN Routes to Multicast Sources

To perform IDF election procedure in this document, the multi-homed root PEs MUST include an IDF negotiation Community in the originating VPN routes to multicast sources. The negotiation mode (Passive or Active) is determined by the connection type of the Client network / CE, and MUST be configured consistently on each multi-homed root PE.

In order to perform endogenous mechanism of IDF election and fast failure detection, the BFD Discriminator Attribute described in section 4.2 MUST also be carried when each multi-homed root PE originates a UMH routes, with the MD field filled with a local configured BFD discriminator and the IP address field of the Source IP TLV filled with the local IP of the interface connecting to the Client network / CE, from which the prefix of the originating UMH route is learned. If the UMH prefix is learned from more than one local interface, the one chosen to fill the Source IP TLV of the BFD Discriminator Attribute MUST be consistent with the one selected as RPF interface for the multicast stream sent by the corresponding multicast source. In this document, the filled Source IP address is the failure detection point, if the corresponding root PE is selected as the IDF of a specific multicast stream, it is used to establish a BFD session to do fast tracking of failure of IDF. In IPv6 scenarios, a global IPv6 address SHOULD be configured on the client facing interfaces to succeed in the establishment of multi-hop IPv6 BFD sessions.

5.1.2. Originating C-Multicast Routes

If a leaf PE decides to send C-Multicast routes to upstream PEs for a given (C-S,C-G), it follows the procedure described in [RFC6514] excepting that the RPF route of the c-root has an IDF negotiation community. According to the negotiation community, a distinct C-Multicast route for (C-S,C-G) is sent to each multi-homed root PE. Leaf PE installs all P-Tunnels rooted from the multi-homed PEs into the RPF tunnel checklist of the corresponding multicast traffic (C-S,C-G).

If there is a local receiver connected to one of the multi-homed root PEs and the Passive IDF Negotiation Mode is performed between them, the root PE having local receivers sends the specific C-Multicast route (C-S,C-G) joined by the local receivers to the multi-homed others, after which it installs all P-Tunnels rooted from the multi-homed others and local upstream interface into the RPF tunnel checklist of the corresponding multicast traffic (C-S,C-G).

5.1.3. Ingress Designated Forwarder Selection

For Passive IDF election, it is performed by CE routers as described in section 3.1. This section describes two optional solutions for Active IDF election.

5.1.3.1. Out-Of-Band Mechanism

VRRP specifies an election protocol that dynamically assigns responsibility of a virtual router to one of the VRRP routers on a LAN. The VRRP router controlling the IPv4 or IPv6 address(es) associated with a virtual router is called the Master, and it forwards packets sent to these IPv4 or IPv6 addresses. Similarly, the role of the VRRP routers associated with a virtual router can also be that of the upstream PEs in MVPN dual homing upstream PEs deployment.

The method of mapping the role of a VRRP router to that of a MVPN upstream PE is more likely an administrative measure and could be implemented as configurable policies. Both the primary and standby PEs install VRF PIM state corresponding to BGP Source Tree Join route and send C-Join messages to the CE toward C-S. Whereas only the primary upstream PE (Virtual Router Master according to VRRP) forwards (C-S,C-G) flow to downstream PEs through a P-tunnel if IDF election is performing between the upstream PEs.

Other private implementations or similar designated forwarder selection technologies could also be optional. However, a feasible technology should has the ability to be deployed per VRF and be associated with one Multicast VPN instance. All PEs connected to the same customer's layer 3 network area MUST keep a coincident status of whether performing IDF election or not by negotiating dynamically or being configured manually, the dynamic protocol for negotiation of this status is outside the scope of this document.

5.1.3.2. Endogenous Mechanism

Considering a multicast source connecting to the client network area multihoming to the provider network, the prefix of the source can be learned by all multi-homed root PEs, each of which originates a corresponding VPN route with a VRI Extended Community including the originator's IP address to the others and leaf PEs. According to that, each multi-homed root PE can learn all the others' originator IP addresses for a specific multicast source, based on which the IDF can be calculated consistently on each root.

The default procedure for IDF election is at the granularity of (C-S,C-G). There are two options listed below for IDF election of a specific multicast source C-S, a deployment can use each of them and MUST be configured consistently among the multi-homed root PEs:

a. To perform single IDF election for all C-Gs of a specific multicast source C-S, each PE builds an ordered list in ascending order of the IP addresses of all multi-homed PE nodes learning the UMH routes to the multicast source C-S (including itself). As described in the first paragraph of this section, each IP address in this list is extracted from the Global Administrator field of VRI Extended Community carried in those UMH routes related to the specified multicast source C-S. Every PE is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the numerically lowest IP address. The originator IP address with ordinal 0 is the winner, and the corresponding root PE is selected as IDF by every PE. The root PE of which the corresponding originator IP address is sub-optimal is selected as Standby IDF.

b. To perform IDF election for each C-G of a specific multicast source C-S, each PE also builds an ordered list of the IP addresses of all the multi-homed PE nodes at first. The difference between this option and above is that the election of IDF occurs not upon receiving all UMH routes of the other multi-homed PEs of the specified C-S but upon receiving the C-multicast join of the corresponding C-G. Assuming an ordered list of N elements, the PE with ordinal i is the IDF for a C-G when (C-G

mod N) = i. The PE with ordinal j is the Standby IDF when j is (C-G mod (N-1)). The calculation of standby IDF uses the ordered IP addresses list without considering the existence of the elected IDF element.

In order to reduce traffic waste between the Client Network and root PEs, a root PE can only send C-PIM Join messages towards the Client Network if it is the primary or standby DF.

5.1.4. Failure Detection and Fast Failover

For the Passive IDF Negotiation Mode, the CE router is responsible for the failure detection of multihoming links or multi-homed PE nodes using some existing solution, which is out the scope of this document. For the Active IDF Negotiation Mode with Out-Of-Band Mechanism described in section 5.1.3.1, the failure detection solution is always built in the multichassis protocols used for IDF election. This section only details the failure detection and fast failover procedure for the Active IDF negotiation mode with endogenous mechanism. Two methods are proposed to detect the failure.

5.1.4.1. BFD Method

To detect the failure of the node or the client facing link of IDF quickly, after the election of IDF PE and Standby IDF PE, the Standby IDF initializes a BFD session. Several important parameters of the BFD session are introduced as follows. The source IP of the BFD session uses a local configured IP address of the corresponding multicast VRF. The destination IP is extracted from the Source IP TLV of BFD Discriminator Attribute carried in the UMH route sent by the IDF. MD is filled with the MD field of BFD Discriminator Attribute carried in VPN routes originated by current Standby IDF. The YD(Your Discriminator) of the BFD session is dynamically learned through the BFD initialization procedure.

Upon the occasion of the failure, the status of the BFD session goes down. The Standby IDF PE of the C-Gs selecting the failure / affected node as IDF takes over the primary role and sends the multicast traffic belonging to C-Gs to leaf PEs through the backbone. The failure / affected PE withdraws its VPN route advertised before, this will re-trigger the procedure described in section 5.1.3.2 and a new IDF PE (which was the old Standby IDF PE) and Standby IDF PE will be selected. The new standby IDF MUST send C-PIM Join message towards Client Network to receive multicast traffic.

If the previous failure node / link goes up again or a new multi-homed PE of the specified multicast source is coming up and the IDF PE is calculated to be changed, the new IDF will take over the running IDF. To avoid data transfer crash, the running IDF (That should be the new Standby IDF) does not trigger the establishment of BFD session with new IDF until the local configured failback time expires, during which it keeps the IDF role and waits the new IDF completing the establishment of the multicast path from the SDR of the specified multicast source to itself. Upon the occasion of BFD session goes up, the running IDF stops sending multicast traffic to leaf PEs and the new IDF takes over the IDF role to send multicast stream for (C-S,C-G).

5.1.4.2. Monitoring traffic from IDF

The second method needs standby IDF to detect the failure by joining a P-Tunnel rooted at the IDF and monitoring the traffic received from the P-Tunnel. Even though there may be no local receiver connected to the standby IDF, the standby IDF needs to join the P-Tunnel by sending Leaf A-D Route or P-Tree Signaling. Standby IDF then sends the C-Multicast Route to IDF in order to receive traffic from the P-Tunnel. IDF will receive traffic from the IDF and the client facing interface simultaneously. However, it does not forward traffic to leaf PE when failure is not detected. When failures occur between the client network and IDF, the standby IDF will no longer receive any traffic from the P-Tunnel. The detection of interrupted flow will trigger the role transition from standby IDF to IDF. Then the new IDF will forward traffic to leaf PE.

However, the standby IDF may also cannot receive traffic when the failure occurs between the IDF and standby IDF. Under this circumstance, the standby IDF will switch to IDF when the client facing link and IDF still work well. There will be dual IDFs and leaf PE will receive two copies of the same flow. Suggestions about deployment are provided to avoid this situation:

a. Multiple parallel links are suggested to be deployed between the IDF and standby IDF. The probability of dual IDFs due to link failure can be greatly reduced.

b. PMSI tunnel protection can be utilized together. When the link between IDF and standby IDF fails, the underlay local protection of PMSI Tunnel can ensure that standby IDF can still receive traffic from IDF and avoid the dual-IDF situation.

5.2. Data Forwarding

5.2.1. Procedure on Root PEs

For the Passive IDF Negotiation Mode, the set of leaves of P-Tunnel rooted at each multi-homed PE has the others as members if the others have local receivers willing to accept the corresponding C-Flow. The detailed signaling procedure is described in section 5.1.2. When CE sends multicast data performing load balance to only one root PE (Which is the Passive IDF), IDF send this multicast traffic to the leaf PEs and the other multi-homed root PEs. When the multi-homed root PEs receive the C-Flow, it MUST perform RPF Set Checking, by accepting the data from either the client facing interface learning the corresponding route of the multicast source or anyone of the P-Tunnels rooted at the other multi-homed PEs. To avoid multicast traffic loop and duplication, the data received from the P-Tunnels at each root PE MUST NOT send back to P-Tunnels again and can only be forwarded to the local receivers of the receiving PE.

For the Active IDF Negotiation Mode, each multi-homed root PE receives a copy of C-Flow and forwards the multicast traffic to its local receivers. Only DF can send data to leaf PEs through backbone. All of the multi-homed root PEs perform RPF Set checking by matching their client facing interface exactly.

5.2.2. Procedure on Leaf PEs

For either of the two IDF negotiation modes described in this document, leaf PEs install each P-Tunnel rooted at each multi-homed root PE into the RPF Set checklist for the corresponding multicast flow (C-S,C-G), thus the multicast data sent by each of the multi-homed root PEs can be accepted by leaf PEs. Upon the failure of IDF, the Standby IDF takes over the primary role and leaf PEs are ready to receive the data sent by the new primary IDF with no latency thanks to the RPF Set checking mechanism.

5.3. Distinguishing UMH and C-multicast Routes

It was recommended in RFC 6514, on each multi-homed root PE, the UMH VRF of the MVPN MUST use its own distinct RD to support non-congruent unicast and multicast connectivity, the procedure described in above section is also under this premise. However, in [RFC7716], the UMH routes are not sent in the VPN-IP SAFI and there is no RD included in the NLRI key. There are also some other scenarios that the UMH VRF of the MVPN on the multi-homed PEs MUST be configured with a same RD for some deployment reasons, which causing that the IDF negotiation procedure can hardly be performed because that the UMH route originated by each multi-homed root PE cannot be collected reliably

by the other root PEs and leaf PEs because of the route selecting mechanism on BGP RRs. When UMH routes to same multicast source from different root PEs carry same RD or no RD, they will be same from the perspective of leaf PEs. Because Originating Router's IP Address is not the key field when the UMH route is processed.

For the scenarios of the same RD, this document introduces a new type of UMH route to be sent in MVPN SAFI, of which the NLRI key consists of the following fields:

```
+-----------------------------------------------+
|  RD (8 octets)                                |
+-----------------------------------------------+
|  IP Prefix Length (1 octet, 0 to 32 / 128)    |
+-----------------------------------------------+
|  IP Prefix (4 / 16 octets )                   |
+-----------------------------------------------+
|  Originating Router's IP Addr (4 / 16 octets) |
+-----------------------------------------------+
```

Figure 3: MVPN UMH Routes

The length of the IP Prefix field is determined by the address family of MVPN. If IPv4 is being used, it will be 4 octets. Otherwise it will be 16 octets for IPv6. After determining the length of IP Prefix field, the length of the Originating Router's IP Addr field is judged by NLRI key length. The type of this route will be allocated in IANA.

If the RDs of the UMH VRFs on the multi-homed root PEs are same, the root PEs import the routes of the client multicast sources to their local UMH VRFs and send above UMH routes to all other PEs of the MVPN. The UMH routes will carry a VRI Extended Community described in [RFC6514], an IDF negotiation Community and a BFD Discriminator Attribute described in this document. All the procedure applied to the VPN-IP routes described in [RFC6513] and [RFC6514] SHOULD be inherited by this UMH route. The receivers (which should be MVPN PEs) of this route MUST install it into their local multicast RIB as UMH route and it has a higher priority than other existing UMH route type while a MVPN PE using it to determine the upstream PE of a specified (C-S,C-G) or (C-*,C-G). The Originating Router's IP Addr will be used to identify UMH routes from different upstream PEs.

For the non-segmented Inter-AS P-Tunnel over IPv6 infrastructure scenarios, the length of Source AS field of C-Multicast routes cannot hold an IPv6 address, causing that it is hard to distinguish the two C-Multicast routes with a same granularity of (C-S,C-G) or (C-*,C-G)

sent to two ingress PEs individually. To solve this problem, this document introduces a Root Distinguisher Extended Community, which is an IP-address-specific Extended Community with a fixed type of IPv4. The Global Administrator field of this Extended Community is filled with a 4-octet global unique value configured. This 4-octet value and the IPv6 Originating Router's IP Addresses of each MVPN PE needs not to be a routable IPv4 address. The Local Administrator field of the Extended Community is filled with 0. The type and sub type of this Extended Community will be allocated in IANA.

The Root Distinguishing Extended Community is carried in the Intra-AS AD routes or the wildcard S-PMSI AD routes. According to [RFC6514] and [RFC6515], the non-segmented Inter-AS and IPv6 infrastructure scenarios are determined on MVPN leaf PEs. The Source AS field of the C-Multicast routes will be filled with the root distinguishing value of root PEs which the route is sent to.

5.4. Segmented Inter-AS Scenario

In the regular procedure of [RFC6514], Intra-AS AD route is only used in non-segmented Inter-AS scenario. In the segmented Inter-AS scenario, different Intra-AS AD routes originated by different PEs in the same AS are aggregated to a single Inter-AS AD route on ASBRs with the granularity of <AS, MVPN>. The specific original root PE's information is substituted with source AS during the aggregation, which results in that leaf PEs located in downstream ASes cannot differentiate two multicast traffic sent by different root PEs in the same original AS.

In this document, two approaches are proposed to facilitate the root PE selection of leaf PEs in downstream ASes.

This first approach is to use the wildcard S-PMSI AD route described in [RFC6625] instead of Intra-AS AD route. As described in [RFC6514], the S-PMSI AD route will not be aggregated by ASBR while being used to set up Inter-AS segmented S-PMSI tunnels, result in that Leaf PE in downstream AS can do explicit tracking of those tunnels established from the redundant PEs located in upstream AS. The propagation procedure between ASes follows the description in section 12.2 of [RFC6514].

The second method is to use PE Distinguisher Labels attribute defined in [RFC6514] to carry PE address and corresponding upstream assigned label at the segmentation point. In the segmented scenario, the PE Distinguisher Labels attribute SHOULD be distributed with the Inter-AS A-D route. When the multicast traffic is received from the Intra-AS P-Tunnel from an ingress root PE on ASBR, the ASBR will switch Intra-AS P-Tunnel to Inter-AS P-Tunnel and add the corresponding PE

Distinguisher Label as an inner label into the label stack. In this way, leaf PE will recognise the root PE of the multicast traffic and RPF Checking can be performed accordingly. As described in [RFC6514], leaf PE will find corresponding Inter-AS A-D route while sending C-multicast route. It fills source-AS field of the C-multicast route with the corresponding PE Distinguisher Label carried in the Inter-AS A-D route.

6. Backward Compatibility

When some devices do not support MVPN Upstream IDF Election, the following procedures are introduced to support the backward compatibility and end-to-end MVPN service. This section mainly discuss the situation that root PE and leaf PE does not support the IDF election function simultaneously.

6.1. Root PE Not Support IDF Election

When a root PE does not support MVPN Upstream IDF Election, the UMH route sent by the root PE will not carry IDF Negotiation Community. Other root PEs will check all received UMH routes with the same prefix. As long as one of these routes does not carry the community, the IDF election procedures will not be executed and hot root standby will be conducted.

Leaf PE will also check the received UMH routes. When one of the received UMH route does not carry the IDF Negotiation Community, the RPF Set checklist will not be used. The RPF Checking will be based on the normal procedure that only one upstream interface will be considered as the valid upstream interface for certain (C-S,C-G).

6.2. Leaf PE Not Support IDF Election

When leaf PE does not support the IDF Election or it cannot become the leaf of PMSI Tunnel rooted at the main IDF, it must revert back to join the normal PMSI P-Tunnel to receive multicast traffic. A new extra PMSI attribute called "Secondary PMSI Tunnel Attribute" is appended after the existing PTA in the x-PMSI A-D route sent by the primary and standby IDF to identify the warm PMSI Tunnel. The format and content of "Secondary PMSI Tunnel Attribute" are same as PTA defined in [RFC6514]. The attribute type (Attr Type) field will be allocated by IANA.

When leaf PE does not support the IDF Election function, it cannot recognise these Secondary attribute. Therefore, the leaf PE will join the PMSI Tunnel identified by the normal PTA defined in [RFC6514]. The explicit tracking defined in [RFC6514] will be conducted. Even if leaf PE supports the IDF Election function, it

may not be able to join the "Warm Tunnel" due to certain reasons, such as local policy. Under this condition, leaf PE can also join the original PMSI Tunnel.

When leaf PE decide to join the "Warm PMSI Tunnel" identified by the Secondary PTA, it will send C-Multicast route to root PEs. The x-PMSI Leaf A-D route will carry a "Secondary Indication Community". The value of this community will be allocated by IANA.

When root PEs are performing the IDF Election, only IDF can forward corresponding traffic into the PMSI Tunnel identified by "Secondary PTA". One multicast traffic can be carried by both the normal PMSI Tunnel and warm PMSI Tunnel simultaneously. Only the warm PMSI Tunnel is controlled by the aforementioned IDF Negotiation Status.

7. Security Considerations

This document follows the security considerations specified in [RFC6513] and [RFC6514]. In addition, because the establishment of segmented Inter-AS PMSI tunnel is introduced by using Intra-AS AD routes in this document, the Originator’s IP addresses are exposed between ASes which may cause some security risks in the scenarios of different service providers for different ASes. In order to reduce the impact, the Intra-AS AD routes to be leaked between ASes MUST be controlled under security policies so that the numbers of the leaked Originator’s IP addresses can be reduced.

8. IANA Considerations

This document defines a new BGP Community called IDF negotiation Community, of which the value will be allocated from IANA for each negotiation mode individually. The BFD Discriminator Attribute defined in [RFC9026] is reused and the value of BFD Mode is recommended to be 2 in this document, which will be reviewed by IANA.

This document defines a new UMH route type for MVPN, of which the value is recommended to be 8 and will be reviewed by IANA. This document defines a new BGP Extended Community called "Root Distinguisher", this Community is of an extended type and is transitive, the Type and Sub-Type are TBD and will be allocated from IANA.

This document defines a new PMSI Tunnel Attribute called "Secondary PMSI Tunnel Attribute" and a new community called "Secondary Indication Community". Their attribute types will be allocated by IANA.

9. Acknowledgements

The authors wish to thank Jingrong Xie and Jeffrey Zhang, for their reviews, comments and suggestions.

10. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5798] Nadas, S., Ed., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, DOI 10.17487/RFC5798, March 2010, <https://www.rfc-editor.org/info/rfc5798>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC6515] Aggarwal, R. and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, DOI 10.17487/RFC6515, February 2012, <https://www.rfc-editor.org/info/rfc6515>.

[RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, DOI 10.17487/RFC6625, May 2012, <https://www.rfc-editor.org/info/rfc6625>.

[RFC7524] Rekhter, Y., Rosen, E., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area Point-to-Multipoint (P2MP) Segmented Label Switched Paths (LSPs)", RFC 7524, DOI 10.17487/RFC7524, May 2015, <https://www.rfc-editor.org/info/rfc7524>.

[RFC7716] Zhang, J., Giuliano, L., Rosen, E., Ed., Subramanian, K., and D. Pacella, "Global Table Multicast with BGP Multicast VPN (BGP-MVPN) Procedures", RFC 7716, DOI 10.17487/RFC7716, December 2015, <https://www.rfc-editor.org/info/rfc7716>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC9026] Morin, T., Ed., Kebler, R., Ed., and G. Mirsky, Ed., "Multicast VPN Fast Upstream Failover", RFC 9026, DOI 10.17487/RFC9026, April 2021, <https://www.rfc-editor.org/info/rfc9026>.

Authors' Addresses

Fanghong Duan
Huawei Technologies
Email: duanfanghong@huawei.com

Siyu Chen
Huawei Technologies
Email: chensiyu27@huawei.com

Yisong Liu
China Mobile
Email: liuyisong@chinamobile.com

Heng Wang
Ruijie Networks Co., Ltd.
Email: wangheng1@ruijie.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 24 July 2023

H. Wang
J. Dong
Huawei
G. Mirsky
Ericsson
Y. Huang
Huawei

20 January 2023

Advertising S-BFD Discriminators in BGP
draft-wang-bess-sbfd-discriminator-05

Abstract

Seamless Bidirectional Forwarding Detection (S-BFD) is a simplified BFD mechanism. It eliminates most negotiation aspects and provides advantages such as fast configuration injection. S-BFD is especially useful in multi-homing PE scenarios and reduces resource overheads on the dual-homing PEs. Although S-BFD is simpler than BFD, a large number of manual configurations are required when there are a large number of PEs.

This document provides the mechanism of distributing S-BFD discriminators with VPN service routes, which simplifies S-BFD deployment for VPN services.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft will expire on 24 July 2023.

Copyright Notice

Table of Contents

1. Introduction

In deploying some network services, carriers usually deploy active and standby nodes to improve network reliability. In this way, a protection switchover can be performed quickly when there is a fault at the active node or a failure on the working path between PEs. To accelerate fault detection, BFD is usually used. Traditionally, BFD sessions need to be provisioned on both ends of the BFD session, which occupies the resources of both PEs.

[RFC7880] defines Seamless Bidirectional Forwarding Detection (S-BFD), a simplified mechanism for using BFD with a large proportion of negotiation aspects eliminated, thus providing benefits such as quick provisioning, as well as improved control and flexibility for network nodes initiating path monitoring. This mechanism is especially useful in asymmetric scenarios, such as the 3PE scenarios. At the dual-homing PE nodes, BFD does not need to be used to detect the single-homed PE node. In this scenario, S-BFD greatly saves resources on the dual-homing nodes.

To deploy S-BFD, the initiator needs to know the reflector’s ip address and discriminator . When a large number of PEs need to be deployed, even the deployment of S-BFD can become complicated. [RFC7883] and [RFC7884] introduced an IGP-based S-BFD discriminator advertisement mechanism to simplify S-BFD deployment. Since VPN services may be deployed across an area or AS boundaries, the IGP-based S-BFD mechanism does not apply in this case.

This document proposes a mechanism to distribute S-BFD discriminator information with BGP service routes. It allows advertising S-BFD discriminator across multiple domains and achieves on-demand end-to-end S-BFD session provisioning for specific BGP service routes.

2. Terminology

BFD : Bidirectional Forwarding Detection

S-BFD : Seamless Bidirectional Forwarding Detection

3PE : One PE connect to dual-homed PEs scenario

APE : Access PE, used to access users

SPE: Service PE, used to support service for users

UCE: User CE

SCE: Service CE

3. Typical Scenarios

In some inter-domain VPN service deployments, only one of a pair of interconnected PEs benefits from monitoring the status of the connection. In such a case, using S-BFD [RFC7880] is advantageous as it reduces the load on one of the PEs while providing the benefit of faster convergence. The following subsections provide some examples of SRv6-based VPN services to show the benefits of using S-BFD.

For SRv6 services, two types of paths are used to support the service. One is service over SRv6 BE, the other is service over SRv6 Policy. For the service over SRv6 BE case, the VPNSID is used to resolve the forwarding information. Thus an S-BFD session is needed to detect the reachability of the VPNSID. The session is an IP-routed S-BFD, and the SRv6 locator of the remote VPN SID can be used as the destination identifier of the S-BFD session. For the service over SRv6 Policy, the <nexthop, color> in the service route is used to resolve to an SRv6 Policy. Then an S-BFD session is needed to detect the reachability of the SRv6 Policy.

3.1. EVPN Layer 3 Service Over SRv6 BE Use Case

```
          /---------------------\  /--------------------\
          |                     |  |                    |
+----+    | +----+      +-----+ |  |+-----+      +----+ |
|UCE1|---|-|APE1|------|ASBR1|-|--|ASBR3|------|SPE1| |
+----+    | +----+ \    ,-----+ |  |+-----+\    /+----+ |
          |         \  /        |  |        \  /       `|
          | ....     \/         |  |         \/         |', +----+
          |          /\         |  |         /\         |  .|SCE1|
          |         /  \        |  |        /  \        |-` +----+
+----+    | +----+ /    '-----+ |  |+-----+-    '+-----`|
|UCEn|---|-|APEn|------|ASBR2|-|--|ASBR4|------|SPE2| |
+----+    | +----+      +-----+ |  |+-----+      +----+ |
          |                     |  |                    |
          |      AS65001        |  |       AS65002      |
          \---------------------/  \--------------------/
```

Figure 1: EVPN Layer 3 Service Over SRv6 BE

Figure 1 shows a SRv6 BE based seamless scenario. UCE is single-homed to APE, and SCE is dual-homed to SPE1 and SPE2. The service is across multiple ASes.

SCE1 accesses SPE1 and SE2 through Layer 3 and advertises its private network routes to them. SPE1 and SPE2 encapsulate the routes into Type 5 routes in the EVPN format and sends them to APE1. After receiving Type 5 routes advertised by SPE1 and SPE2, APE1 generates primary and backup entries for the routes to speed up service switchover. In this scenario, the SRv6 BE service mode is used. APE1 will resolve SPE1's VPN routes reachability through the VPN SID. To ensure that APE1 can properly route to PE1, PE1 needs to advertise its own locator route. The advertisement of the locator route is not in the scope of this document.

To speed up the fault detection, we may configure an S-BFD session on APE1 to detect SPE1 or SPE2's reachability. In traditional mode, a discriminator needs to be assigned by SPE1 and SPE2, and two S-BFD

sessions need to be configured on APE1 to detect the VPN SID's reachability of SPE1 and SPE2. It needs to generate an S-BFD session with the destination set to the VPN SID. To reduce the number of S-BFD sessions, locator-based S-BFD sessions can be used instead of S-BFD sessions for VPNSIDs.

There are a large number of such APEs that exist on the network. Each APE is configured with several S-BFD sessions to detect PE1 and PE2, which increases the deployment complexity.

3.2. EVPN Layer 3 Service Over SPv6 Policy Use Case

```
            /---------------------\  /--------------------\
            |                     |  |                    |
+----+      | +----+      +-----+ |  |+-----+      +----+ |
|UCE1|---|-|APE1|------|ASBR1|-|--|ASBR3|------|SPE1| |
+----+      | +----+ \    ,-----+ |  |+-----+\    /+----+ |
            |         \  /        |  |        \  /       `|
            | ....     \/         |  |         \/         |', +----+
            |          /\         |  |         /\         |  .|SCE1|
            |         /  \        |  |        /  \        |-` +----+
+----+      | +----+ /    '-----+ |  |+-----+-    '+-----`|
|UCEn|---|-|APEn|------|ASBR2|-|--|ASBR4|------|SPE2| |
+----+      | +----+      +-----+ |  |+-----+      +----+ |
            |                     |  |                    |
            |      AS65001        |  |       AS65002      |
            \---------------------/  \--------------------/
```

Figure 2: EVPN Layer 3 Service Over SRv6 Policy

Figure 2 shows an SRv6 Policy scenario. SCE1 is dual-homed to SPE1 and SPE2, and UCE1 is accessed to APE1. SPE1, SPE2, and APE1 are cross BGP ASes.

SCE1 accesses SPE1 and SPE2 through Layer 3 and advertises its private network routes to APE1. SPE1 and SPE2 encapsulate the routes into Type 5 routes in the EVPN format and sends them to APE1.

After receiving Type 5 routes advertised by SPE1 and SPE2, APE1 generates primary and backup entries for the routes, speeding up service switchover. APE1 parses the tunnel based on the <nexthop, color> of the service routes advertised by SPE1 and SPE2, and matches an SRv6 Policy. After receiving the traffic from UCE1 to SCE1, APE1 encapsulates and forwards the traffic based on the SRv6 Policy.

An S-BFD session needs to be established for these SRv6 Policy-based forwarding paths to swiftly detect the availability of the paths. When detecting a fault on the SRv6 Policy path of the primary service route, services can be swiftly switched to the backup path, providing more reliable protection for services.

There are a large number of such PEs that exist on the network. Each PE is configured with several S-BFD sessions to detect PE1 and PE2, which increases the deployment complexity.

Certainly, this scenario may also be implemented in other methods. For example, when provisioning an SRv6 policy, an S-BFD session can be provisioned. While in some cases, it would be more efficient if the S-BFD session could be provisioned based on the demand of the services.

4. Procedure

4.1. BGP Encoding

[RFC9026] specifies the "BFD Discriminators" (38) attribute, which is an optional transitive BGP attribute that conveys the Discriminators and other optional attributes used to establish BFD sessions.

In [RFC9026], the BFD Discriminators attribute is used to transmit P2MP BFD session creation information MVPN scenarios. For non-multicast services, such as L3VPN services, L2VPN services, EVPN services and native IP services, BFD discriminators are also required for creating an S-BFD session. This document reuses the BFD Discriminators attribute and defines new BFD modes for some of these services.

The format of BFD Discriminator attribute is shown as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+
|    BFD Mode   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       BFD Discriminator                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                         Optional TLVs                         ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Format of the BFD Discriminator Attribute

o BFD Mode:

The BFD Mode field is 1 octet. [RFC9026] defines only the P2MP BFD session for MVPN. This document defines two new types of S-BFD session types for the scenarios in Section 3.

As described in the preceding scenarios, there are two types of S-BFD sessions for SRv6 services. For service over SRv6 BE, an IP-routed S-BFD session needs to be created to detect the reachability of the SRv6 locator. For service over SRv6 Policy, an S-BFD session for SRv6 Policy needs to be created to detect the reachability of the SRv6 Policy. Thus two new BFD modes are introduced:

* S-BFD for SRv6 Locator Session Mode, which is dedicated to detecting the locator. The type value is to be allocated as described in Section 6.

* S-BFD for Common Session Mode, which is for general S-BFD session. The type value is to be allocated as described in Section 6. This mode is not only for SRv6, but also can be used for other scenarios.

o BFD Discriminators:

The field length is 4 octets. Used to specify the discriminator for S-BFD session.

o Optional TLVs:

Variable-length fields are optional. Indicates the additional information required for creating a S-BFD session. The format is as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |     Length    |            Value              ...
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Format of the Optional TLV

If a transit node changes the next hop or reassigns a VPN SID when advertising a route, the transit node needs to use the locally allocated S-BFD discriminator for the S-BFD discriminator attribute. Suppose the transit node does not recognize the S-BFD Discriminator attribute in the received route and continues to advertise the route to the remote PE. In that case, the receiver may use incorrect information when creating an S-BFD session. Therefore, the advertised S-BFD Discriminator attribute also needs to carry the IP address of the originator of the discriminator for receiver side verification.

For the two BFD modes defined in this document, the "Source IP Address" TLV as defined in [RFC9026] MUST be carried in the BFD Discriminator attribute. If the mode is "S-BFD for SRv6 Locator Session", the SRv6 Locator address MUST be used for the service is carried in the TLV. If the mode is "S-BFD for Common Session", the next-hop address MUST be used for the service is carried in the TLV.

4.2. Router Procedure

In BGP address families, such as L3VPN or EVPN, routes can carry the S-BFD Discriminator attribute as required so that S-BFD sessions can be established based on the attribute. The following uses S-BFD for SRv6 Locator as an example. If mode is set to S-BFD for Common Session, the processing method is similar.

4.2.1. Egress Node Process

As shown in figure 1, the S-BFD discriminator is configured on PE1. After obtaining the information, BGP encapsulates the attribute into the EVPN route and sets the BFD Mode to S-BFD for Locator Session, when advertising the EVPN route. The Discriminator value is local discriminator value. The optional TLV carries the local PE’s locator address used by the VPN.

4.2.2. Transit Node Process

Here is the end-to-end SRv6 BE scenario. The ASBR does not re-allocate the VPN SID. Thus, the ASBR does not require to modify the VPN SID, and not to alter the BFD discriminator attribute.

4.2.3. Ingress Node Process

After receiving the EVPN Type 5 routes from PE1 and PE2, PE3 imports the routes to the VRF of PE3 based on the route targets. Routes triggers establish the S-BFD sessions based on <S-BFD discriminator, locator ip>.

Then, routes with the same prefix from PE1 and PE2 form primary and backup paths. When the primary path or the egress node is in fault, S-BFD detects that fault and forms switch to backup path quickly.

To avoid the waste of redundant resources, assume that the ASBR re-assigns the SID in Option B and the ASBR does not recognize the attribute. In this case, the SID and locator carried in the route received by PE3 do not match the Source IP carried in the Optional TLV in the BFD attribute. Therefore, PE3 does not need to establish an S-BFD session to remote PE, which can avoid resource waste.

5. Error handling

Error handling complies with [RFC7606]. In this document, the BFD discriminator information is used only to establish an S-BFD session. Therefore, if the BFD discriminator information is invalid, the BFD attribute will be discard and not transmit to other devices.

For BFD discriminator attribute, the following case will be processed:

o The BFD Discriminator attribute length is smaller than 5 octets, the attribute is invalid.

o The BFD Discriminator value in receiving BFD Discriminator attribute is 0, the attribute is invalid.

For the BFD mode type "S-BFD for SRv6 Locator Session", the following case will be processed:

o If the BFD discriminator attribute doesn’t contain optional TLV with type set to 1, the attribute is invalid.

o If the optional TLV type is 1 but the length is not 16, the attribute is invalid.

o If the optional TLV type is 1 but the value is all 0, the attribute is invalid.

o If multiple Source IP Optional TLVs are carried, the first source IP address should be used as the destination to establish an S-BFD session. For EVPN type 2 MAC-IP routes may use the first and the second IP address because it may carry two SRv6 SIDs with different locators. Other source IP addresses should be ignored.

o If a non-Source IP Optional TLV is carried, the Optional TLV will be ignored.

For the BFD mode type "S-BFD for Common Session", the following case will be processed:

o If the BFD discriminator attribute doesn’t contain optional TLV with type set to 1, the attribute is invalid.

o If the optional TLV type is 1 but the length is not 4 or 16, the attribute is invalid.

o It the optional TLV type is 1 but the value is all 0, the attribute is invalid.

o If multiple Source IP Optional TLVs are carried, only the first source IP address should be used as the destination to establish an S-BFD session. Other source IP addresses should be ignored.

o If a non-Source IP Optional TLV is carried, the Optional TLV will be ignored.

6. IANA Considerations

IANA is requested to assign two new code points from the "BFD Mode" subregistry in the "Border Gateway Protocol (BGP) Parameters" registry.

Value	Description
176	S-BFD for SRv6 Locator Session
177	S-BFD for Common Session

7. Security Considerations

The new S-BFD modes introduced in this document does not introduce any new security risks for BGP.

The BFD attribute is an optional attribute and is mainly used for network services within a single administrative domain. The operator SHOULD ensure this attribute does not propagate across the boundary of the administrative domain. For VPN services, the advertisement range of this attribute is the same as that of VPN routes. Therefore, this attribute is not advertised outside the management domain. For public IPv4 and IPv6 services, the border node of the administrative domain SHOULD be configured not to propagate the BFD attribute to other domains.

When creating an S-BFD session, the initiator verifies the S-BFD session based on routing information. This reduces the number of invalid S-BFD sessions and avoid attribute attack.

8. Acknowledgements

NA

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

9.2. References

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC7880] Pignataro, C., Ward, D., Akiya, N., Bhatia, M., and S. Pallagatti, "Seamless Bidirectional Forwarding Detection (S-BFD)", RFC 7880, DOI 10.17487/RFC7880, July 2016, <https://www.rfc-editor.org/info/rfc7880>.

[RFC7883] Ginsberg, L., Akiya, N., and M. Chen, "Advertising Seamless Bidirectional Forwarding Detection (S-BFD) Discriminators in IS-IS", RFC 7883, DOI 10.17487/RFC7883, July 2016, <https://www.rfc-editor.org/info/rfc7883>.

[RFC7884] Pignataro, C., Bhatia, M., Aldrin, S., and T. Ranganath, "OSPF Extensions to Advertise Seamless Bidirectional Forwarding Detection (S-BFD) Target Discriminators", RFC 7884, DOI 10.17487/RFC7884, July 2016, <https://www.rfc-editor.org/info/rfc7884>.

[RFC9026] Morin, T., Ed., Kebler, R., Ed., and G. Mirsky, Ed., "Multicast VPN Fast Upstream Failover", RFC 9026, DOI 10.17487/RFC9026, April 2021, <https://www.rfc-editor.org/info/rfc9026>.

Authors' Addresses

Haibo Wang
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: rainsword.wang@huawei.com

Jie Dong
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: jie.dong@huawei.com

Greg Mirsky
Ericsson
Email: gregimirsky@gmail.com

Yang Huang
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China
Email: yang.huang@huawei.com