BESS   Yisong Liu
Internet Draft   China Mobile
Intended status: Standards Track   C. Lin
Expires: April 5, 2024   M. Chen
                                               New H3C Technologies
                                                          Y. Liu
                                                            ZTE
                                                  October 8, 2023

SRv6 Service SID Flag Extension for Multi-homed SRv6 BGP Services
draft-liu-bess-multihome-srv6-service-sid-flag-02

Abstract

   In some multihoming SRv6 L3VPN and EVPN scenarios, there are
   requirements for the egress PE to advertise multiple SRv6 Service
   SIDs for the same service, such as egress fast reroute and anycast
   load balancing. This document defines No-Further-FRR-flag and
   Anycast-flag for SRv6 Service SIDs carried in BGP messages.

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), its areas, and its working groups. Note that
   other groups may also distribute working documents as Internet-
   Drafts.

   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."

   The list of current Internet-Drafts can be accessed at
   http://www.ietf.org/ietf/1id-abstracts.txt

   The list of Internet-Draft Shadow Directories can be accessed at
   http://www.ietf.org/shadow.html

   This Internet-Draft will expire on April 5, 2024.

Copyright Notice

Table of Contents

1. Introduction

[RFC9252] defines procedures and messages for SRv6-based BGP
services, including Layer 3 Virtual Private Network (L3VPN),
Ethernet VPN (EVPN), and Internet services. In some multihoming
scenarios, there are requirements for the egress PE to advertise
multiple SRv6 Service SIDs for the same service, such as egress fast
reroute and anycast load balancing. And those SIDs need to be
identified in the BGP messages.

This document defines No-Further-FRR-flag and Anycast-flag for SRv6
Service SIDs carried in BGP messages.

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. No-Further-FRR Service SID

2.1. Use Case

2.1.1. SRv6 L3VPN Multihoming

   In the multihoming SRv6 L3VPN scenarios, two egress PEs may
   establish a backup path between them and use it as the protection of
   PE-CE link failure.

   Take the network in Figure 1 as an example. When traffic goes from
   CE1 to CE2, it may be load-balanced between PE2 and PE3 or only
   forwarded to the main egress PE. If the link PE2-CE2 fails, PE2 can
   still forward the traffic for CE2 by sending it over the backup path
   to PE3 (and similarly for PE3 if link2 fails).

```
                         +-----+
                         | CE1 |
                         +-----+
                            |
                            |
                         +-----+
     ------------------- | PE1 |***************
        ^                +-----+              *
        |           /          \             *
        |          /            \           *
        |       P1                P2        *
        |       .                 .    +------+
     SRv6 VPN   .        *************.*******|BGP-RR|
        |       .      *              .   +------+
        |       P3   *             P4       *
        |       |   *              |        *
        |       |  *               |        *
        v     +-----+   Backup   +-----+    *
     --------- | PE2 |############| PE3 |*****
              +-----+    Path     +-----+
                  \             /
                   \           /
                   +-----+
                   | CE2 |
                   +-----+


                         Figure 1
```

   Examples of BGP routes advertised by PE2 and PE3 are as following:

```
     BGP Route by PE2:
       VPN Prefix of CE2:
         BGP Prefix SID Attr:
           SRv6 L3 Service TLV:
             SRv6 SID Information sub-TLV:
               SID: SID-2
                 Behavior: End.DT46

     BGP Route by PE3:
       VPN Prefix of CE2:
         BGP Prefix SID Attr:
           SRv6 L3 Service TLV:
             SRv6 SID Information sub-TLV:
               SID: SID-3
                 Behavior: End.DT46
```

Examples of FIB entries for L3VPN service SID on PE2 and PE3 are as following:

```
     FIB on PE2:
       SID-2:
         Primary Next-hop: CE2
         Backup Next-hop: Service SRv6 SID-3

     FIB on PE3:
       SID-3:
         Primary Next-hop: CE2
         Backup Next-hop: Service SRv6 SID-2
```

However, suppose CE2 is down. PE2 will think PE2-CE2 link is down and send traffic to PE3 over the backup path. PE3 will also think PE3-CE3 link is down and send the traffic back to PE2 over the backup path. So, traffic will loop between PE2 and PE3 until BGP convergence.

The traffic forwarding when CE2 fails is as following:

```
+======+=============+=======+==============+
| Node | Packet      | Next  | Comment      |
+======+=============+=======+==============+
| PE1  | <SID-2> pkt | PE2   |              |
+------+-------------+-------+--------------+
| PE2  | pkt         | CE2   | PE2-CE2 down |
+------+-------------+-------+--------------+
| PE2  | <SID-3> pkt | PE3   | FRR          |
+------+-------------+-------+--------------+
| PE3  | pkt         | CE2   | PE3-CE2 down |
+------+-------------+-------+--------------+
| PE3  | <SID-3> pkt | PE2   | FRR          |
+------+-------------+-------+--------------+
| PE2  | --          | CE2   | PE2-CE2 down |
+------+-------------+-------+--------------+
| PE2  | <SID-3> pkt | PE3   | FRR          |
+------+-------------+-------+--------------+
| ...  |             |       | Loop!        |
+------+-------------+-------+--------------+
```

2.1.2. SRv6 EVPN Multihoming

The EVPN services include Designated Forwarder (DF) election
procedure.

In All-Active mode, all PEs are allowed to forward unicast traffic,
which is similar with the L3VPN case in Section 2.1.

In Single-Active mode, only DF is allowed to forward unicast
traffic, and it requires additional considerations in FRR.

2.2. Solution

Each egress PE advertises an additional SRv6 Service SID in BGP
routes which is called No-Further-FRR SID.

The owner of No-Further-FRR SID will not provide local FRR for it.
When the next-hop of No-Further-FRR SID is down, like PE-CE link
failure or CE node failure, the PE will drop packets rather than
apply FRR.

The No-Further-FRR SID can used by other PE as the protection of
local PE-CE link failure, without worrying about the looping
problem.

To support backwards compatibility and BGP RR deployment, both the
normal SRv6 Service SID and the No-Further-FRR SID MAY be advertised

together. A No-Further-FRR flag is used to indicate the No-Further-FRR SID.

Detailed BGP extensions will be described in Section 4.
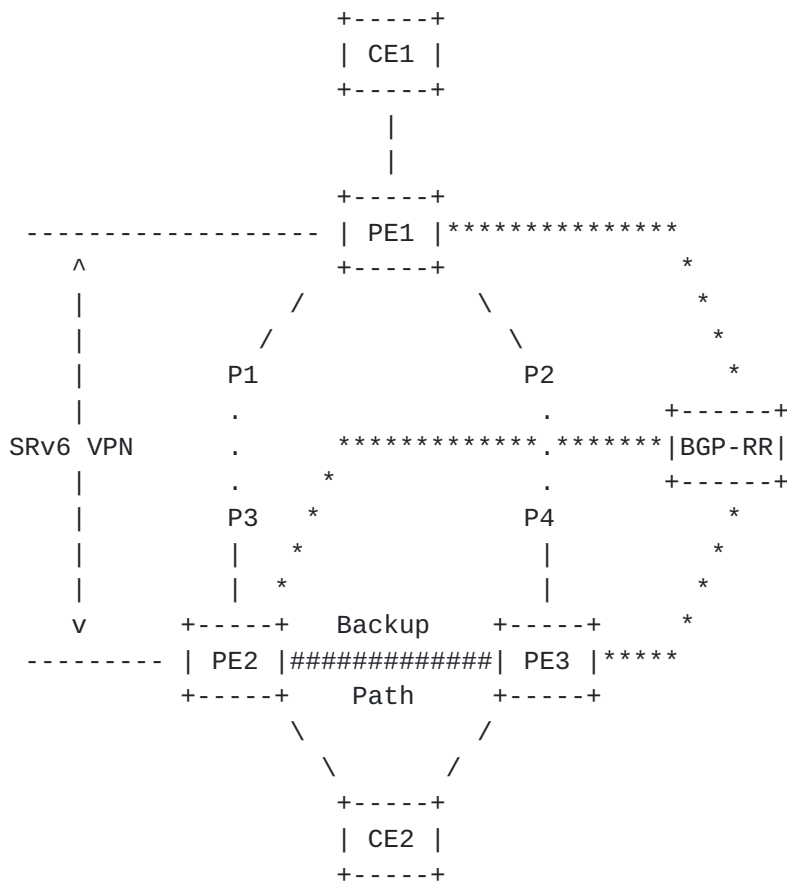
Still taking the network in Figure 1 as an example, the BGP routes advertised by PE2 and PE3 are as following:

```
BGP Route by PE2:
  VPN Prefix of CE2:
    BGP Prefix SID Attr:
      SRv6 L3 Service TLV:
        SRv6 SID Information sub-TLV:
          SID: SID-21
            Behavior: End.DT46(L3VPN) or End.DX2/End.DT2U(EVPN)
        SRv6 SID Information sub-TLV:
          SID: SID-22
            Behavior: End.DT46(L3VPN) or End.DX2/End.DT2U(EVPN)
        Flag: No-Further-FRR

BGP Route by PE3:
  VPN Prefix of CE2:
    BGP Prefix SID Attr:
      SRv6 L3 Service TLV:
        SRv6 SID Information sub-TLV:
          SID: SID-31
            Behavior: End.DT46(L3VPN) or End.DX2/End.DT2U(EVPN)
        SRv6 SID Information sub-TLV:
          SID: SID-32
            Behavior: End.DT46(L3VPN) or End.DX2/End.DT2U(EVPN)
        Flag: No-Further-FRR
```

The FIB entries for L3VPN service SID on PE2 and PE3 are as following:

```
     FIB on PE2:
       SID-21:
         Primary Next-hop: CE2
         Backup Next-hop: Service SRv6 SID-32
       SID-22 (No-Further-FRR):
         Primary Next-hop: CE2

     FIB on PE3:
       SID-31:
         Primary Next-hop: CE2
         Backup Next-hop: Service SRv6 SID-22
       SID-32 (No-Further-FRR):
         Primary Next-hop: CE2
```

   After adopting the proposed solution, if CE fails, PE2 will think
   PE2-CE2 link is down and send traffic to PE3 by using the No-
   Further-FRR SID-32. PE3 will also think PE3-CE3 link is down, but
   PE3 will drop the packets rather than apply FRR.

   The traffic forwarding when CE2 fails is as following:

```
    +======+==============+=======+==============+
    | Node | Packet       | Next  | Comment      |
    +======+==============+=======+==============+
    | PE1  | <SID-21> pkt | PE2   |              |
    +------+--------------+-------+--------------+
    | PE2  | pkt          | CE2   | PE2-CE2 down |
    +------+--------------+-------+--------------+
    | PE2  | <SID-32> pkt | PE3   | FRR          |
    +------+--------------+-------+--------------+
    | PE3  | pkt          | CE2   | PE3-CE2 down |
    +------+--------------+-------+--------------+
    | PE3  | -            | -     | Drop         |
    +------+--------------+-------+--------------+
```

2.2.1. Consideration for EVPN Single-Active Mode

   The processing of the No-Further-FRR SID should apply an override to
   EVPN DF-Election and bypass the local blocking state on the AC,
   until EVPN control plane reconverges.

3. Anycast Service SID

   In the multihoming SRv6 L3VPN and EVPN scenarios, anycast Service
   SID may be used to advertise the same service at different egress
   PEs, which can improve service reliability and load balancing.
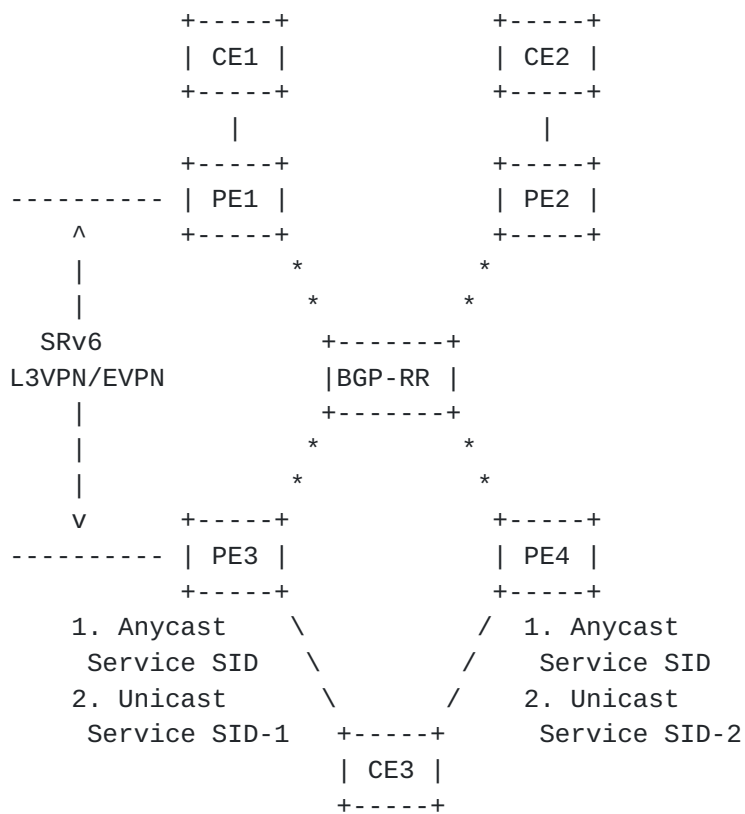
```
             +-----+              +-----+
             | CE1 |              | CE2 |
             +-----+              +-----+
                |                    |
             +-----+              +-----+
 ----------  | PE1 |              | PE2 |
      ^      +-----+              +-----+
      |              *          *
      |               *        *
    SRv6               +-------+
    L3VPN/EVPN         |BGP-RR |
      |                +-------+
      |               *        *
      |              *          *
      v      +-----+              +-----+
 ----------  | PE3 |              | PE4 |
             +-----+              +-----+
     1. Anycast    \          /  1. Anycast
      Service SID   \        /    Service SID
     2. Unicast      \      /   2. Unicast
      Service SID-1   +-----+      Service SID-2
                      | CE3 |
                      +-----+


   PE1:
     VPN Traffic Policy:
       PE3 & PE4 Load Balancing
     FIB Entry for VPN Traffic:
       Next-hop: Anycast Service SID

   PE2:
     VPN Traffic Policy:
       PE3 Active, PE4 Backup
     FIB Entry for VPN Traffic:
       Primary Next-hop: Unicast Service SID-1
       Backup Next-hop: Unicast Service SID-2
```

                         Figure 2

   As shown in Figure 2, PE3 and PE4 use the same anycast SRv6 Service
   SID for the VPN service of CE3. The ingress PE1 encapsulates the
   payload in an outer IPv6 header where the destination address is
   that anycast SRv6 Service SID. The packets from CE1 can reach CE3
   through either PE3 or PE4. Assume that the path from PE1 to PE3 and
   the path from PE1 to PE4 have the same cost. The traffic flows will
   be load balanced between PE3 and PE4.

PE3 and PE4 also have unicast SRv6 Service SIDs, which are SID-1 and
SID-2, for the VPN service of CE3. The ingress PE2 uses SID-1 as the
primary SRv6 Service SID, and SID-2 as backup. The packets from CE2
will be forwarded to CE3 through PE3. If any failure occurs on the
path to PE3, service will be switched to PE4.

Since ingress PE1 and PE2 have different strategies for the control
of VPN traffics, egress PE3 and PE4 each need to advertise two SRv6
Service SIDs, an anycast SID for ingress PE1 and a unicast SID for
ingress PE2. Local export policy may be used by egress PE3 and PE4
to control which SID is advertised to ingress PE1 and which is
advertised to ingress PE2. However, if BGP Route Reflector is
deployed, both the anycast Service SID and the unicast Service SID
will be advertised to RR and reflected to ingress PEs, and the
receiver has to choose which Service SID to use. So, it is required
to identify which Service SID is anycast and which Service SID is
unicast, when both two SIDs are advertised in BGP messages.

4. Extensions for BGP

This document defines two new flags in the SRv6 Service SID Flags
field of the SRv6 SID Information Sub-TLV [RFC9252]:

```
 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
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | SRv6 Service  |    SRv6 Service                |              |
 | Sub-TLV       |    Sub-TLV                     |              |
 | Type=1        |    Length                      |   RESERVED1  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   SRv6 SID Value (16 octets)                                //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 | Svc SID Flags |    SRv6 Endpoint Behavior      |    RESERVED2 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   SRv6 Service Data Sub-Sub-TLVs                            //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

   Svc SID Flags:

    0 1 2 3 4 5 6 7
   +-+-+-+-+-+-+-+-+
   |N|A|           |
   +-+-+-+-+-+-+-+-+
```

o N-flag: No-Further-FRR flag. When set, the associated SID has no
   fast reroute protection.

o A-flag: Anycast flag. When set, the associated SID is anycast.

The new-defined flags can be used for the SRv6 Service SIDs of L3
and L2 services, such as End.DX4, End.DT4, End.DX6, End.DT6,
End.DT46. End.DX2, End.DX2V, End.DT2U, etc.

5. Backward Compatibility

According to [RFC9252],

o Any unknown flags in the SRv6 Service SID Flags field MUST be
  ignored by the receiver.

o When multiple SRv6 SID Information Sub-TLVs are present, the
  ingress PE SHOULD use the SRv6 SID from the first instance of the
  Sub-TLV.

If there are PE routers not supporting the flags defined in this
document, the egress PE may expect those routers to use the first
SID and ignore the new-defined flags.

6. Security Considerations

TBD.

7. IANA Considerations

This document defines the following bits in the SRv6 Service SID
Flags field of SRv6 SID Information Sub-TLV:

```
TLV Code Point    Value
-----------------------------------------------------------
TBD               N-flag
TBD               A-flag
```

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
          Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
          2119 Key Words", BCP 14, RFC 8174, May 2017

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene,
          B., Zhuang, S., and J. Rabadan, "BGP Overlay Services
          Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI
          10.17487/RFC9252, July 2022, <https://www.rfc-
          editor.org/info/rfc9252>.

Authors' Addresses

    Yisong Liu
    China Mobile
    China
    Email: liuyisong@chinamobile.com

    Changwang Lin
    New H3C Technologies
    China
    Email: linchangwang.04414@h3c.com

    Mengxiao Chen
    New H3C Technologies
    China
    Email: chen.mengxiao@h3c.com

    Yao Liu
    ZTE
    China
    Email: liu.yao71@zte.com.cn