Network Working Group
Internet Draft
Intended status: Informational
Expires: May 2, 2021

L. Dunbar
J. Guichard
Futurewei
Ali Sajassi
Cisco
J. Drake
Juniper
B. Najem
Bell Canada
Ayan Barnerjee
D. Carrel
Cisco

November 2, 2020

BGP Usage for SDWAN Overlay Networks
draft-ietf-bess-bgp-sdwan-usage-01

Abstract

The document describes three distinct SDWAN scenarios and discusses the applicability of BGP for each of those scenarios. The goal of the document is to demonstrate how BGP-based control plane is used for large scale SDWAN overlay networks with little manual intervention.

SDWAN edge nodes are commonly interconnected by multiple underlay networks which can be owned and managed by different network providers.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Here are some key characteristics of "SDWAN" networks:

- Augment of transport, which refers to utilizing overlay paths over different underlay networks. Very often there are multiple parallel overlay paths between any two SDWAN edges, some of which are private networks over which traffic can traverse with or without encryption, others require encryption, e.g. over untrusted public networks.
- Enable direct Internet access from remote sites, instead hauling all traffic to Corporate HQ for centralized policy control.
- Some traffic flows can be forwarded based on their application identifiers instead of based on destination IP addresses, by the edge nodes placing the traffic flows onto specific overlay paths based on their application requirement.
- The traffic flows forwarding can also be based on specific performance criteria (e.g. packets delay, packet loos, jitter) to provide better application performance by choosing the right underlay that meets or exceeds the specified criteria.

[Net2Cloud-Problem] describes the network related problems that enterprises face to connect enterprises' branch offices to dynamic workloads in different Cloud DCs, including using SDWAN to aggregate multiple paths provided by different service providers to achieve better performance and to accomplish application ID based forwarding.

Even though SDWAN has been positioned as a flexible way to reach dynamic workloads in third party Cloud data centers over different underlay networks, scaling becomes a major issue when there are hundreds or thousands of nodes to be interconnected by an SDWAN overlay networks.

BGP is widely used by underlay networks. This document describes using BGP for edge nodes to exchange information across the SDWAN overlay networks.

2. Conventions used in this document

Cloud DC: Third party data centers that host applications and workloads owned by different organizations or tenants.

Controller: Used interchangeably with SDWAN controller to manage SDWAN overlay path creation/deletion and monitor the path conditions between sites.

CPE: Customer Premise Equipment

CPE-Based VPN: Virtual Private Secure network formed among CPEs. This is to differentiate from more commonly used PE-based VPNs [RFC 4364].

Homogeneous SDWAN: A type of SDWAN network in which all traffic to/from the SDWAN edge nodes has to be encrypted regardless of underlay networks. For lack of better terminology, we call this Homogeneous SDWAN throughout this document.

ISP: Internet Service Provider

NSP: Network Service Provider. NSP usually provides more advanced network services, such as MPLS VPN, private leased lines, or managed Secure WAN connections, many times within a private trusted domain, whereas an ISP usually provides plain internet services over public untrusted domains.

PE: Provider Edge

SDWAN Edge Node: an edge node, which can be physical or virtual, maps the attached clients' traffic to the wide area network (WAN) overlay tunnels.

SDWAN: Software Defined Wide Area Network. In this document, "SDWAN" refers to networks that allow augment of transport, application ID and/or performance based forwarding, and direct internet breakout at remote sites.

SDWAN IPsec SA: IPsec Security Association between two SDWAN ports or nodes.

SDWAN over Hybrid Networks: SDWAN over Hybrid Networks typically have edge nodes utilizing bandwidth resources from multiple service providers. In Hybrid SDWAN network, packets over private networks can go natively without encryption and are encrypted over the untrusted network, such as the public Internet.

WAN Port: A Port or Interface facing an ISP or Network Service Provider (NSP), with address (usually public routable address) allocated by the ISP or the NSP.

C-PE: SDWAN Edge node, which can be CPE for customer managed SDWAN, or PE that is for provider managed SDWAN services).

ZTP: Zero Touch Provisioning

3. Use Case Scenario Description and Requirements

SDWAN networks can have different topologies and have different traffic patterns. To make it easier for the focused discussion in subsequent drafts on SDWAN control plane and data plane, this section describes several SDWAN scenarios that may have different impact on their corresponding control planes & data planes.

3.1. Requirements

3.1.1. Supporting Multiple SDWAN Segmentations

The term "network segmentation", a.k.a. SDWAN instances, is referring to the process of dividing the network into logical sub-networks using isolation techniques on a forwarding device such as a switch, router, or firewall. For a homogeneous network, such as MPLS VPN or Layer 2 network, VRF or VLAN are used to achieve the network segmentation.

As SDWAN is an overlay network arching over multiple types of networks, MPLS L2VPN/L3VPN or pure L2 underlay can continue using the VRF, VN-ID or VLAN to differentiate SDWAN network segmentations. For public internet, the IPsec inner encapsulation header can carry the SDWAN Instance Identifier to differentiate the packets belonging to different SDWAN instances.

BGP already has the capability to differentiate control packets for different network instances. When using BGP for SDWAN, the SDWAN segmentations can be differentiated by the SDWAN Target ID in the BGP Extended Community in the same way as VPN instances being represented by the Route Target. Same as Route Target, need to use a different name to differentiate from VPN if a CPE supports traditional VPN with multiple VRFs and supports multiple SDWAN Segmentations (instances). The actual SDWAN Target ID encoding is proposed by [SDWAN-EDGE-Discovery].

3.1.2. Client Service Requirement

Client interface of SDWAN nodes can be IP or Ethernet based.

For Ethernet based client interfaces, SDWAN edge should support VLAN-based service interfaces (EVI100), VLAN bundle service interfaces (EVI200), or VLAN-Aware bundling service interfaces. EVPN service requirements are applicable to the Client traffic, as described in the Section 3.1 of RFC8388.

For IP based client interfaces, L3VPN service requirements are applicable.

3.1.3. Application Flow Based Segmentation

Application Flow based Segmentation, also known as SDWAN Traffic Segmentation, enables the separation of the traffic based on the business and the security needs for different users' groups and/or application requirements. Each user group and/or applications may need different isolated topology and/or policies to fulfill the business requirements.

The Application Flow based Segmentation concept is analogous to VLAN (in L2 network) and VRF (in L3 network).

One can think about the Application Flow based Segmentation as a feature that can be provided or enabled on a single SDWAN service (or domain) to a single Subscriber. Each SDWAN Service can have one or more overlay Segments to support the business requirement; each Segment has its own policy, topology and application/user groups. Applications/users' group can belong to more than one Segment.

For example, a retail business requires the point-of-sales (PoS) application in all stores to be isolated from other applications AND routed only to the payment processing entity at a hub site (i.e. hub and spoke); however, the same retail business requires the other applications to be routed to all sites (i.e. multipoint-to multipoint) AND isolated from the PoS application.

In the figure below, the traffic from the PoS application follows a Tree topology, whereas other traffic can be multipoint-to-multipoint topology.

```
                              +--------+
              Payment traffic |Payment |
                +------+----+-+gateway +------+----+-----+
               /      /     | +----+---+      |     \     \
              /      /      |      |          |      \     \
           +-+--+  +-+--+  +-+--+  |   +-+--+  +-+--+  +-+--+
           |Site|  |Site|  |Site|  |   |Site|  |Site|  |Site|
           | 1  |  |  2 |  | 3  |  |   |4   |  |  5 |  | 6  |
           +--+-+  +--+-+  +--|-+  |   +--|-+  +--|-+  +--|-+
              |       |       |    |      |       |       |
            ==+=======+=======+====+======+=======+=======+===
                  Multi-point connection for Other traffic
```

Another example is an enterprise who wants to isolate the traffic for each department and have different topology and policy for different department; the HR department may need to access certain applications that are NOT accessible by the engineering department. In addition, the contractors may have a limited access to the enterprise resources.

3.1.4. Zero Touch Provisioning

Unlike traditional EVPN or L3VPN whose PEs are deployed for long term, SDWAN edge nodes (virtual or physical) deployment at a specific location can be ephemeral. Therefore, Zero Touch Provisioning (ZTP), or Plug and Play, is a common requirement for SDWAN. When an SDWAN edge is physically installed at a location or instantiated on a VM in a Cloud DC, ZTP automates follow-up steps, including updates to the OS, software version, and configuration prior to connection. From network control perspective, ZTP includes the following:

- Upon power up, an SDWAN node can establish transport layer secure connection (such as TLS, SSL, etc.) to its controller whose address can be burned or preconfigured on the device.

- The SDWAN Controller can designate a Local Network Controller in the proximity of the SDWAN node; the Local Network Controller manages and monitor the communication policies of the edge node.

3.1.5. Constrained Propagation of SDWAN Edge Properties

One SDWAN edge node may only be authorized to communicate with a small number of other SDWAN edge nodes. Under this circumstance, the property of the SDWAN edge node cannot be propagated to any other nodes who are not authorized to communicate. But a remote SDWAN edge node upon powering up might not have the proper policies to know who the authorized peers are. Therefore, it is very essential for SDWAN deployment have a central point to distribute the properties of each SDWAN edge node to its authorized peers.

BGP is well suited for this purpose. RFC 4684 has specified the procedure to constrain the distribution of BGP UPDATE to only a subset of SDWAN edges. Basically, each edge node informs the Route Reflector (RR) [RFC4456] on its interested SDWAN instances. The RR only propagates the BGP UPDATE for the relevant SDWAN instances to the edge.

Usually the connection between a SDWAN edge node and its RR is over insecure network. Therefore, upon power up, a SDWAN node needs to establish a secure transport layer connection (TLS, SSL, etc.) to its designated RR. The BGP UPDATE messages need to be sent over the secure channel (TLS, SSL, etc.) to the RR.

```
                              +---+
                 Peer Group 1 |RR |   Peer Group 2
                +======+====+=+   +======+====+=====+
               /      /     | +---+      |     \     \
              /      /      |            |      \     \
           +-+--+  +-+--+  +-+--+      +-+--+  +-+--+  +-+--+
           |C-PE|  |C-PE|  |C-PE|      |C-PE|  |C-PE|  |C-PE|
           | 1  |  |  2 |  | 3  |      |4   |  |  5 |  | 6  |
           +----+  +----+  +----+      +----+  +----+  +----+
                Tenant 1                   Tenant 2
```

Figure 1: Peer Groups managed by RR

Tenant separation is achieved by the SDWAN instance identification represented in control plane and data plane, respectively.

3.2. Scenario #1: Homogeneous WAN

This is referring to a type of SDWAN network with edge nodes encrypting all traffic over WAN to other edge nodes, regardless of whether the underlay is private or public. For lack of better terminology, we call this Homogeneous SDWAN throughout this document.

Some typical scenarios for the use of a Homogeneous SDWAN network are as follows:

- A small branch office connecting to its HQ offices via the Internet. All sensitive traffic to/from this small branch office has to be encrypted, which is usually achieved using IPsec SAs.

- A store in a shopping mall may need to securely connect to its applications in one or more Cloud DCs via the Internet. A common way of achieving this is to establish IPsec SAs to the Cloud DC gateway to carry the sensitive data to/from the store.

As described in [SECURE-EVPN], the granularity of the IPsec SAs for Homogeneous SDWAN can be per site, per subnet, per tenant, or per address. Once the IPsec SA is established for a specific subnet/tenant/site, all traffic to/from the subnets/tenants/site are encrypted.

```
                                       +---+
                        +--------------|RR |------------+
                       /  Untrusted    +-+-+             \
                      /                                   \
                     /                                     \
         +----+  +---------+                             +------+  +----+
         | CN3|--|         P1-----+ -------------+------ P1     |--| CN1|
         +----+  | C-PE1   P2-----+              |       | C-PE2|  +----+
         +----+  |         P3-----+     Wide     +------ P2     |  +----+
         | CN2|--|         |      |     area     +------ P3     |--| CN3|
         +----+  +---------+      |   network    |       +------+  +----+
                              |              |
         +----+  +---------+      | all packets  |       +------+  +----+
         | CN1|--|         P1-----+ encrypted    +------ P1     |--| CN1|
         +----+  | C-PE3   P2-----+     by       |       | C-PE4|  +----+
         +----+  |         P3-----+ IPsec SAs    +------ P2     |  +----+
         | CN2|--|         P4-----+--------------+       |      |--| CN2|
         +----+  +---------+                             +------+  +----+
```

CN: Client Networks, which is same as Tenant Networks used by NVo3

Figure 2: Homogeneous SDWAN

One of the key properties of homogeneous SDWAN is that the SDWAN Local Network Controller (RR)is connected to C-PEs via untrusted public network, therefore, requiring secure connection between RR and C-PEs (TLS, DTLS, etc.).

Homogeneous SDWAN has some similarity to commonly deployed IPsec VPN, albeit the IPsec VPN is usually point-to-point among a small number of nodes and with heavy manual configuration for IPsec between nodes, whereas an SDWAN network can have a large number of edge nodes with an SDWAN controller to manage requiring zero touch provisioning upon powering up.

Existing Private VPNs (e.g. MPLS based) can use homogeneous SDWAN to extend over public network to remote sites to which the VPN operator does not own or lease infrastructural connectivity, as described in [SECURE-EVPN] and [SECURE-L3VPN]

3.3. Scenario #2: CPE based SDWAN over Hybrid WAN Underlay

In this scenario, SDWAN edge nodes (a.k.a. C-PEs) have some WAN ports connected to PEs of Private VPNs over which packets can be forwarded natively without encryption, and some WAN ports connected to the public Internet over which sensitive traffic have to be encrypted (usually by IPsec SA).

In this scenario, the SDWAN edge nodes’ egress WAN ports are all IP/Ethernet based, either egress to PEs of the VPNs or egress to the public Internet. Even if the VPN is a MPLS network, the VPN’s PEs have IP/Ethernet links to the SDWAN edge (C-PEs). Throughout this document, this scenario is also called CPE based SDWAN over Hybrid Networks.

Even though IPsec SA can secure the packets traversing the Internet, it does not offer the premium SLA commonly offered by Private VPNs, especially over long distance. Clients need to have policies to specify criteria for flows only traversing private VPNs or traversing either as long as encrypted when over the Internet. For example, client can have those polices for the flows:

1. A policy or criteria for sending the flows over a private network without encryption (for better performance),
2. A policy or criteria for sending the flows over any networks as long as the packets of the flows are encrypted when traversing untrusted networks, or
3. A policy of not needing encryption at all.

If a flow traversing multiple segments, such as A<->B<->C<->D, has either Policy 2 or 3 above, the flow can traverse different underlays in different network segments, such as over Private network underlay between A<->B without encryption, or over the public internet between B<->C in an IPsec SA.

As shown in the figure below, C-PE-1 has two different types of interfaces (A1 to Internet and A2 & A3 to VPN). The C-PEs' loopback addresses and addresses attached to C-PEs may or may not be visible to the ISPs/NSPs. The addresses for the WAN ports can have addresses allocated by service providers or dynamically assigned (e.g. by DHCP). One WAN port shown in the figure below (e.g. A1, A2, A3 etc.) is a logical representation of potential multiple physical ports on the C-PEs.

```
                                  +---+
                   +--------------|RR |----------+
                  /  Untrusted    +-+-+           \
                 /                                 \
                /                                   \
+----+  +---------+  packets encrypted over     +------+  +----+
| CN3|--|         A1-----+ Untrusted    +------ B1     |--| CN1|
+----+  | C-PE1   A2-\                          | C-PE2|  +----+
+----+  |         A3--+--+              +---+---B2     |  +----+
| CN2|--|         |   |PE+--------------+PE |---B3     |--| CN3|
+----+  +---------+   +--+   trusted    +---+   +------+  +----+
                         |      WAN     |
+----+  +---------+   +--+   packets    +---+   +------+  +----+
| CN1|--|         C1--|PE| go natively  |PE |-- D1     |--| CN1|
+----+  | C-PE3   C2--+--+ without encry+---+   | C-PE4|  +----+
        |         |      +--------------+       |      |
        |         |                             |      |
+----+  |         |      without encrypt over   |      |  +----+
| CN2|--|         C3--+---- Untrusted  --+------D2     |--| CN2|
+----+  +---------+                             +------+  +----+

CN: Client Network
```

Figure 3: Hybrid SDWAN

Some key characteristics of a Hybrid SDWAN overlay network are as follows:

- one C-PE may be connected to different ISPs/NSPs, with some of its WAN ports addresses being assigned by different ISPs/NSPs.

- The WAN ports connected to PEs of trusted private networks (e.g. MPLS VPN) hand off IP/Ethernet packets, just like today's CPE that do not handle MPLS packets and do not participate in the underlay VPN networks' control plane. Traffic can flow natively without encryption when be forwarded out through those WAN ports for better performance.

- The WAN ports connected to untrusted networks, e.g. the Internet, requires sensitive traffic to be encrypted, i.e. encrypted by IPsec SA.

- An SDWAN local Network Controller (RR) is connected to C-PEs via the untrusted public network, therefore, requiring secure connection between RR and C-PEs via TLS, DTLS, etc.

- The SDWAN nodes' [loopback] addresses might not be routable nor visible in the underlay ISP/NSP networks. Routes & services attached to SDWAN edges at the SDWAN overlay layer are in different address spaces than the underlay networks.

- There could be multiple SDWAN devices sharing a common property, such as a geographic location. Some applications over SDWAN may need to traverse specific geographic locations for various reasons, such as to comply with regulatory rules, to utilize specific value added services, or others.

- The underlay path selection between sites can be a local decision. Some policies allow one service from C-PE1 -> C-PE2 -> C-PE3 using one ISP/NSP underlay in the first segment (C-PE1 -> C-PE2) and using a different ISP/NSP in the second segment (C-PE2-> CPE3).

- Services may not be congruent, i.e. the packets from A-> B may traverse one underlay network, and the packets from B -> A may traverse a different underlay.

- Different services, routes, or VLANs attached to SDWAN nodes can be aggregated over one underlay path; same service/routes/VLAN can spread over multiple SDWAN underlays at different times depending on the policies specified for the service. For example, one tenant's packets to HQ need to be encrypted when sent over the Internet or have to be sent over private networks, while the same tenant’s packets to Facebook can be sent over the Internet without encryption.

3.4. Scenario #3: Private VPN PE based SDWAN

This scenario refers to existing VPN (e.g. MPLS based VPN, such as EVPN or IPVPN) adding extra ports facing untrusted public networks allowing PEs to offload some low priority traffic to ports facing public networks when the VPN MPLS paths are congested. Throughout this document, this scenario is also called Internet Offload for Private VPN, or PE based SDWAN.

In this scenario, the packets offloaded to untrusted public network must be encrypted.

PE based SDWAN can be used by VPN service providers to temporarily increase bandwidth between sites when they are not sure if the demand will sustain for long period of time or as a temporary solution before the permanent infrastructure is built or leased.

```
                                   +---+
                           +-------|PE2|
                          /        +---+
                Internet /          ^
                offload /           || VPN
                       /     VPN    v
                      ++--+        ++-+       +---+
                      |PE1| <====> |RR| <===> |PE3|
                      +-+-+        +--+       +-+-+
                        |                       |
                        +--- Public Internet -- +
```

Figure 4: Additional Internet paths added to the VPN

Here are some key properties for PE based SDWAN:

- For MPLS based VPN, PEs continue having MPLS encapsulation handoff to existing paths.

- The BGP RR is connected to PEs in the same way as VPN, i.e. via the trusted network.
- For the added Internet ports, PEs have IP packets handoff, i.e. sending and receiving IP data frames. Internally, PEs can have the option to encapsulate the MPLS payload in IP, as specified by RFC4023.
- The ports facing public internet might get IP addresses assigned by ISPs, which may not be in the same address domain as PEs'.
- Ports facing public internet are not as secure as the ports facing private infrastructure. There could be spoofing, or DDOS attacks to the ports facing public internet. Extra consideration must be given when injecting the new routes learned from public network into VRFs.
- Even though packets are encrypted over public internet, the performance SLA is not guaranteed over public internet. Therefore, clients may have policies only allowing some flows to be offloaded to internet path.

4. BGP Walk Through

4.1. BGP Walk Through for Homogeneous SDWAN

In the figure below, packets destined towards multiple routes attached to the C-PE2 can be carried by one IPsec tunnel. Then one BGP UPDATE can be announced by C-PE2 to its RR.

```
                                  +---+
                        +---------|RR |----------+
                       / Untrusted+---+            \
                      /                             \
                C-PE1/                               \
             +---------+                        +------+
           --+---+---------------------------------->  |-10.1.x.x/16
             |  /      |                        |C-PE2 |- VLAN = 15
             | /       |                      +----->  |
           --|/1.1.1.1 |                      | |      |-12.1.1.x/24
             +---------+                      | +------+
                                              |  2.2.2.2
                                              |
               C-PE3                          |
             +---------+                      |
           --|---+----------------------------+
             |  /      |
             | /       |
           --|/3.3.3.3 |
             +---------+
```

Figure 5: Homogeneous SDWAN

The BGP UPDATE Message from C-PE2 to RR should have the client routes encoded in the MP-NLRI Path Attribute and the IPsec Tunnel associated information encoded in the Tunnel-Encap Path Attributes as described in the [SECURE-EVPN].

Alternatively, two separate BGP UPDATEs can be used to optimize the BGP UPDATE packet size, as described by Section 4 and 8 of [Tunnel-encap]. UPDATE U1 has its Nexthop to the node loopback address and is reclusively resolved to the IPsec tunnel detailed attributes advertised by the UPDATE U2 for the Node Loopback address:

Suppose that:

- a given packet P destined towards the client addresses attached to C-PE2 (e.g. prefix 10.1.x.x/16) can be carried by any IPsec tunnels terminated at C-PE2;
- The path along which P is to be forwarded is determined by BGP UPDATE U1;
- UPDATE U1 does not have a Tunnel Encapsulation attribute;
- UPDATE U1 can include Encapsulation Extended Community and/or Color Extended Community;
- The address of the next hop of UPDATE U1 is router C-PE2;

- The best route to router C-PE2 is a BGP route that was advertised in UPDATE U2;
- UPDATE U2 has a Tunnel Encapsulation attribute to further describe the IPsec detailed attributes.

```
   UPDATE U1:

     - MP-NLRI Path Attribute:
         10.1.x.x/16
         VLAN #15
         12.1.1.x/24
         Nexthop: 2.2.2.2 (C-PE2)
     - Encapsulation Extended Community: Type = IPsec



   UPDATE U2:
     - MP-NLRI Path Attribute:
         2.2.2.2 (C-PE2)
     - Tunnel Encapsulation Path Attributes (as described in the
     [SECURE-EVPN])
```

If different client routes attached to C-PE2 needs to be reached by separate IPsec tunnels, then The Color-Extended-Community [Tunnel-encap] is used to associate routes with the tunnels. See Section 8 of [Tunnel-encap].

If C-PE2 doesn't have the policy on authorized peers for the specific client routes, RR needs to check the client routes policies to propagate the BGP UPDATE messages to the remote authorized edge nodes.

4.2. BGP Walk Through for Hybrid WAN Underlay

In this scenario, some client routes can be forwarded by any tunnels terminated at the edge node and some client routes can be forwarded by some specific tunnels (such as only MPLS VPN).

Color Extended Community (Section 4 & 8 of [Tunnel-Encap]) can be used to represent specific tunnels for the client routes.

For example, in the Figure 5 above, suppose that Route 10.1.x.x/16 can be carried by either MPLS or IPsec, and Route 12.1.1.x/24 can only be carried by MPLS, the following UDPATE messages can be used:

```
UPDATE #1a for Route Route 10.1.x.x/16:

  - MP-NLRI Path Attribute:
      10.1.x.x/16
      Nexthop: 2.2.2.2 (C-PE2)
  - Encapsulation Extended Community: Type = IPsec
  - Encapsulation Extended Community: Type= MPLS-in-GRE
  - Color Extended Community: RED

UPDATE #1b for Route Route 12.1.1.x/24:
  - MP-NLRI Path Attribute:
      12.1.1.x/24
      Nexthop: 2.2.2.2 (C-PE2)
  - Encapsulation Extended Community: Type= MPLS-in-GRE
  - Color Extended Community: YELLOW


UPDATE #2a: for IPsec tunnels terminated at the node:
  - MP-NLRI Path Attribute:
      2.2.2.2 (C-PE2)
  - Tunnel Encapsulation Path Attributes: TYPE=IPsec (as described
  in the [SECURE-EVPN])
  - Color Extended Community: RED



UPDATE #2b: for MPLS-in-GRE terminated at the node:
  - MP-NLRI Path Attribute:
      2.2.2.2 (C-PE2)
  - Tunnel Encapsulation Path Attributes: TYPE=MPLS-in-GRE [Tunnel-
  Encap] Section 3
  - Color Extended Community: YELLOW
```

4.3. BGP Walk Through for Application Flow Based Segmentation

If the applications are assigned with unique IP addresses, the Application Flow based Segmentation described in Section 3.1.2 can be achieved by advertising different BGP UPDATE messages to different nodes. In the Figure below, the following BGP Updates can be advertised to ensure that Payment Application only communicates with the Payment Gateway:

BGP UPDATE #1a from C-PE2 to RR for the P2P topology that is only propagated to Payment GW node:

UPDATE #1a (only to the Payment GW node):

- MP-NLRI Path Attribute:
    - 30.1.1.x/24
    - Nexthop: 2.2.2.2
- Encapsulation extended community: Tunneltype=IPSEC
- Color Extended Community: BLUE

BGP UPDATE #1b from C-PE2 to RR for the routes to be reached by C-PE1 and C-PE2:

- MP-NLRI Path Attribute:
    - 10.1.x.x
    - 12.4.x.x
    - Nexthop:2.2.2.2
- Encapsulation extended community: Tunnel-type=IPSEC
- Color Extended Community: RED

BGP UPDATE #2 describes the IPsec detailed attributes for IPsec tunnels terminated at C-PE2 2.2.2.2.

UPDATE #2a: for all IPsec SAs terminated at the node:

- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)
- Tunnel Encapsulation Path Attributes: TYPE=IPsec (for all IPsec SAs)
- Color Extended Community: RED

UPDATE #2b: for the IPsec SA to the Payment Gateway:

- MP-NLRI Path Attribute:
  2.2.2.2 (C-PE2)

- Tunnel Encapsulation Path Attributes: TYPE=IPsec (for the IPsec SA to Payment GW).
- Color Extended Community: Blue

```
                                  +-------+
                                  |Payment|
                        +-------->|  GW   |<-------+
                       /Hub-spoke +-------+         \
                      /for Payment App               \
               C-PE1 /                                \ C-PE2
             +------/--+                          +----\-+
           --|-----/   |                          |     -|- 30.1.1.x/24
             + --------------------------------------->  |-10.1.x.x/16
             |         |                          |      |-
             |         |                 +------------>  |- 12.1.1.x/24
           --|---------------------------+        |      |
             +---------+                       +------>  |- VLAN=25;
                                              /   +------+  22.1.1.x/24
             +---------+                     /
           --| -----------------------------+
             | C-PE3   |                   /
             |         |                  /
           --| --------------------------+
             +---------+
```

Figure 6: Application Based SDWAN Segmentation

4.4. Client Service Provisioning Model

The provisioning tasks described in Section 4 of RFC8388 are the same for the SDWAN client traffic. When client traffic is multi-homed to two (or more) C-PEs, the Non-Service-Specific parameters need to be provisioned per the Section 4.1.1 of RFC8388.

Since some SDWAN nodes are ephemeral and have small number of IP subnets or VLANs attached to their client ports, it is recommended to have default and simplified Service-specific parameters for each client port, remotely managed by the SDWAN Network Controller via the secure channel (TLS/DTLS) between the controller and the C-PEs.

4.5. Underlay Network Properties Advertisement

Since the deployment of PEs to MPLS VPN are for relatively long term, the common provisioning procedure for PE’s WAN ports is via CLI.

A SDWAN node deployment can be ephemeral and its location can be in remote locations, manual provisioning for its WAN ports is not acceptable. In addition, a SDWAN WAN port’s IP address can be dynamically assigned or using private addresses. Therefore, it is necessary to have a separate control protocol; something like NHRP did for ATM, for a SDWAN node to advertise its directly connected underlay network properties to its peers.

Unlike a PE to MPLS based VPN where its WAN ports are homogeneously facing MPLS private network and all traffic are egressed in MPLS data frames through its WAN ports, the WAN ports of a SDWAN node can be connected to a PE of VPN with Ethernet/IP, MPLS private network directly via MPLS headers, or the public Internet.

For Scenario #1 described in Section 3.2, the WAN ports can face public internet or VPN.

For Scenario #2 described in Section 3.3, WAN ports are either configured as connecting to PEs of VPN where traffic can be sent as IP/Ethernet without encryption, or configured as connecting to public Internet that requires encryption for packets egress out.

For Scenario #3 described in Section 3.4, the WAN ports are either configured as VPN egress ports (hand off MPLS data frames), or as connecting to the public internet that requires MPLS in IP in IPsec encapsulation.

4.6. Why BGP as Control Plane for SDWAN?

For a small sized SDWAN network, traditional hub & spoke model using NHRP or DSVPN/DMVPN with a hub node (or controller) managing SDWAN node WAN ports mapping (e.g. local & public addresses and tunnel identifiers mapping) can work reasonably well. However, for a large SDWAN network, say more than 100 nodes with different types of topologies, the traditional approach becomes very messy, complex and error prone.

Here are some of the compelling reasons of using BGP instead of extending NHRP/DSVPN/DMVPN. (Same as the reasons quoted by LSVR on why using BGP):

- BGP has the built-in capability to constrain the propagation of SDWAN edge node properties to a small number of edge nodes [RFC4684].

- RR already has the capability to apply policies to communications among peers.

- BGP is widely deployed as sole protocol (see RFC 7938)

- Robust and simple implementation

- Wide acceptance - minimal learning

- Reliable transport

- Guaranteed in-order delivery

- Incremental updates

- Incremental updates upon session restart

- No flooding and selective filtering

5. SDWAN Traffic Forwarding Walk Through

BGP based EVPN control plane are still applicable to routes attached to the client ports of SDWAN nodes. Section 5 of RFC8388 describes the BGP EVPN NLRI Usage for various routes of client traffic. The procedures described in the Section 6 of RFC8388 are same for the SDWAN client traffic.

The only additional consideration for SDWAN is to control how traffic egress the SDWAN edge node to various WAN ports.

5.1. SDWAN Network Startup Procedures

A SDWAN network can add or delete SDWAN edge nodes on regular basis depending on user requests.

- For Scenario #1: a SDWAN edge node in a shopping mall or Cloud DC can be added or removed on demand. The Zero Touch Provisioning described in 3.1.2 are required for the node startup.
- For Scenario #2: this can be Data Centers or enterprises upgrading their CPEs to add extra bandwidth via public internet in addition to VPN services that they already purchased. Before the node powers up

or upgraded, there should be links connected to the PEs of a provider VPNs.

- For Scenario #3, the Internet facing WAN ports are added to (or removed from) existing VPN PEs.

5.2. Packet Walk-Through for Scenario #1

Upon power up, a SDWAN node can learn client routes from the Client facing ports, in the same way as EVPN described in RFC8388. Controller facilitates the IPsec SA establishment and rekey management as described in [SECURE-EVPN]. Controller manages how client's routes are associated with individual IPSec SA.

[SECURE-EVPN] describes a solution for SDWAN Scenario #1. It utilizes the BGP RR to facilitate the key and policy exchange among PE devices to create private pair-wise IPsec Security Associations without IKEv2 point-to-point signaling or any other direct peer-to-peer session establishment messages.

When C-PEs do not support MPLS, the approaches described by RFC8365 can be used, with addition of IPsec encrypting the IP packets when sending packets over the Black Interfaces.

5.3. Packet Walk-Through for Scenario #2

In this scenario, C-PEs have some WAN ports connected to the public internet and some WAN ports with direct connect to PEs of trusted VPN. The C-PEs in Scenario #2 have the plain IP/Ethernet data frames egress to the PEs of the VPN, encrypted data frames egress the WAN ports facing the public Internet.

Users specify the policy or criteria on which flows can only egress WAN ports facing the trusted VPN without encryption, which can egress the WAN ports facing the public Internet with encryption, or which can egress WAN ports facing the public Internet without encryption.

The internet facing WAN ports can face potential DDoS attacks, additional anti-DDoS mechanism has to be enabled on those WAN ports and the Control Plane should not learn routes from the Public Network facing WAN ports.

For the Scenario #2, if a client route can be reached by MPLS VPN and IPsec Tunnel via public network, the BGP UPDATE for the client route should indicate all available tunnels in the Tunnel Path Attribute of the BGP NLRI.

```
                                       +---+
                        +--------------|RR |----------+
                       /  Untrusted    +-+-+           \
                      /    Network                      \
                     /                                   \
     +----+  +---------+  packets encrypted over     +--------+  +----+
     | CN3|--|         A1-----+ Untrusted    +------ B1       |--| CN1|
     +----+  | C-PE-a  A2-----+              +-------B2 C-PE-b|  +----+
             |10.1.1.1 |                             |10.1.2.1|
     +----+  |         |   +--+              +---+   |        |  +----+
     | CN2|--|         A3  |PE+--------------+PE |---B3       |--| CN3|
     +----+  +---------+   +--+   trusted    +---+   +--------+  +----+
                              |     VPN      |
                              +--------------+
```

Figure 8: SDWAN Scenario #2

For example, if the CN1 route can be reached by both VPN and Public internet, the CN1’s BGP route UPDATE should include the following:

- MP-NLRI Path Attribute:

  CN1

- Tunnel-Encap Path Attribute:

  Tunnel 1: MPLS-in-GRE encapsulation
    With the MPLS-in-GRE Sub-TLV specified by Tunnel-Encap;

  Tunnel 2: IPsec-GRE encapsulation
    With the IPsec Sub-TLVs specified by the [SECURE-EVPN] and [BGP-EDGE-DISCOVERY]

There could be multiple IPsec SA tunnels terminated at the edge node loopback address or terminated at WAN ports. For the Scenario #2, there can be policies to determine which IPsec SA tunnels that the client route can be carried. When a client route can be carried by multiple IPsec SA tunnels terminated by two different WAN ports, multiple Tunnel Path Attributes with different Tunnel-end-point Sub-TLVs need to be included in the NLRI of the BGP UPDATE for the client route.

5.4. Packet Walk-Through for Scenario #3

The behavior described in [SECURE-L3VPN] applies to this scenario.

[SECURE-L3VPN] describes how to extend the RFC4364 VPN to allow some PEs being connected to other PEs via public networks. In this scenario, the PEs is the SDWAN Edge nodes. [SECURE-L3VPN] introduces the concept of RED Interface & Black Interface on those PEs. RED interfaces face the VPN over which packets can be forwarded natively without encryption. Black Interfaces face public network over which only IPsec-protected packets are forwarded. [SECURE-L3VPN] assumes PEs terminate MPLS packets, and use MPLS over IPsec when sending over the Black Interfaces.

The C-PEs not only have RED interfaces facing clients but also have RED interface facing MPLS backbone, with additional BLACK interfaces facing the untrusted public networks for the WAN side. The C-PEs cannot mix the routes learned from the Black Interfaces with the Routes from RED Interfaces. The routes learned from core-facing RED interfaces are for underlay and cannot be mixed with the routes learned over access-facing RED interfaces that are for overlay. Furthermore, the routes learned over core-facing interfaces (both RED and BLACK) can be shared in the same GLOBAL route table.

There may be some added risks of the packets from the ports facing the Internet. Therefore, special consideration has to be given to the routes from WAN ports facing the Internet. RFC4364 describes using an RD to create different routes for reaching same system. A similar approach can be considered to force packets received from the Internet facing ports to go through special security functions before being sent over to the VPN backbone WAN ports.

6. Manageability Considerations

SDWAN overlay networks utilize the SDWAN controller to facilitate route distribution, central configurations, and others. SDWAN Edge nodes need to advertise the attached routes to their controller (i.e. RR in BGP case).

7. Security Considerations

Having WAN ports facing the public Internet introduces the following security risks:

1) Potential DDoS attack to the C-PEs with ports facing internet.

2) Potential risk of provider VPN network being injected with illegal traffic coming from the public Internet WAN ports on the C-PEs.

8. IANA Considerations

None

9. References

9.1. Normative References

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

[RFC4364] E. rosen, Y. Rekhter, "BGP/MPLS IP Virtual Private networks (VPNs)", Feb 2006.

[RFC7296] C. Kaufman, et al, "Internet Key Exchange Protocol Version 2 (IKEv2)", Oct 2014.

[RFC7432] A. Sajassi, et al, "BGP MPLS-Based Ethernet VPN", Feb 2015.

[RFC8365] A. Sajassi, et al, "A network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", March 2018.

9.2. Informative References

[RFC8192] S. Hares, et al, "Interface to Network Security Functions (I2NSF) Problem Statement and Use Cases", July 2017

[RFC5521] P. Mohapatra, E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", April 2009.

[BGP-SDWAN-Port] L. Dunbar, H. Wang, W. Hao, "BGP Extension for SDWAN Overlay Networks", draft-dunbar-idr-bgp-sdwan-overlay-ext-03, work-in-progress, Nov 2018.

[Net2Cloud-Gap] L. Dunbar, A. Malis, C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work in progress, Oct. 2018.

[SDWAN-EDGE-Discovery] L. Dunbar, S. Hares, R. Raszuk, K. Majumdar, "BGP UPDATE for SDWAN Edge Discovery", draft-dunbar-idr-sdwan-edge-discovery-00, work-in-progress, July 2020.

[VPN-over-Internet] E. Rosen, "Provide Secure Layer L3VPNs over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, work-in-progress, July 2018

[DMVPN] Dynamic Multi-point VPN: https://www.cisco.com/c/en/us/products/security/dynamic-multipoint-vpn-dmvpn/index.html

[DSVPN] Dynamic Smart VPN: http://forum.huawei.com/enterprise/en/thread-390771-1-1.html

[SECURE-EVPN] A. Sajassi, et al, "Secure EVPN", draft-sajassi-bess-secure-evpn-01, Work-in-progress, March 2019.

[SECURE-L3VPN] E. Rosen, R. Bonica, "Secure Layer L3VPN over Public Infrastructure", draft-rosen-bess-secure-l3vpn-00, Work-in-progress, June 2018.

[ITU-T-X1036] ITU-T Recommendation X.1036, "Framework for creation, storage, distribution and enforcement of policies for network security", Nov 2007.

[Net2Cloud-Problem] L. Dunbar and A. Malis, "Seamless Interconnect Underlay to Cloud Overlay Problem Statement", draft-dm-net2cloud-problem-statement-02, June 2018

[Net2Cloud-gap] L. Dunbar, A. Malis, and C. Jacquenet, "Gap Analysis of Interconnecting Underlay with Cloud Overlay", draft-dm-net2cloud-gap-analysis-02, work-in-progress, Aug 2018.

[Tunnel-Encap] E. Rosen, et al "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10, Aug 2018.

10. Acknowledgments

Acknowledgements to Adrian Farrel, Joel Halpern, John Scudder, Darren Dukes, Andy Malis and Donald Eastlake for their review and contributions.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Linda Dunbar
Futurewei
Email: ldunbar@futurewei.com

James Guichard
Futurewei
Email: james.n.guichard@futurewei.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper
Email: jdrake@juniper.net

Basil Najem
Bell Canada
Email: basil.najem@bell.ca

David Carrel
Cisco
Email: carrel@cisco.com

Ayan Banerjee
Cisco
Email: ayabaner@cisco.com

INTERNET-DRAFT
Intended status: Proposed Standard

V. Govindan
M. Mudigonda
A. Sajassi
Cisco Systems
G. Mirsky
ZTE
D. Eastlake
Futurewei Technologies

Expires: May 1, 2021
November 2, 2020

Fault Management for EVPN networks
draft-ietf-bess-evpn-bfd-02

Abstract

This document specifies proactive, in-band network OAM mechanisms to detect loss of continuity and miss-connection faults that affect unicast and multi-destination paths (used by Broadcast, Unknown Unicast, and Multicast traffic) in an Ethernet VPN (EVPN) network. The mechanisms specified in the draft are based on the widely adopted Bidirectional Forwarding Detection (BFD) protocol.

Status of this Memo

Distribution of this document is unlimited. Comments should be sent to the authors or the BESS working group mailing list: bess@ietf.org.

Table of Contents

1. Introduction

[ietf-bess-evpn-oam-req-frmwk] outlines the OAM requirements of Ethernet VPN networks (EVPN [RFC7432]). This document specifies mechanisms for proactive fault detection at the network (overlay) layer of EVPN. The mechanisms proposed in the draft use the widely adopted Bidirectional Forwarding Detection (BFD [RFC5880]) protocol.

EVPN fault detection mechanisms need to consider unicast traffic separately from Broadcast, Unknown Unicast, and Multicast (BUM) traffic since they map to different Forwarding Equivalency Classes (FECs) in EVPN. Hence this document proposes different fault detection mechanisms to suit each type. For unicast traffic and BUM traffic via MP2P tunnels, using BFD [RFC5880], and for BUM traffic via a P2MP tunnel, using BFD Multipoint Active Tails [RFC8563] [mirsky-mpls-p2mp-bfd].

Packet loss and packet delay measurement are out of scope for this document.

1.1 Terminology

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.

The following acronyms are used in this document.

BFD - Bidirectional Forwarding Detection [RFC5880]

BUM - Broadcast, Unknown Unicast, and Multicast

CC - Continuity Check

CV - Connectivity Verification

EVI - EVPN Instance

EVPN - Ethernet VPN [RFC7432]

FEC - Forwarding Equivalency Class

GAL - Generic Associated Channel Label [RFC5586]

LSM - Label Switched Multicast (P2MP)

LSP - Label Switched Path

MP2P - Multi-Point to Point

OAM - Operations, Administration, and Maintenance

P2MP - Point to Multi-Point (LSM)

PE - Provider Edge

VXLAN - Virtual eXtesible Local Area Network (VXLAN) [RFC7348]

2. Scope of this Document

This document specifies BFD based mechanisms for proactive fault detection for EVPN both as specified in [RFC7432] and also for EVPN using VXLAN encapsulation [ietf-vxlan-bfd]. It covers the following:

- Unicast traffic.
- BUM traffic using Multi-point-to-Point (MP2P) tunnels (ingress replication).
- BUM traffic using Point-to-Multipoint (P2MP) tunnels (Label Switched Multicast (LSM)).
- MPLS and VXLAN encapsulation.

This document does not discuss BFD mechanisms for:

- EVPN variants like PBB-EVPN [RFC7623]. It is intended to address this in future versions.
- Integrated Routing and Bridging (IRB) solution based on EVPN [ietf-bess-evpn-inter-subnet-forwarding]. It is intended to address this in future versions.
- EVPN using other encapsulations such as NVGRE or MPLS over GRE [RFC8365].
- BUM traffic using MP2MP tunnels.

This document specifies procedures for BFD asynchronous mode. BFD demand mode is outside the scope of this specification except as it is used in [RFC8563]. The use of the Echo function is outside the scope of this specification.

3. Motivation for Running BFD at the EVPN Network Layer

The choice of running BFD at the network layer of the OAM model for EVPN [ietf-bess-evpn-oam-req-frmwk] was made after considering the following:

- In addition to detecting link failures in the EVPN network, BFD sessions at the network layer can be used to monitor the successful setup of MP2P and P2MP EVPN tunnels transporting Unicast and BUM traffic such as label programming. The scope of reachability detection covers the ingress and the egress EVPN PE nodes and the network connecting them.

- Monitoring a representative set of path(s) or a particular path among multiple paths available between two EVPN PE nodes could be done by exercising entropy mechanisms such as entropy labels, when they are used, or VXLAN source ports. However, paths that cannot be realized by entropy variations cannot be monitored. The fault monitoring requirements outlined by [ietf-bess-evpn-oam-req-frmwk] are addressed by the mechanisms proposed by this draft.

BFD testing between EVPN PE nodes does not guarantee that the EVPN service is functioning. (This can be monitored at the service level, that is CE to CE.) For example, an egress EVPN-PE could understand EVPN labeling received but could switch data to an incorrect interface. However, BFD testing in the EVPN Network Layer does provide additional confidence that data transported using those tunnels will reach the expected egress node. When BFD testing in the EVPN overlay fails, that can be used as an indication of a Loss-of-Connectivity defect in the EVPN underlay that would cause EVPN service failure.

4. Fault Detection for Unicast Traffic

The mechanisms specified in BFD for MPLS LSPs [RFC5884] [RFC7726] are applied to test the handling of unicast EVPN traffic. The discriminators required for de-multiplexing the BFD sessions are advertised through BGP. This is needed for MPLS since the label stack does not contain enough information to identify the sender of the packet.

The usage of MPLS entropy labels or various VXLAN source ports takes care of the requirement to monitor various paths of the multi-path server layer network [RFC6790]. Each unique realizable path between the participating PE routers MAY be monitored separately when such entropy is used. At least one path of multi-path connectivity between two PE routers MUST be tracked with BFD, but in that case the granularity of fault-detection will be coarser.

To support unicast OAM to a PE node, that PE MUST allocate a BFD discriminator to be used for BFD messages to it and MUST advertise this discriminator with BGP using the BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover] in an EVPN MAC/IP Advertisement Route [RFC7432]. If configured to do so, once a PE knows a unicast route and discriminator for another PE, it endeavors to bring UP and maintain a BFD session to that other PE. Once the BFD session is UP, the ends of the BFD session MUST NOT change the local discriminator values of the BFD Control packets they generate, unless they first bring down the session as specified in [RFC5884]. The session is brought down if no route or discriminator is available due to withdrawal.

5. Fault Detection for BUM Traffic

Section 5.1 below discusses fault detection for MP2P tunnels using ingress replication and Section 5.2 discusses fault detection for P2MP tunnels.

5.1 Ingress Replication

Ingress replication uses separate MP2P tunnels for transporting BUM traffic from the ingress PE (head) to a set of one or more egress PEs (tails). The fault detection mechanism specified by this document takes advantage of the fact that the head makes a unique copy for each tail.

Another key aspect to be considered in EVPN is the advertisement of the Inclusive Multicast Ethernet Tag Route [RFC7432]. The BUM traffic flows from a head node to a particular tail only after the head receives the inclusive multicast route. This contains the BUM EVPN MPLS label (downstream allocated) corresponding to the MP2P tunnel for MPLS encapsulation and contains the IP address of the PE originating the inclusive multicast route for use in VXLAN encapsulation. It also contains a BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover].

There MAY exist multiple BFD sessions between a head PE and an individual tail due to (1) the usage of MPLS entropy labels [RFC6790] or VXLAN source ports for an inclusive multicast FEC and (2) due to multiple MP2P tunnels indicated by different tail labels or IP addresses for MPLS or VXLAN. If configured to do so, once a PE knows an inclusive multicast route and discriminator for another PE it endeavors to bring UP and maintain a BFD session to that other PE. Once a BFD session for an MP2P path is UP, the ends of the BFD session MUST NOT change the local discriminator values of the BFD Control packets they generate, unless they first bring down the session as specified in [RFC5884]. The session is brought down if no route or discriminator is available due to withdrawal.

5.2 P2MP Tunnels (Label Switched Multicast)

Fault detection for BUM traffic distributed using a P2MP tunnel uses BFD Multipoint Active Tails in one of the three methods providing head notification. Sections 5.2.2 and 5.2.3 of [RFC8563] describe two of these scenarios ("Head Notification and Tail Solicitation with Multipoint Polling" and "Head Notification with Composite Polling"). [mirsky-mpls-p2mp-bfd] describes the third ("Head Notification without Polling"). All three of these modes assume the existence of unicast paths from the tails to the head. In addition, Head Notification with Composite Polling assumes a head to tail unicast path.

The BUM traffic flows from a head node to the tails after the head receives an inclusive multicast route [RFC7432]. This contains the BUM EVPN MPLS label (upstream allocated) corresponding to the P2MP tunnel for MPLS encapsulation. It also contains a BFD Discriminator Attribute [ietf-bess-mvpn-fast-failover]. The BFD discriminator advertised by a tail in the inclusive multicast route MUST be used in any reverse unicast traffic so the head can determine which tail is responding. If configured to do so, once a PE knows an inclusive multicast route, it brings UP and maintains a BFD session to the tails. The session is brought down if no such route is available due to their withdrawal.

For MPLS encapsulation of the head to tails BFD, Label Switched Multicast is used. For VXLAN encapsulation, BFD is delivered to the tails through underlay multicast using an outer multicast IP address.

6. BFD Packet Encapsulation

The sections below describe the MPLS and VXLAN encapsulations of BFD for EVPN OAM use.

6.1 MPLS Encapsulation

This section describes use of the Generic Associated Channel Label (GAL) for BFD encapsulation in MPLS based EVPN OAM.

6.1.1 MPLS Unicast

As shown in Figure 1, the packet initially contains the following labels: LSP label (transport), the optional entropy label, the EVPN Unicast label, and then the Generic Associated Channel label with the G-ACh type set to TBD1. The G-ACh payload of the packet MUST contain the destination L2 header (in overlay space) followed by the IP header that encapsulates the BFD packet. The MAC address of the inner packet is used to validate the <EVI, MAC> in the receiving node.

- The destination MAC MUST be the dedicated MAC TBD-A (see Section 8) or the MAC address of the destination PE.
- The destination IP address MUST be 127.0.0.1/32 for IPv4 [RFC1812] or ::1/128 for IPv6 [RFC4291].
- The destination IP port MUST be 3784 [RFC5881].
- The source IP port MUST be in the range 49152 through 65535.
- The discriminator values for BFD are obtained through BGP as discussed in Section 4 or are exchanged out-of-band or through some other means outside the scope of this document.

```
       <---------- 4 bytes ---------->
      +-------------------------------+  -----
      |          LSP Label            |      |
      +-------------------------------+      |
      :      entropy label indicator  :      |
      + (optional)                    +  MPLS Label Stack
      :      entropy label            :      |
      +-------------------------------+      |
      |      EVPN Unicast label       |
      +-------------------------------+      |
      | Generic Assoc. Channel Label  |      |
      +-------------------------------+  -----
      |  ACH word, Type TBD1 no TLVs  |
      +-------------------------------+  ---     -------
      |    Destination MAC Address    |   |           |
      +               +---------------+   |           |
      |   TBD-A       |               |   |           |
      +---------------+               +  L2 Header    |
      |       Source MAC Address      |   |           |
      +---------------+---------------+   |           |
      | VLAN Ethertype|     VLAN-ID   |   |           |
      +---------------+---------------+   |           |
      |IP4/6 Ethertype|                   |           |
      +---------------+---------------+  ---          |
      /                               /         G-ACh Payload
      /...      IP4/6 Header       .../                |
      /                               /                |
      +-------------------------------+                |
      |                               |                |
      +           UDP Header          +                |
      |                               |                |
      +-------------------------------+                |
      |                               |                |
      +       BFD Control Packet      +                |
      /                               /                |
      /...                         .../  ---------------
```

Figure 1. MPLS Unicast Encapsulation

6.1.2 MPLS Ingress Replication

The packet initially contains the following labels: LSP label (transport), the optional entropy label, the BUM label, and the split horizon label [RFC7432] (where applicable). The G-ACh type is set to TBD1. The G-ACh payload of the packet is as described in Section 6.1.1.

6.1.3 MPLS LSM (Label Switched Multicast, P2MP)

The encapsulation is the same as in Section 6.1.2 for ingress replication except that the transport label identifies the P2MP tunnel, in effect the set of tail PEs, rather than identifying a single destination PE at the end of an MP2P tunnel.

6.2 VXLAN Encapsulation

This section describes the use of the VXLAN [RFC7348] for BFD encapsulation in VXLAN based EVPN OAM. This specification conforms to [ietf-bfd-vxlan]. [Some or all of this section may be removed as being redundant with [ietf-bfd-vxlan].]

6.2.1 VXLAN Unicast

Figure 2 below shows the unicast VXLAN encapsulation. The outer and inner IP headers have a unicast source IP address of the BFD message source and a destination IP address of the BFD message destination

The destination UDP port MUST be 3784 [RFC5881]. The source port MUST be in the range 49152 through 65535. If the BFD source has multiple IP addresses, entropy MAY be further obtained by using any of those addresses assuming the source is prepared for responses directed to the IP address used.

The Your BFD discriminator is the value distributed for this unicast OAM purpose by the destination using BGP as discussed in Section 4 or is exchanged out-of-band or through some other means outside the scope of this document.

```
       <---------- 4 bytes ---------->
      +-------------------------------+  ---
      |    Destination MAC Address    |    |
      +               +---------------+    |
      |               |               |    |
      +---------------+               +  L2 Header
      |       Source MAC Address      |    |
      +-------------------------------+    |
      |           VLAN Tag            |    |
      +---------------+---------------+    |
      |IP4/6 Ethertype|                    |
      +---------------+---------------+  ---
      /                               /
      /...      IP4/6 Header       .../
      /                               /
      +-------------------------------+
      |                               |
      +          UDP Header           +
      |                               |
      +-------------------------------+
      |                               |
      +         VXLAN Header          +
      |                               |
      +-------------------------------+  ---
      |    Destination MAC Address    |    |
      +               +---------------+    |
      |               |               |    |
      +---------------+               +  L2 Header
      |       Source MAC Address      |    |
      +---------------+---------------+    |
      | IP4 Ethertype |                    |
      +---------------+---------------+  ---
      /                               /
      /...       IP4 Header        .../
      /                               /
      +-------------------------------+
      |                               |
      +          UDP Header           +
      |                               |
      +---------------+---------------+
      |                               |
      +       BFD Control Packet      +
      |                               |
      /...                         .../
```

Figure 2. VXLAN Unicast Encapsulation

6.2.2 VXLAN Ingress Replication

The BFD packet construction is as given in Section 6.2.1 except as follows:

(1) The destination IP address used by the BFD message source is that advertised by the destination PE in its Inclusive Multicast EVPN route for the MP2P tunnel in question; and

(2) The Your BFD discriminator used is the one advertised by the BFD destination using BGP as discussed in Section 5.1 for the MP2P tunnel in question or is exchanged out-of-band or through some other means outside the scope of this document.

6.2.3 VXLAN LSM (Label Switched Multicast, P2MP)

The VXLAN encapsulation for the head-to-tails BFD packets uses the multicast destination IP corresponding to the VXLAN VNI.

The destination port MUST be 3784. For entropy purposes, the source port can vary but MUST be in the range 49152 through 65535 [RFC5881]. If the head PE has multiple IP addresses, entropy MAY be further obtained by using any of those addresses.

The Your BFD discriminator is the value distributed for this multicast OAM purpose by the BFD message using BGP as discussed in Section 5.2 or is exchanged out-of-band or through some other means outside the scope of this document.

7. Scalability Considerations

The mechanisms proposed by this draft could affect the packet load on the network and its elements especially when supporting configurations involving a large number of EVIs. The option of slowing down or speeding up BFD timer values can be used by an administrator or a network management entity to maintain the overhead incurred due to fault monitoring at an acceptable level.

8. IANA Considerations

The following IANA Actions are requested.

8.1 Pseudowire Associated Channel Type

IANA is requested to assign a channel type from the "Pseudowire Associated Channel Types" registry in [RFC4385] as follows.

Value	Description	Reference
TBD1	BFD-EVPN OAM	[this document]

8.2 MAC Address

IANA is requested to assign a multicast MAC address under the IANA OUI [0x01005E900004 suggested] as follows:

Address	Usage	Reference
TBD-A	EVPN OAM	[this document]

9. Security Considerations

Security considerations discussed in [RFC5880], [RFC5883], and [RFC8029] apply.

MPLS security considerations [RFC5920] apply to BFD Control packets encapsulated in a MPLS label stack. When BPD Control packets are routed, the authentication considerations discussed in [RFC5883] should be followed.

VXLAN BFD security considerations in [ietf-vxlan-bfd] apply to BFD packets encapsulate in VXLAN.

Acknowledgements

The authors wish to thank the following for their comments and suggestions:

Mach Chen

Normative References

[ietf-bess-evpn-inter-subnet-forwarding] Sajassi, A., Salam, S., Thoria, S., Rekhter, Y., Drake, J., Yong, L., and L. Dunbar, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-08, work in progress, March 2019.

[ietf-bess-mvpn-fast-failover] Morin, T., Kebler, R., Mirsky, G., "Multicast VPN fast upstream failover", draft-ietf-bess-mvpn-fast-failover-05 (work in progress), February 2019.

[ietf-bfd-vxlan] Pallagatti, S., Paragiri, S., Govindan, V., Mudigonda, M., G. Mirsky, "BFD for VXLAN", draft-ietf-bfd-vxlan (work in progress), October 2020.

[mirsky-mpls-p2mp-bfd] G. Mirsky, S. Mishra, "BFD for Multipoint Networks over Point-to-Multi-Point MPLS LSP", draft-mirsky-mpls-p2mp-bfd (work in progress), October 2020.

[RFC1812] Baker, F., Ed., "Requirements for IP Version 4 Routers", RFC 1812, DOI 10.17487/RFC1812, June 1995, <https://www.rfc-editor.org/info/rfc1812>.

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

[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>.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <http://www.rfc-editor.org/info/rfc4385>.

[RFC5586] Bocci, M., Ed., Vigoureux, M., Ed., and S. Bryant, Ed., "MPLS Generic Associated Channel", RFC 5586, DOI 10.17487/RFC5586, June 2009, <https://www.rfc-editor.org/info/rfc5586>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI

10.17487/RFC5881, June 2010, <https://www.rfc-editor.org/info/rfc5881>.

[RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, DOI 10.17487/RFC5883, June 2010, <https://www.rfc-editor.org/info/rfc5883>.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <https://www.rfc-editor.org/info/rfc5884>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

[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, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[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, <http://www.rfc-editor.org/info/rfc7432>.

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <http://www.rfc-editor.org/info/rfc7623>.

[RFC7726] Govindan, V., Rajaraman, K., Mirsky, G., Akiya, N., and S. Aldrin, "Clarifying Procedures for Establishing BFD Sessions for MPLS Label Switched Paths (LSPs)", RFC 7726, DOI 10.17487/RFC7726, January 2016, <https://www.rfc-editor.org/info/rfc7726>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

[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>.

[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>.

[RFC8563] Katz, D., Ward, D., Pallagatti, S., Ed., and G. Mirsky, Ed., "Bidirectional Forwarding Detection (BFD) Multipoint Active Tails", RFC 8563, DOI 10.17487/RFC8563, April 2019, <https://www.rfc-editor.org/info/rfc8563>.

Informative References

[ietf-bess-evpn-oam-req-frmwk] Salam, S., Sajassi, A., Aldrin, S., J. Drake, and D. Eastlake, "EVPN Operations, Administration and Maintenance Requirements and Framework", draft-ietf-bess-evpn-oam-req-frmwk-00, work in progress, February 2019.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <https://www.rfc-editor.org/info/rfc5920>.

Authors' Addresses

Vengada Prasad Govindan
Cisco Systems

Email: venggovi@cisco.com

Mudigonda Mallik
Cisco Systems

Email: mmudigon@cisco.com

Ali Sajassi
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134, USA

Email: sajassi@cisco.com

Gregory Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Donald Eastlake, 3rd
Futurewei Technologies
2386 Panoramic Circle
Apopka, FL 32703 USA

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Copyright, Disclaimer, and Additional IPR Provisions

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: March 4, 2021

Ali. Sajassi
Mankamana. Mishra
Samir. Thoria
Cisco Systems
Jorge. Rabadan
Nokia
John. Drake
Juniper Networks
August 31, 2020

Per multicast flow Designated Forwarder Election for EVPN
draft-ietf-bess-evpn-per-mcast-flow-df-election-04

Abstract

[RFC7432] describes mechanism to elect designated forwarder (DF) at the granularity of (ESI, EVI) which is per VLAN (or per group of VLANs in case of VLAN bundle or VLAN-aware bundle service). However, the current level of granularity of per-VLAN is not adequate for some applications.[I-D.ietf-bess-evpn-df-election-framework] improves base line DF election by introducing HRW DF election. [I-D.ietf-bess-evpn-igmp-mld-proxy] introduces applicability of EVPN to Multicast flows, routes to sync them and a default DF election. This document is an extension to HRW base draft [I-D.ietf-bess-evpn-df-election-framework] and further enhances HRW algorithm for the Multicast flows to do DF election at the granularity of (ESI, VLAN, Mcast flow).

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

EVPN based All-Active multi-homing is becoming the basic building block for providing redundancy in next generation data center deployments as well as service provider access/aggregation networks. [RFC7432] defines the role of a designated forwarder as the node in the redundancy group that is responsible to forward Broadcast, Unknown unicast, Multicast (BUM) traffic on that Ethernet Segment (CE device or network) in All-Active multi-homing.

The default DF election mechanism allows selecting a DF at the granularity of (ES, VLAN) or (ES, VLAN bundle) for BUM traffic. While [I-D.ietf-bess-evpn-df-election-framework] improve on the default DF election procedure, some service provider residential applications require a finer granularity, where whole multicast flows are delivered on a single VLAN.

```
                               (Multicast sources)
                                        |
                                        |
                                      +---+
                                      |CE4|
                                      +---+
                                        |
                                        |
                                  +-----+-----+
                     +------------|   PE-1    |------------+
                     |            |           |            |
                     |            +-----------+            |
                     |                                     |
                     |                   EVPN              |
                     |                                     |
                     |                                     |
                     | (DF)                           (NDF)|
               +-----------+                        +-----------+
               |  |EVI-1|  |                        |  |EVI-1|  |
               |   PE-2    |------------------------|   PE-3    |
               +-----------+                        +-----------+
                      AC1  \                       / AC2
                            \                     /
                             \      ESI-1        /
                              \                 /
                               \               /
                               +---------------+
                               |    CE2        |
                               +---------------+
                                      |
                                      |
                             (Multiple receivers)
```

Figure 1: Multi-homing Network of EVPN
for IPTV deployments

Consider the above topology, which shows a typical residential deployment scenario, where multiple receivers are behind an all-active multihoming segments. All of the multicast traffic is provisioned on EVI-1. Assume PE-2 get elected as DF. According to [RFC7432], PE-2 will be responsible for forwarding multicast traffic to that Ethernet segment.

o Forcing sole data plane forwarding responsibility on PE-2 is a limitation in the current DF election mechanism. The topology at Figure 1 would always have only one of the PE to be elected as DF irrespective of which current DF election mechanism is in use defined in [RFC7432] or [I-D.ietf-bess-evpn-df-election-framework].

- The problem may also manifest itself in a different way. For example, AC1 happens to use 80% of its available bandwidth to forward unicast data. And now there is need to serve multicast receivers where it would require more than 20% of AC1 bandwidth. In this case, AC1 becomes oversubscribed and multicast traffic drop would be observed even though there is already another link (AC2) present in network which can be used more efficiently load balance the multicast traffic.

In this document, we propose an extension to the HRW base draft to allow DF election at the granularity of (ESI, VLAN, Mcast flow) which would allow multicast flows to be better distributed among redundancy group PEs to share the load.

2. Terminology

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] .

With respect to EVPN, this document follows the terminology that has been defined in [RFC7432] and [RFC4601] for multicast terminology.

3. The DF Election Extended Community

[I-D.ietf-bess-evpn-df-election-framework] defines an extended community, which would be used for PEs in redundancy group to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs in redundancy group about its willingness to support Per multicast flow base DF election capability by signaling a DF election extended community along with Ethernet-Segment Route (Type-4). The current proposal extends the existing extended community defined in [I-D.ietf-bess-evpn-df-election-framework]. This draft defines new a DF type.

- DF type (1 octet) - Encodes the DF Election algorithm values (between 0 and 255) that the advertising PE desires to use for the ES.

  * Type 0: Default DF Election algorithm, or modulus-based algorithms in [RFC7432].

  * Type 1: HRW algorithm defined in [I-D.ietf-bess-evpn-df-election-framework]

  - Type 2: Handshake defines in [I-D.ietf-bess-evpn-fast-df-recovery]
  - Type 3: Time-Synch defined in [I-D.ietf-bess-evpn-fast-df-recovery]
  - Type 4: HRW base per (S,G) multicast flow DF election (explained in this document)
  - Type 5: HRW base per (*,G) multicast flow DF election (explained in this document)
  - Type 6 - 254: Unassigned
  - Type 255: Reserved for Experimental Use.

- The [I-D.ietf-bess-evpn-df-election-framework] describes encoding of capabilities associated to the DF election algorithm using Bitmap field. When these capabilities bits are set along with the DF type-4 and type-5, they need to be interpreted in context of this new DF type-4 and type-5. For example, consider a scenario where all PEs in the same redundancy group (same ES) can support both AC-DF, DF type-4 and DF type-5 and receive such indications from the other PEs in the ES. In this scenario, if a VLAN is not active in a PE, then the DF election procedure on all PEs in the ES should factor that in and exclude that PE in the DF election per multicast flow.

- A PE SHOULD attach the DF election Extended Community to ES route and Extended Community MUST be sent if the ES is locally configured for DF type Per Multicast flow DF election. Only one DF Election Extended community can be sent along with an ES route.

- When a PE receives the ES Routes from all the other PEs for the ES, it checks if all of other PEs have advertised their desire to proceed by Per multicast flow DF election. If all peering PEs have done so, it performs DF election based on Per multicast flow procedure. But if:

  - There is at least one PE which advertised route-4 (AD per ES Route) which does not indicate its capability to perform Per multicast flow DF election. OR
  - There is at least one PE signaling single active in the AD per ES route

it MUST be considered as an indication to support of only Default DF election [RFC7432] and DF election procedure in [RFC7432] MUST be used.

4. HRW base per multicast flow EVPN DF election

This document is an extension of [I-D.ietf-bess-evpn-df-election-framework], so this draft does not repeat the description of HRW algorithm itself.

EVPN PE does the discovery of redundancy groups based on [RFC7432]. If redundancy group consists of N peering EVPN PE nodes, after the discovery all PEs build an unordered list of IP address of all the nodes in the redundancy group. The procedure defined in this draft does not require the list of PEs to be ordered. Address [i] denotes the IP address of the [i]th EVPN PE in redundancy group where (0 < i <= N).

4.1. DF election for IGMP (S,G) membership request

The DF is the PE who has maximum weight for (S, G, V, Es) where

- S - Multicast Source

- G - Multicast Group

- V - VLAN ID.

- Es - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

1. Weight

   * The weight of PE(i) to (S,G,VLAN ID, Es) is calculated by function, weight (S,G,V, Es, Address(i)), where (0 < i <= N), PE(i) is the PE at ordinal i.

   * Weight (S,G,V, Es, Address(i)) = (1103515245. ((1103515245.Address(i) + 12345) XOR D(S,G,V,ESI))+12345) (mod 2^31)

   * In case of tie, the PE whose IP address is numerically least is chosen.

2. Digest

- D(S,G,V, Es) = CRC_32(S,G,V, Es)
- Here D(S,G,V,Es) is the 31-bit digest (CRC_32 and discarding the MSB) of the Source IP, Group IP, Vlan ID and Es. The CRC MUST proceed as if the architecture is in network byte order (big-endian).

4.2. DF election for IGMP (*,G) membership request

The DF is the PE who has maximum weight for (G, V, Es) where

- o G - Multicast Group
- o V - VLAN ID.
- o Es - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

1. Weight
   - The weight of PE(i) to (G,VLAN ID, Es) is calculated by function, weight (G,V, Es, Address(i)), where (0 < i <= N), PE(i) is the PE at ordinal i.
   - Weight (G,V, Es, Address(i)) = (1103515245. ((1103515245.Address(i) + 12345) XOR D(G,V,ESI))+12345) (mod 2^31)
   - In case of tie, the PE whose IP address is numerically least is chosen.
2. Digest
   - D(G,V, Es) = CRC_32(G,V, Es)
   - Here D(G,V,Es) is the 31-bit digest (CRC_32 and discarding the MSB) of the Group IP, Vlan ID and Es. The CRC MUST proceed as if the architecture is in network byte order (big-endian).

4.3. Default DF election procedure

Per multicast DF election procedure would be applicable only when host behind Attachment Circuit (of the Es) start sending IGMP membership requests. Membership requests are synced using procedure defined in [I-D.ietf-bess-evpn-igmp-mld-proxy], and each of the PE in redundancy group can use per flow DF election and create DF state per multicast flow. The HRW DF election "Type 1" procedure defined in [I-D.ietf-bess-evpn-df-election-framework] MUST be used for the Es DF election and SHOULD be performed on Es even before learning multicast membership request state. This default election procedure MUST be used at port level but will be overwritten by Per flow DF election as and when new membership request state are learnt.

5. Procedure to use per multicast flow DF election algorithm

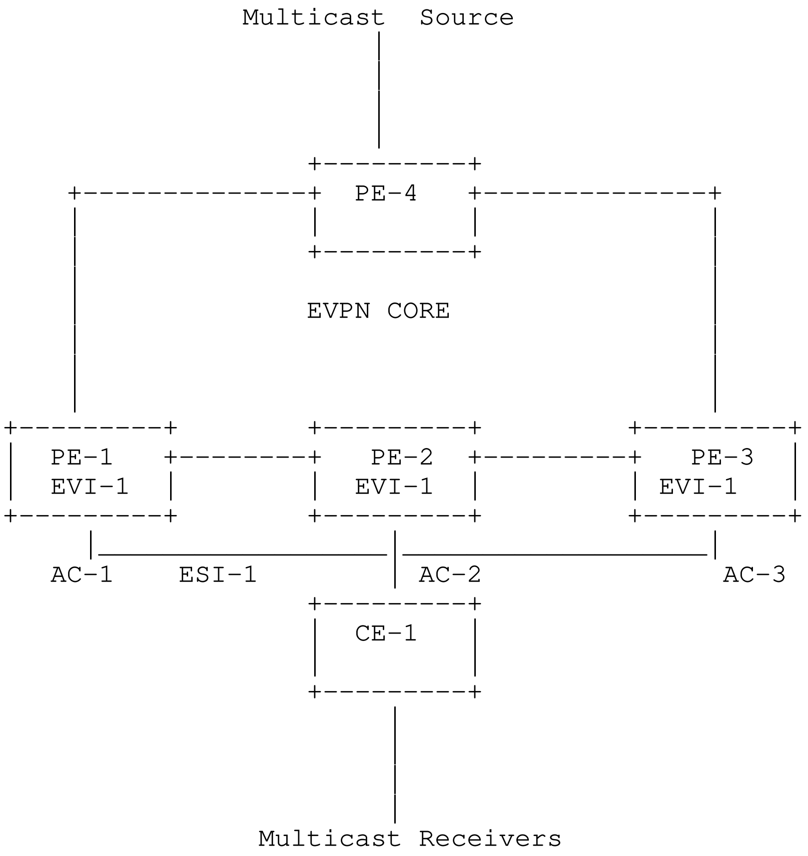

Figure-2 : Multihomed network

Figure-2 shows multihomed network. Where EVPN PE-1, PE-2, PE-3 are multihomed to CE-1. Multiple multicast receivers are behind all active multihoming segment.

1. PEs connected to the same Ethernet segment can automatically discover each other through exchange of the Ethernet Segment

Route. This draft does not change any of this procedure, it still uses the procedure defined in [RFC7432].

2. Each of the PEs in redundancy group advertise Ethernet segment route with extended community indicating their ability to participate in per multicast flow DF election procedure. Since Per multicast flow would not be applicable unless PE learns about membership request from receiver, there is a need to have the default DF election among PEs in redundancy group for BUM traffic. Until multicast membership state are learnt, we use the the DF election procedure in Section 4.3, namely HRW per (v,Es) as defined in [I-D.ietf-bess-evpn-df-election-framework] .

3. When a receiver starts sending membership requests for (s1,g1), where s1 is multicast source address and g1 is multicast group address, CE-1 could hash membership request (IGMP join) to any of the PEs in redundancy group. Let’s consider it is hashed to PE-2. [I-D.ietf-bess-evpn-igmp-mld-proxy] defines a procedure to sync IGMP join state among redundancy group of PEs. Now each of the PE would have information about membership request (s1,g1) and each of them run DF election procedure Section 4.1 to elect DF among participating PEs in redundancy group. Consider PE-2 gets elected as DF for multicast flow (s1,g1).

    1. PE-1 forwarding state would be nDF for flow (s1,g1) and DF for rest other BUM traffic.

    2. PE-2 forwarding state would be DF for flow (s1,g1) and nDF for rest other BUM traffic.

    3. PE-3 forwarding state would be nDF for flow (s1,g1) and rest other BUM traffic.

4. As and when new multicast membership request comes, same procedure as above would continue.

5. If Section 3 has DF type 4, For membership request (S,G) it MUST use Section 4.1 to elect DF among participating PEs. And membership request (*,G) MUST use Section 4.2 to elect DF among participating PEs.

6. Triggers for DF re-election

There are multiple triggers which can cause DF re-election. Some of the triggers could be

1. Local ES going down due to physical failure or configuration change triggers DF re-election at peering PE.

2. Detection of new PE through ES route.

3. AC going up / down

4. ESI change

5. Remote PE removed / Down

6. Local configuration change of DF election Type and peering PE consensus on new DF Type

This document does not provide any new mechanism to handle DF re-election procedure. It uses the existing mechanism defined in [RFC7432]. Whenever either of the triggers occur, a DF re-election would be done. and all of the flows would be redistributed among existing PEs in redundancy group for ES.

7. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

8. IANA Considerations

Allocation of DF type in DF extended community for EVPN.

9. Acknowledgement

Authors would like to acknowledge helpful comments and contributions of Luc Andre Burdet.

10. Normative References

[HRW1999] IEEE, "Using name-based mappings to increase hit rates", IEEE HRW, February 1998.

[I-D.ietf-bess-evpn-df-election-framework] Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for EVPN Designated Forwarder Election Extensibility", draft-ietf-bess-evpn-df-election-framework-03 (work in progress), May 2018.

[I-D.ietf-bess-evpn-fast-df-recovery] Sajassi, A., Badoni, G., Rao, D., Brissette, P., Drake, J., and J. Rabadan, "Fast Recovery for EVPN DF Election", draft-ietf-bess-evpn-fast-df-recovery-00 (work in progress), June 2018.

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-00 (work in progress), March 2017.

[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>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <https://www.rfc-editor.org/info/rfc4601>.

[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>.

Authors' Addresses

Ali Sajassi
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sajassi@cisco.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: mankamis@cisco.com

Samir Thoria
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sthoria@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
UNITED STATES

Email: jorge.rabadan@nokia.com

John Drake
Juniper Networks

Email: jdrake@juniper.net

BESS Working Group A. Sajassi, Ed.
Internet-Draft P. Brissette
Intended status: Standards Track Cisco Systems
Expires: March 12, 2021 J. Uttaro
AT&T
J. Drake
Juniper Networks
S. Boutros
Ciena
J. Rabadan
Nokia
September 8, 2020

EVPN VPWS Flexible Cross-Connect Service
draft-ietf-bess-evpn-vpws-fxc-02

Abstract

This document describes a new EVPN VPWS service type specifically for multiplexing multiple attachment circuits across different Ethernet Segments and physical interfaces into a single EVPN VPWS service tunnel and still providing Single-Active and All-Active multi-homing. This new service is referred to as flexible cross-connect service. After a description of the rationale for this new service type, the solution to deliver such service is detailed.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC8214] describes a solution to deliver P2P services using BGP constructs defined in [RFC7432]. It delivers this P2P service between a pair of Attachment Circuits (ACs), where an AC can designate on a PE, a port, a VLAN on a port, or a group of VLANs on a port. It also leverages multi-homing and fast convergence capabilities of [RFC7432] in delivering these VPWS services. Multi-homing capabilities include the support of single-active and all-active redundancy mode and fast convergence is provided using "mass withdraw" message in control- plane and fast protection switching using prefix independent convergence in data-plane upon node or link failure [I-D.ietf-rtgwg-bgp-pic]. Furthermore, the use of EVPN BGP constructs eliminates the need for multi-segment PW auto-discovery and signaling if the VPWS service need to span across multiple ASes.

Some service providers have very large number of ACs (in millions) that need to be back hauled across their MPLS/IP network. These ACs may or may not require tag manipulation (e.g., VLAN translation). These service providers want to multiplex a large number of ACs across several physical interfaces spread across one or more PEs (e.g., several Ethernet Segments) onto a single VPWS service tunnel in order to a) reduce number of EVPN service labels associated with EVPN-VPWS service tunnels and thus the associated OAM monitoring, and b) reduce EVPN BGP signaling (e.g., not to signal each AC as it is the case in [RFC8214]).

These service provider want the above functionality without scarifying any of the capabilities of [RFC8214] including single-active and all-active multi-homing, and fast convergence.

This document presents a solution based on extensions to [RFC8214] to meet the above requirements.

1.1. Terminology

MAC: Media Access Control

MPLS: Multi Protocol Label Switching

OAM: Operations, Administration and Maintenance

PE: Provider Edge device

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

EVPL: Ethernet Virtual Private Line

EPL: Ethernet Private Line

ES: Ethernet Segment

VPWS: Virtual private wire service

EVI: EVPN Instance

RT: Route Target

VPWS Service Tunnel: It is represented by a pair of EVPN service labels associated with a pair of endpoints. Each label is downstream assigned and advertised by the disposition PE through an Ethernet A-D per-EVI route. The downstream label identifies the endpoint on the disposition PE. A VPWS service tunnel can be associated with many VPWS service identifiers where each identifier is a normalized VID.

Single-Active Redundancy Mode: When a device or a network is multi-homed to two or more PEs and when only a single PE in such redundancy group can forward traffic to/from the multi-homed device or network for a given VLAN, then such multi-homing or redundancy is referred to as "Single-Active".

All-Active Redundancy Mode: When a device is multi-homed to two or more PEs and when all PEs in such redundancy group can forward traffic to/from the multi-homed device for a given VLAN, then such multi-homing or redundancy is referred to as "All-Active".

1.2. 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].

2. Requirements

Two of the main motivations for service providers seeking a new solution are: 1) to reduce number of VPWS service tunnels by multiplexing large number of ACs across different physical interfaces instead of having one VPWS service tunnel per AC, and 2) to reduce the signaling of ACs as much as possible. Besides these two requirements, they also want multi-homing and fast convergence capabilities of [RFC8214].

In [RFC8214], a PE signals an AC indirectly by first associating that AC to a VPWS service tunnel (e.g., a VPWS service instance) and then signaling the VPWS service tunnel via a Ethernet A-D per EVI route with Ethernet Tag field set to a 24-bit VPWS service instance identifier (which is unique within the EVI) and ESI field set to a 10-octet identifier of the Ethernet Segment corresponding to that AC.

Therefore, a PE device that receives such EVPN routes, can associate the VPWS service tunnel to the remote Ethernet Segment, and when the remote ES fails and the PE receives the "mass withdraw" message associated with the failed ES per [RFC7432], it can update its BGP path list for that VPWS service tunnel quickly and achieve fast convergence for multi-homing scenarios. Even if fast convergence were not needed, there would still be a need for signaling each AC failure (via its corresponding VPWS service tunnel) associated with the failed ES, so that the BGP path list for each of them gets updated accordingly and the packets are sent to backup PE (in case of single- active multi-homing) or to other PEs in the redundancy group (in case of all-active multi-homing). In absence of updating the BGP path list, the traffic for that VPWS service tunnel will be black-holed.

When a single VPWS service tunnel multiplexes many ACs across number of Ethernet Segments (number of physical interfaces) and the ACs are not signaled via EVPN BGP to remote PE devices, then the remote PE devices neither know the association of the received Ethernet Segment to these ACs (and in turn to their local ACs) nor they know the association of the VPWS service tunnel (e.g., EVPN service label) to the far-end ACs - i.e, the remote PEs only know the association of their local ACs to the VPWS service tunnel but not the far-end ACs. Thus upon a connectivity failure to the ES, they don’t know how to redirect traffic via another multi-homing PE to that ES. In other words, even if an ES failure is signaled via EVPN to the remote PE devices, they don’t know what to do with such message because they don’t know the association among the remote ES, the remote ACs, and the VPWS service tunnel.

In order to address this issue when multiplexing large number of ACs onto a single VPWS service tunnel, two mechanisms are devised: one to support VPWS services between two single-homed endpoints and another one to support VPWS services where one of the endpoints is multi-homed. An endpoint can be an AC, MAC-VRF, IP-VRF, global table, or etc.

For single-homed endpoints, it is OK not to signal each AC in BGP because upon connection failure to the ES, there is no alternative path to that endpoint. However, the ramification for not signaling an AC failure is that the traffic destined to the failed AC, is sent over MPLS/IP core and then gets discarded at the destination PE - i.e., it can waste network resources. However, when there is a connection failure, the application layer will eventually stop sending traffic making transient this waste of network resources. Section 3.2 describes a solution for such single-homing VPWS service.

For VPWS services where one of the endpoints is multi-homed, there are two options:

1) to signal each AC via BGP so that the path list can be updated upon a failure that impacts those ACs. This solution is described in Section 3.3 and it is called VLAN-signaled flexible cross-connect service.

2) to bundle several ACs on an ES together per destination end-point (e.g., ES, MAC-VRF, etc.) and associated such bundle to a single VPWS service tunnel. This is similar to VLAN-bundle service interface described in [RFC8214]. This solution is described in Section 3.2.1.

3. Solution

This section describes a solution for providing a new VPWS service between two PE devices where a large number of ACs (e.g., VLANs) that span across many Ethernet Segments (i.e., physical interfaces) on each PE are multiplex onto a single P2P EVPN service tunnel. Since multiplexing is done across several physical interfaces, there can be overlapping VLAN IDs across these interfaces; therefore, in such scenarios, the VLAN IDs (VIDs) MUST be translated into unique VIDs to avoid collision. Furthermore, if the number of VLANs that are getting multiplex onto a single VPWS service tunnel exceed 4095, then a single tag to double tag translation MUST be performed. This translation of VIDs into unique VIDs (either single or double) is referred to as "VID normalization".

When single normalized VID is used, the lower 12-bit of Ethernet tag field in EVPN routes is set to that VID and when double normalized VID is used, the lower 12-bit of Ethernet tag field is set to inner VID and the higher 12-bit is set to the outer VID. As in [RFC8214], 12-bit and 24-bit VPWS service instance identifiers representing normalised VIDs MUST be right-aligned.

Since there is only a single EVPN VPWS service tunnel associated with many normalized VIDs (either single or double) across multiple physical interfaces, MPLS lookup at the disposition PE is no longer sufficient to forward the packet to the right egress endpoint/interface. Therefore, in addition to an EVPN label lookup corresponding to the VPWS service tunnel, a VID lookup (either single or double) is also required. On the disposition PE, one can think of the lookup of EVPN label results in identification of a VID-VRF, and the lookup of normalized VID(s) in that table, results in identification of egress endpoint/interface. The tag manipulation (translation from normalized VID(s) to local VID) can be performed either as part of the VID table lookup or at the egress interface itself.

Since VID lookup (single or double) needs to be performed at the disposition PE, then VID normalization MUST be performed prior to the

MPLS encapsulation on the ingress PE. This requires that both imposition and disposition PE devices be capable of VLAN tag manipulation, such as re-write (single or double), addition, deletion (single or double) at their endpoints (e.g., their ES's, MAC-VRFs, IP-VRFs, etc.).

3.1. VPWS Service Identifiers

In [RFC8214], a unique value in the context of each PE's EVI is signaled. The 32-bit Ethernet Tag ID field MUST be set to this VPWS service instance identifier value.
For FXC, Ethernet Tag ID field value may represent:

- VLAN-Bundle : a unique value for a group of VLANs ;

- VLAN-Aware Bundle : a unique value for individual VLANs, and may be considered same as the normalised VID

Both the VPWS service instance identifier and normalised VID are carried in the Ethernet Tag ID field of the Ethernet A-D per EVI route. For FXC, in the case of a 12-bit ID the VPWS service instance identifier is the same as the single-tag normalised VID and will be the same on both PEs. Similarly in the case of a 24-bit ID, the VPWS service instance identifier is the same as the double-tag normalised VID.

3.2. Flexible Xconnect

In this mode of operation, many ACs across several Ethernet Segments are multiplex into a single EVPN VPWS service tunnel represented by a single VPWS service ID. This is the default mode of operation for FXC and the participating PEs do not need to signal the VLANs (normalized VIDs) in EVPN BGP.

With respect to the data-plane aspects of the solution, both imposition and disposition PEs are aware of the VLANs as the imposition PE performs VID normalization and the disposition PE does VID lookup and translation. In this solution, there is only a single P2P EVPN VPWS service tunnel between a pair of PEs for a set of ACs.

As discussed previously, since the EVPN VPWS service tunnel is used to multiplex ACs across different ES's (e.g., physical interfaces), the EVPN label alone is not sufficient for proper forwarding of the received packets (over MPLS/IP network) to egress interfaces. Therefore, normalized VID lookup is required in the disposition direction to forward packets to their proper egress end-points - i.e., the EVPN label lookup identifies a VID-VRF and subsequently, the normalized VID lookup in that table, identifies the egress interface.

This mode of operation is only suitable for single-homing because in multi-homing the association between EVPN VPWS service tunnel and remote AC changes during the failure and therefore the VLANs (normalized VIDs) need to be signaled.

In this solution, on each PE, the single-homing ACs represented by their normalized VIDs are associated with a single EVPN VPWS service tunnel (in a given EVI). The EVPN route that gets generated is an Ethernet A-D per EVI route with ESI=0, Ethernet Tag field set to VPWS service instance ID, MPLS label field set to dynamically generated EVPN service label representing the EVPN VPWS service tunnel. This route is sent with an RT representing the EVI. This RT can be auto-generated from the EVI per section 5.1.2.1 of [RFC8365]. Furthermore, this route is sent with the EVPN Layer-2 Extended Community defined in section 3.1 of [RFC8214] with two new flags (defined in Section 4) that indicate: 1) this VPWS service tunnel is for default Flexible Cross-Connect, and 2) normalized VID type (single versus double). The receiving PE uses these new flags for consistency check and MAY generate an alarm if it detects inconsistency but doesn’t bring down the VPWS service.

It should be noted that in this mode of operation, a single Ethernet A-D per EVI route is sent upon configuration of the first AC (ie, normalized VID). Later, when additional ACs are configured and associated with this EVPN VPWS service tunnel, the PE does not advertise any additional EVPN BGP routes. The PE only associates locally these ACs with the already created VPWS service tunnel.

3.2.1. Default mode FXC with Multi-homing

The default FXC mode can be used for multi-homing. In this mode, a group of normalized VIDs (ACs) on a single Ethernet segment that are destined to a single endpoint are multiplexed into a single EVPN VPWS service tunnel represented by a single VPWS service ID. When the default FXC mode is used for multi-homing, instead of a single EVPN VPWS service tunnel, there can be many service tunnels per pair of PEs - i.e, there is one tunnel per group of VIDs per pair of PEs and there can be many groups between a pair of PEs, thus resulting in many EVPN service tunnels.

3.3. VLAN-Signaled Flexible Xconnect

In this mode of operation, just as the default FXC mode in Section 3.2, many normalized VIDs (ACs) across several different ES’s/interfaces are multiplexed into a single EVPN VPWS service tunnel; however, this single tunnel is represented by many VPWS service IDs (one per normalized VID) and these normalized VIDs are signaled using EVPN BGP.

In this solution, on each PE, the multi-homing ACs represented by their normalized VIDs are configured with a single EVI. There is no need to configure VPWS service instance ID in here as it is the same as the normalized VID. For each normalized VID on each ES, the PE generates an Ethernet A-D per EVI route where ESI field represents the ES ID, the Ethernet Tag field is set to the normalized VID, MPLS label field is set to dynamically generated EVPN label representing the P2P EVPN service tunnel and it is the same label for all the ACs that are multiplexed into a single EVPN VPWS service tunnel. This route is sent with an RT representing the EVI. As before, this RT can be auto-generated from the EVI per section 5.1.2.1 of [RFC8365]. Furthermore, this route is sent with the EVPN Layer-2 Extended Community defined in section 3.1 of [RFC8214] with two new flags (defined in Section 4) that indicate: 1) this VPWS service tunnel is for VLAN-signaled Flexible Cross-Connect, and 2) normalized VID type (single versus double). The receiving PE uses these new flags for consistency check and MAY generate an alarm if it detects inconsistency but doesn’t bring down the VPWS service.

It should be noted that in this mode of operation, the PE sends a single Ethernet A-D per EVI route for each AC that is configured - i.e., each normalized VID that is configured per ES results in generation of an EVPN Ethernet A-D per EVI.

This mode of operation provides automatic cross checking of normalized VIDs used for EVPL services because these VIDs are signaled in EVPN BGP. For example, if the same normalized VID is configured on three PE devices (instead of two) for the same EVI, then when a PE receives the second Ethernet A-D per EVI route, it generates an error message unless the two Ethernet A-D per EVI routes include the same ESI. Such cross-checking is not feasible in default FXC mode because the normalized VIDs are not signaled.

3.3.1. Local Switching

When cross-connection is between two ACs belonging to two multi-homed Ethernet Segments on the same set of multi-homing PEs, then forwarding between the two ACs MUST be performed locally during normal operation (e.g., in absence of a local link failure) - i.e., the traffic between the two ACs MUST be locally switched within the PE.

In terms of control plane processing, this means that when the receiving PE receives an Ethernet A-D per-EVI route whose ESI is a local ESI, the PE does not alter its forwarding state based on the received route. This ensures that the local switching takes precedence over forwarding via MPLS/IP network. This scheme of locally switched preference is consistent with baseline EVPN [RFC7432] where it describes the locally switched preference for MAC/IP routes.

In such scenarios, the Ethernet A-D per EVI route should be advertised with the MPLS label either associated with the destination Attachment Circuit or with the destination Ethernet Segment in order to avoid any ambiguity in forwarding. In other words, the MPLS label cannot represent the same VID-VRF used in Section 3.3 because the same normalized VID can be reachable via two Ethernet Segments. In case of using MPLS label per destination AC, then this same solution can be used for VLAN-based VPWS or VLAN-bundle VPWS services per [RFC8214].

3.4. Service Instantiation

The V field defined in Section 4 is OPTIONAL. However, when transmitted, its value could be flagging an error condition which may result in an operational issue. Notification to operator of an error is not sufficient, the VPWS service tunnel must not be established.

If both PEs of a VPWS tunnel are signaling a matching Normalised VID in control plane, yet one is operating in single tag and the other in double tag mode, the signaling of V-bit allows for detecting and preventing this tunnel instantiation.

If single VID normalisation is signaled in the Ethernet Tag ID field (12-bits) yet dataplane is operating based double tags, the VID normalisation applies only to outer tag. If double VID normalisation is signaled in the Ethernet Tag ID field (24-bits), VID normalisation applies to both inner and outer tags.

4. BGP Extensions

This draft uses the EVPN Layer-2 attribute extended community defined in [RFC8214] with two additional flags added to this EC as described below. This EC is sent with Ethernet A-D per EVI route per Section 3, and SHALL be sent for Single-Active and MAY be sent for All-Active redundancy mode .

```
        +-------------------------------------------+
        | Type (0x06) / Sub-type (0x04) (2 octets)  |
        +-------------------------------------------+
        | Control Flags (2 octets)                  |
        +-------------------------------------------+
        | L2 MTU (2 octets)                         |
        +-------------------------------------------+
        | Reserved (2 octets)                       |
        +-------------------------------------------+

                             1 1 1 1 1 1
         0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        | MBZ           | V | M |-|C|P|B|    (MBZ = MUST Be Zero)
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following bits in the Control Flags are defined; the remaining bits MUST be set to zero when sending and MUST be ignored when receiving this community.

Name	Meaning
B,P,C	per definition in [RFC8214]
-	reserved for Flow-label (RFC7432-bis)
M	00 mode of operation as defined in [RFC8214] 01 VLAN-Signaled FXC 10 Default FXC
V	00 operating per [RFC8214] 01 single-VID normalization 10 double-VID normalization

The M and V fields are OPTIONAL. The M field is ignored at reception for forwarding purposes and is used for error notifications.

5. Failure Scenarios

Two examples will be used as an example to analyze the failure scenarios.

The first scenario is depicted in Figure 1 and shows the VLAN-signaled FXC mode with Multi-Homing. In this example:

- CE1 is connected to PE1 and PE2 via (port,vid)=(p1,1) and (p3,3) respectively. CE1's VIDs are normalized to value 1 on both PEs, and CE1 is Xconnected to CE3's VID 1 at the remote end.

- CE2 is connected to PE1 and PE2 via ports p2 and p4 respectively:

  * (p2,1) and (p4,3) identify the ACs that are used to Xconnect CE2 to CE4's VID 2, and are normalized to value 2.

  * (p2,2) and (p4,4) identify the ACs that are used to Xconnect CE2 to CE5's VID 3, and are normalized to value 3.

In this scenario, PE1 and PE2 advertise an Ethernet A-D per EVI route per normalized VID (values 1, 2 and 3), however only two VPWS Service Tunnels are needed: VPWS Service Tunnel 1 (sv.T1) between PE1's FXC service and PE3's FXC, and VPWS Service Tunnel 2 (sv.T2) between PE2's FXC and PE3's FXC.

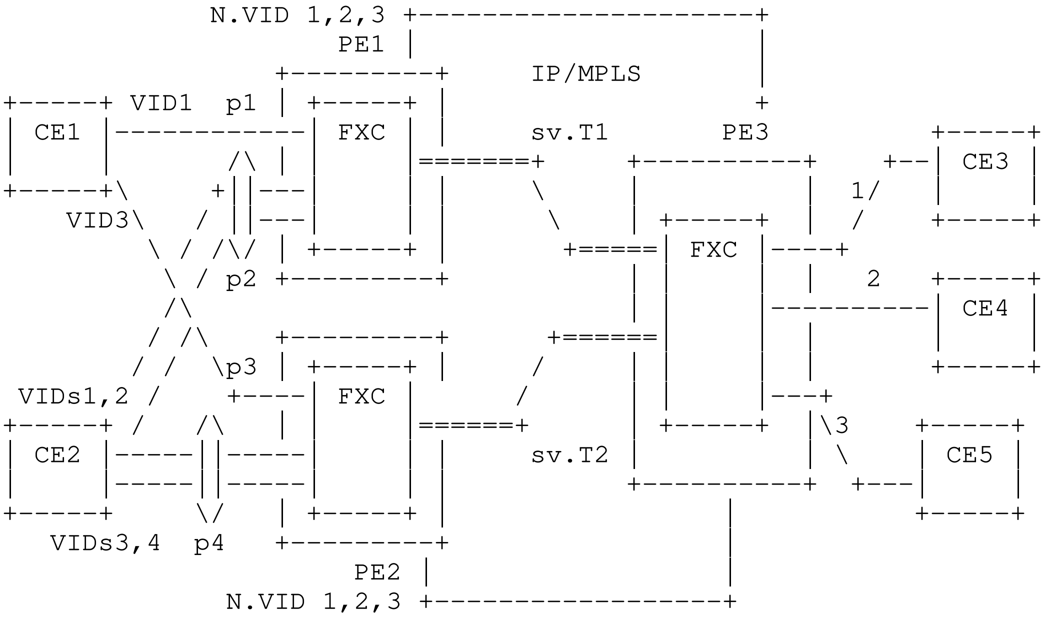

Figure 1: VLAN-Signaled Flexible Xconnect

The second scenario is a default Flexible Xconnect with Multi- Homing solution and it is depicted in Figure 2. In this case, the same VID Normalization as in the previous example is performed, however there is not an individual Ethernet A-D per EVI route per normalized VID, but per bundle of ACs on an ES. That is, PE1 will advertise two

Ethernet A-D per EVI routes: the first one will identify the ACs on p1’s ES and the second one will identify the AC2 in p2’s ES. Similarly, PE2 will advertise two Ethernet A-D per EVI routes.

```
                     N.VID 1,2,3 +---------------------+
                             PE1 |                     |
                         +---------+     IP/MPLS       |
       +-----+ VID1   p1 | +-----+ | sv.T1             +
       | CE1 |-------------| FXC |======+            PE3         +-----+
       |     |        /\ | |     | |    \     +----------+    +--| CE3 |
       +-----+\      +||---|     | sv.T2 \    |          |  1/   |     |
           VID3\    / ||---|     |=====+  \   | +-----+  |  /    +-----+
                \  // \/ | +-----+ |    \  +====| FXC |----+
                 \ /  p2 +---------+     +======|     |  |   2   +-----+
                 /\                           | |     |----------| CE4 |
                / /\    +---------+       +=====|     |  |       |     |
               / /  \p3 | +-----+ sv.T3  / +====|     |  |       +-----+
        VIDs1,2 /    +----| FXC |=======+ /   | |     |---+
       +-----+ /   /\   | |     | |      /    | +-----+  |\3    +-----+
       | CE2 |-----||---| |     | sv.T4 /     |          | \    | CE5 |
       |     |-----||---| |     |======+      +----------+  +---|     |
       +--VIDs3,4  \/   | +-----+ |                  |          +-----+
                   p4   +---------+                  |
                             PE2 |                   |
                     N.VID 1,2,3 +-------------------+
```

Figure 2: Default Flexible Xconnect

5.1. EVPN VPWS Service Failure

The failure detection of an EVPN VPWS service can be performed via OAM mechanisms such as VCCV-BFD and upon such failure detection, the switch over procedure to the backup S-PE is the same as the one described above.

5.2. Attachment Circuit Failure

In case of AC Failure, the VLAN-Signaled and default FXC modes behave in a different way:

o VLAN-signaled FXC (Figure 1): a VLAN or AC failure, e.g. VID1 on CE2, triggers the withdrawal of the Ethernet A-D per EVI route for the corresponding Normalized VID, that is, Ethernet-Tag 2. When PE3 receives the route withdrawal, it will remove PE1 from its path-list for traffic coming from CE4.

- o Default FXC (Figure 2): a VLAN or AC failure is not signaled in the default mode, therefore in case of an AC failure, e.g. VID1 on CE2, nothing prevents PE3 from sending CE4's traffic to PE1, creating a black-hole. Application layer OAM may be used if per-VLAN fault propagation is required in this case.

5.3. PE Port Failure

In case of PE port Failure, the failure will be signaled and the other PE will take over in both cases:

- o VLAN-signaled FXC (Figure 1): a port failure, e.g. p2, triggers the withdrawal of the Ethernet A-D per EVI routes for Normalized VIDs 2 and 3, as well as the withdrawal of the Ethernet A-D per ES route for p2's ES. Upon receiving the fault notification, PE3 will withdraw PE1 from its path-list for the traffic coming from CE4 and CE5.

- o Default FXC (Figure 2): a port failure, e.g. p2, is signaled by route for sv.T2 will also be withdrawn. Upon receiving the fault notification, PE3 will remove PE1 from its path-list for traffic coming from CE4 and CE5.

5.4. PE Node Failure

In the case of PE node failure, the operation is similar to the steps described above, albeit that EVPN route withdrawals are performed by the Route Reflector instead of the PE.

6. Security Considerations

Since this document describes a muxing capability which leverages EVPN-VPWS signaling, no additional functionality beyond the muxing service is added and thus no additional security considerations are needed beyond what is already specified in [RFC8214].

7. IANA Considerations

This document requests allocation of bits 4-7 in the "EVPN Layer 2 Attributes Control Flags" registry with names M and V:

M	Signaling mode of operation (2 bits)
V	VLAN-ID normalisation (2 bits)

8. References

8.1. 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>.

[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>.

[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>.

8.2. Informative References

[I-D.ietf-rtgwg-bgp-pic] Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", draft-ietf-rtgwg-bgp-pic-11 (work in progress), February 2020.

[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>.

Appendix A. Contributors

In addition to the authors listed on the front page, the following co-authors have also contributed substantially to this document:

Wen Lin
Juniper Networks

EMail: wlin@juniper.net

Luc Andre Burdet
Cisco

EMail: lburdet@cisco.com

Authors' Addresses

Ali Sajassi (editor)
Cisco Systems

Email: sajassi@cisco.com

Patrice Brissette
Cisco Systems

Email: pbrisset@cisco.com

James Uttaro
AT&T

Email: uttaro@att.com

John Drake
Juniper Networks

Email: jdrake@juniper.net

Sami Boutros
Ciena

Email: sboutros@ciena.com

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com

BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2021

J. Rabadan, Ed.
S. Sathappan
K. Nagaraj
Nokia
M. Miyake
T. Matsuda
Softbank

October 30, 2020

PBB-EVPN ISID-based CMAC-Flush
draft-ietf-bess-pbb-evpn-isid-cmacflush-01

Abstract

Provider Backbone Bridging (PBB) can be combined with Ethernet VPN (EVPN) to deploy Ethernet Local Area Network (ELAN) services in large Multi-Protocol Label Switching (MPLS) networks (PBB-EVPN). Single-Active Multi-homing and per-ISID Load-Balancing can be provided to access devices and aggregation networks. In order to speed up the network convergence in case of failures on Single-Active Multi-Homed Ethernet Segments, PBB-EVPN defines a flush mechanism for Customer MACs (CMAC-flush) that works for different Ethernet Segment Backbone MAC (BMAC) address allocation models. This document complements those CMAC-flush procedures for cases in which no PBB-EVPN Ethernet Segments are defined (the attachment circuit is associated to a zero Ethernet Segment Identifier) and a Service Instance Identifier based (ISID-based) CMAC-flush granularity is required.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

[RFC7623] defines how Provider Backbone Bridging (PBB) can be combined with Ethernet VPN (EVPN) to deploy ELAN services in very large MPLS networks. [RFC7623] also describes how Single-Active Multi-homing and per-ISID Load-Balancing can be provided to access devices and aggregation networks. When Access Ethernet/MPLS Networks exists, [I-D.ietf-bess-evpn-virtual-eth-segment] describes how virtual Ethernet Segments can be associated to a group of Ethernet Virtual Circuits (EVCs) or even Pseudowires (PWs). In order to speed up the network convergence in case of failures on Single-Active Multi-Homed Ethernet Segments, [RFC7623] defines a CMAC-flush mechanism that works for different Ethernet Segment BMAC address allocation models.

In some cases, the administrative entities that manage the access devices or aggregation networks don't demand Multi-Homing Ethernet Segments (ES) from the PBB-EVPN provider, but simply multiple single-homed ES. If that is the case, the PBB-EVPN network is no longer aware of the redundancy offered by the access administrative entity. Figure 1 shows an example where the PBB-EVPN network provides four different Attachment Circuits (ACs) for ISID1, with those ACs not being part of any ES or vES (therefore they are referred to as null vES).

```
                     <--PBB-EVPN Network--->
       ISID1     vES +-----+         +-----+
       +----+    null| PE1 +---------+ PE3 |vES null
       |CE1 +--------+ BM1 |         | BM3 | +---------+
       +-+--+     act|     |         |     |=====      |
         |   G.8032  +-+---+         +---+-+ |   \act  | ISID1
         |   Access    |                 |   |    \  +-+--+
         |    Ring     |     IP/MPLS     |   |     ==|CE3 |
         |             |                 |   |    /  +-+--+
         |stb    vES +-+---+         +---+-+ |   /stb  |
       +-+--+    null| PE2 |         | PE4 +-----      |
       |CE2 +--------+ BM2 |         | BM4 | +---------+
       +----+     act|     +---------+     |vES null
       ISID1         +-----+         +-----+ <-MPLS Ag->
                                               Network
```

Figure 1: PBB-EVPN and non-ES based redundancy

In the example in Figure 1, CE1 and CE2 provide redundant connectivity for ISID1 through the use of G.8032 Ethernet Ring Protection Switching. CE3 provides redundant active-standby PW connectivity for ISID1. In the two cases the ACs are connected to null ES, hence the PEs will keep their ACs active and the CEs will be responsible for the per-ISID load balancing while avoiding loops.

For instance, CE2 will block its link to CE1 and CE3 will block its forwarding path to PE4. In this situation, a failure in one of the redundant ACs will make the CEs to start using their redundant paths, however those failures will not trigger any CMAC-flush procedures in the PEs that implement [RFC7623]. For example, if the active PW from CE3 fails, PE3 will not issue any CMAC-flush message and therefore the remote PEs will continue pointing at PE3's BMAC to reach CE3's CMACs, until the CMACs age out in the ISID1 forwarding tables.

[RFC7623] provides a CMAC-flush solution based on a shared BMAC update along with the MAC Mobility extended community where the sequence number is incremented. However, the procedure is only used along with Ethernet Segments. Even if that procedure could be used for null Ethernet Segments, as in the example of Figure 1, the [RFC7623] CMAC-flush procedure would result in unnecessary flushing of unaffected ISIDs on the remote PEs, and subsequent flooding of unknown unicast traffic in the network.

This document describes an extension of the [RFC7623] CMAC-flush procedures, so that in the above failure example, PE3 can trigger a CMAC-flush notification that makes PE1, PE2 and PE4 flush all the CMACs associated to PE3's BMAC and (only) ISID1. This new CMAC-flush procedure explained in this document will be referred to as "PBB-EVPN ISID-based CMAC-flush" and can be used in PBB-EVPN networks with null or non-null (virtual) Ethernet Segments.

1.1. 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.

EVPN: Ethernet Virtual Private Networks, as in [RFC7432].

EVI: EVPN Instance.

MAC-VRF: A Virtual Routing and Forwarding table for MAC addresses.

PBB-EVPN: Provider-Backbone-Bridging and EVPN, as in [RFC7623].

PE: Provider Edge router.

CE: Customer Edge router.

CMAC: Customer MAC address.

BMAC or BM: Backbone MAC address.

ISID: Service Instance Identifier.

B-Component: Backbone Component, as in [RFC7623].

I-Component: Service Instance Component, as in [RFC7623].

PW: Pseudowire.

AC: Attachment Circuit.

ES and ESI: Ethernet Segment and Ethernet Segment Identifier.

Act: Active state, used with ACs or PWs that are operationally active.

Stb: Standby state, used with ACs or PWs that are in a state where they cannot transmit traffic.

G.8032: Ethernet Ring Protection.

RD: Route Distinguisher.

RT: Route Target.

BMAC/ISID route: an EVPN MAC/IP Advertisement route that uses a BMAC in the MAC address field and an ISID in the Ethernet Tag field, and it is used to notify remote PEs about the required CMAC-flush procedure for the CMACs associated with the advertised BMAC and ISID.

BMAC/0 route: an EVPN MAC/IP Advertisement route that uses a BMAC in the MAC address field and a zero Ethernet Tag ID.

Familiarity with the terminology in [RFC7623] is expected.

2. Solution requirements

The following requirements are followed by the CMAC-flush solution described in this document:

a. The solution solves black-hole scenarios in case of failures on null ES ACs (Attachment Circuits not associated to ES, that is, ESI=0) when the access device/network is responsible for the redundancy.

b. This extension works with Single-Active non-null ES and virtual ES, irrespective of the PE BMAC address assignment (dedicated per-ES BMAC or shared BMAC, as in [RFC7623]).

c. In case of failure on the egress PE, the solution provides a CMAC-flush notification at BMAC and ISID granularity level.

d. The solution provides a reliable CMAC-flush notification in PBB-EVPN networks that use Route-Reflectors (RRs), without causing "double flushing" or no flushing for certain ISIDs due to the notification messages being aggregated at the RR.

e. The solution coexists in [RFC7623] networks where there are PEs that do not support this specification.

f. The solution SHOULD be enabled/disabled by an administrative option on a per-PE and per-ISID basis.

3. EVPN BGP Encoding for ISID-based CMAC-flush

The solution does not use any new BGP attributes but reuses the MAC Mobility extended community as an indication of CMAC-flush (as in [RFC7623]) and encodes the ISID in the Ethernet Tag field of the EVPN MAC/IP advertisement route. As a reference, Figure 2 shows the MAC Mobility extended community and the EVPN MAC/IP advertisement route that are used specified in [RFC7432] and used in this document as a CMAC-flush notification message.

```
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=0x03 |   Flags       |   Reserved=0  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       Sequence Number                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

            +---------------------------------------+
            |  RD                                   |
            +---------------------------------------+
            |  ESI = 0                              |
            +---------------------------------------+
            |  Ethernet Tag ID = ISID               |
            +---------------------------------------+
            |  MAC Address Length = 48              |
            +---------------------------------------+
            |  BMAC Address                         |
            +---------------------------------------+
            |   IP Address Length = 0               |
            +---------------------------------------+
            |  MPLS Label1                          |
            +---------------------------------------+
```

Figure 2: CMAC-Flush notification encoding: BMAC/ISID route

Where:

o The route's RD and RT are the ones corresponding to its EVI. Alternatively to the EVI's RT, the route MAY be tagged with an RT auto-derived from the Ethernet Tag (ISID) instead. [RFC7623] describes how the EVPN MAC/IP Advertisement routes can be advertised along with the EVI RT or an RT that is derived from the ISID.

- The Ethernet Tag encodes the ISID for which the PE that receives the route must flush the CMACs upon reception of the route.

- The MAC address field encodes the BMAC Address for which the PE that receives the route must flush the CMACs upon reception of the route.

- The MAC Mobility extended community is used as in [RFC7623], where a delta in the sequence number between two updates for the same BMAC/ISID will be interpreted as a CMAC-flush notification for the corresponding BMAC and ISID.

All the other fields are set and used as defined in [RFC7623]. This document will refer to this route as the BMAC/ISID route, as opposed to the [RFC7623] BMAC/0 route (BMAC route sent with Ethernet Tag ID = 0).

Note that this BMAC/ISID route will be accepted and reflected by any [RFC7432] RR, since no new attributes or values are used. A PE receiving the route will process the received BMAC/ISID update only in case of supporting the procedures described in this document.

4. Solution description

Figure 1 will be used in the description of the solution. CE1, CE2 and CE3 are connected to ACs associated to ISID1, where no (Multi-Homed) Ethernet Segments have been enabled, and the ACs and PWs are in active or standby state as per Figure 1.

Enabling or disabling ISID-based CMAC-flush SHOULD be an administrative choice on the system that MAY be configured per ISID (I-Component). When enabled on a PE:

a. The PE will be able to generate BMAC/ISID routes as CMAC-Flush notifications for the remote PEs.

b. he PE will be able to process BMAC/ISID routes received from remote PEs.

When ISID-based CMAC-flush is disabled, the PE will follow the [RFC7623] procedures for CMAC-flush.

This CMAC-flush specification is described in three sets of procedures:

- ISID-based CMAC-flush activation

- CMAC-flush notification generation upon AC failures

- CMAC-flush process upon receiving a CMAC-flush notification

4.1. ISID-based CMAC-Flush activation procedures

The following behavior MUST be followed by the PBB-EVPN PEs following this specification. Figure 1 is used as a reference.

- As in [RFC7623], each PE advertises a shared BMAC in a BMAC/0 route (with BM1, BM2, BM3 and BM4 in the MAC address field, respectively). This is the BMAC that each PE will use as BMAC SA (Source Address) when encapsulating the frames received on any local single-homed AC. Each PE will import the received BMAC/0 routes from the remote PEs and will install the BMACs in its B-component MAC-VRF. For instance, PE1 will advertise BM1/0 and will install BM2, BM3 and BM4 in its MAC-VRF.

- Assuming ISID-based CMAC-flush is activated for ISID 1, the PEs will advertise the shared BMAC with ISID 1 encoded in the Ethernet Tag. That is, PE1 will advertise BM1/1 and will receive BM2/1, BM3/1 and BM4/1. The receiving PEs MUST use these BMAC/ISID routes only for CMAC-flush procedures and they MUST NOT be used them to add/withdraw any BMAC entry in the MAC-VRFs. As per [RFC7623], only BMAC/0 routes can be used to add/withdraw BMACs in the MAC-VRFs.

- The above procedure MAY also be used for dedicated BMACs (BMACs allocated per Ethernet Segment).

4.2. CMAC-Flush generation

If, for instance, there is a failure on PE1's AC, PE1 will generate an update including BM1/1 along with the MAC Mobility extended community where the Sequence Number has been incremented. The reception of the BM1/1 with a delta in the sequence number will trigger the CMAC-flush procedures on the receiving PEs.

- An AC going operationally down MUST generate a BMAC/ISID with a higher Sequence Number. If the AC going down makes the entire local ISID go operationally down, the PE will withdraw the BMAC/ISID route for the ISID.

- An AC going operationally up SHOULD NOT generate any BMAC/ISID update, unless it activates its corresponding ISID, in which case the PE will advertise the BMAC/ISID route.

- An AC receiving a G.8032 flush notification or a flush message in any other protocol from the access network MAY propagate it to the remote PEs by generating a BMAC/ISID route update with higher Sequence Number.

4.3. CMAC-flush process upon receiving a CMAC-flush notification

A PE receiving a CMAC-flush notification will follow these procedures:

- A received BMAC/ISID route (with non-zero ISID) MUST NOT add/ remove any BMAC to/from the MAC-VRF.

- An update of a previously received BMAC/ISID route with a delta Sequence Number, MUST flush all the CMACs associated to that ISID and BMAC. CMACs associated to the same ISID but different BMAC MUST NOT be flushed.

- A received BMAC/ISID withdraw (with non-zero ISID) MUST flush all the CMACs associated to that BMAC and ISID.

Note that the CMAC-flush procedures described in [RFC7623] for BMAC/0 routes are still valid and a PE receiving [RFC7623] CMAC-flush notification messages MUST observe the behavior specified in [RFC7623].

5. Conclusions

The ISID-based CMAC-flush solution described in this document has the following benefits:

a. The solution solves black-hole scenarios in case of failures on null ES ACs, since the CMAC-flush procedures are independent of the Ethernet Segment definition.

b. This extension can also be used with Single-Active non-null ES and virtual ES, irrespective of the PE BMAC address assignment (dedicated per-ES BMAC or shared BMAC).

c. It provides a CMAC-flush notification at BMAC and ISID granularity level, therefore flushing a minimum number of CMACs and reducing the amount of unknown unicast flooding in the network.

d. It provides a reliable CMAC-flush notification in PBB-EVPN networks that use RRs. RRs will propagate the CMAC-flush notifications for all the affected ISIDs and irrespective of the order in which the notifications make it to the RR.

e. The solution can coexist in a network with systems supporting or not supporting this specification.

6. Security Considerations

Security considerations described in [RFC7623] apply to this document.

In addition, this document suggests additional procedures, that can be activated on a per ISID basis, and generate additional EVPN MAC/IP Advertisement routes in the network. The format of these additional EVPN MAC/IP Advertisement routes is backwards compatible with [RFC7623] procedures and should not create any issues on receiving PEs not following this specification, however, the additional routes may consume extra memory and processing resources on the receiving PEs. Because of that, it is RECOMMENDED to activate this feature only when necessary (when multi-homed networks or devices are attached to the PBB-EVPN PEs), and not by default in any PBB-EVPN PE.

7. IANA Considerations

8. Acknowledgments

The authors want to thank Vinod Prabhu, Sriram Venkateswaran, Laxmi Padakanti, Ranganathan Boovaraghavan for their review and contributions.

9. Contributors

10. References

10.1. Normative References

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <https://www.rfc-editor.org/info/rfc7623>.

[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>.

[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>.

10.2. Informative References

[I-D.ietf-bess-evpn-virtual-eth-segment]
Sajassi, A., Brissette, P., Schell, R., Drake, J., and J. Rabadan, "EVPN Virtual Ethernet Segment", draft-ietf-bess-evpn-virtual-eth-segment-06 (work in progress), March 2020.

Authors' Addresses

Jorge Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: senthil.sathappan@nokia.com

Kiran Nagaraj
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: kiran.nagaraj@nokia.com

M. Miyake
Softbank

Email: masahiro.miyake@g.softbank.co.jp

T. Matsuda
Softbank

Email: taku.matsuda@g.softbank.co.jp

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: February 19, 2021

A. Sajassi
M. Mishra
S. Thoria
P. Brissette
Cisco Systems
J. Rabadan
Nokia
J. Drake
Juniper Networks

August 18, 2020

AC-Aware Bundling Service Interface in EVPN
draft-sajassi-bess-evpn-ac-aware-bundling-02

Abstract

EVPN provides an extensible and flexible multi-homing VPN solution over an MPLS/IP network for intra-subnet connectivity among Tenant Systems and End Devices that can be physical or virtual.

EVPN multihoming with IRB is one of the common deployment scenarios. There are deployments which requires capability to have multiple subnets designated with multiple VLAN IDs in single bridge domain.

[RFC7432] defines three different type of service interface which serve different requirements but none of them address the requirement to be able to support multiple subnets within single bridge domain. In this draft we define new service interface type to support multiple subnets in single bridge domain. Service interface proposed in this draft will be applicable to multihoming case only.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

EVPN based All-Active multi-homing is becoming the basic building block for providing redundancy in next generation data center deployments as well as service provider access/aggregation network. For EVPN IRB mode, there are deployments which expect to be able to support multiple subnets within single Bridge Domain. Each subnet would be differentiated by VLAN. Thus, single IRB interface can still serve multiple subnet.

Motivation behind such deployments are

1. Manageability: If there is support to have multiple subnets using single bridge domain, it would require only one Bridge domain and one IRB for "N" subnets compare to "N" Bridge domain and "N" IRB interface to manage.

2. Simplicity: It avoids extra configuration by configuring Vlan Range as compare to individual VLAN, BD and IRB interface per subnet.

Multiple subnet per bridge domain deployments guarantee that there would not be duplicate MAC address across subnet.

[RFC7432] defines three types of service interface. None of them provide flexibility to achieve multiple subnet within single bridge domain. Brief about existing service interface from [RFC7432] are ,

1. VLAN-Based Service Interface: With this service interface, an EVPN instance consists of only a single broadcast domain (e.g., a single VLAN). Therefore, there is a one-to-one mapping between a VID on this interface and a MAC-VRF.

2. VLAN Bundle Service Interface: With this service interface, an EVPN instance corresponds to multiple broadcast domains (e.g., multiple VLANs); however, only a single bridge table is maintained per MAC-VRF, which means multiple VLANs share the same bridge table. The MPLS-encapsulated frames MUST remain tagged with the originating VID. Tag translation is NOT permitted. The Ethernet Tag ID in all EVPN routes MUST be set to 0.

3. VLAN-Aware Bundle Service Interface: With this service interface, an EVPN instance consists of multiple broadcast domains (e.g., multiple VLANs) with each VLAN having its own bridge table -- i.e., multiple bridge tables (one per VLAN) are maintained by a single MAC-VRF corresponding to the EVPN instance.

Though from definition it looks like VLAN Bundle Service Interface does provide flexibility to support multiple subnet within single bridge domain. But it requirement is to have multiple subnet from same ES on multi-homing all active mode, it would not work. For example, lets take the case from Figure-1, If PE1 learns MAC of H1 on Vlan 1 (subnet S1). When MAC route is originated , as per [RFC7432] ether tag would be set to 0. If there is packet coming from IRB interface which is untagged packet, and it reaches to PE2, PE2 does not have associated AC information. In this case PE2 can not forward traffic which is destined to H1.

This draft proposes an extension to existing service interface types defined in [RFC7432] and defines AC-aware Bundling service interface. AC-aware Bundling service interface would provide mechanism to have multiple subnets in single bridge domain. This extension is applicable only for multi-homed EVPN peers..

```
                                 H3
                                  |
                             +----+-----+
                             |          |
                             |   PE3    | EVI-1
                             |          |
           +-----------------+----------+----------------------+
           |                                                   |
           |                                                   |
           |                                                   |
           |                  IP MPLS core                     |
           |                                                   |
           |                                                   |
           |                                                   |
           |                                                   |
           +------+----------------------------------------+--+
                  |                                        |
   +--------------+----+                      +------------+------+
   |                   |                      |                   |
   |       PE1         |                      |           PE2     |
   |                   |                      |                   |
   |      +-----+      |                      |      +-----+      |
   |      | IRB |      |                      |      | IRB |      |
   |   +--+-----+--+   |                      |   +--+-----+--+   |
   |   |  BD & EVI |   |                      |   |  BD & EVI |   |
   |   +--+--+--+--+   |                      |   +-----------+   |
   |   |S1|S2|S3|S4|   |                      |   |S1|S2|S3|S4|   |
   +---+--+-X+--+--+---+                      +---+--+-X+--+--+---+
               X                                    X
                  X                              X
                     X                        X  ESI-100
                        X                  X     EVI-1
                           X            X        BD-1
                              X      X
                                 XX
                              +-------+
                              |  CE   |
                              +-+--+--+
                                |  |
                               H1  H2
```

Figure 1: EVPN topology with multi-homing and non multihoming peer

The above figure shows sample EVPN topology, PE1 and PE2 are multihomed peers. PE3 is remote peer which is part of same EVPN instance (evi1). It is showing four subnets S1, S2, S3, S4 where numeric value provides associated Vlan information.

1.1. Problem with Unicast MAC route processing for multihome case

BD-1 has multiple subnet where each subnet is distinguished by Vlan 1, 2 ,3 and 4. PE1 learns MAC address MAC-1 from AC associated with subnet S1. PE1 uses MAC route to advertise MAC-1 presence to peer PEs. As per [RFC7432] MAC route advertisement from PE1 does not carry any context which can provide information about MAC address association with AC. When PE2 receives MAC route with MAC-2 it can not determine which AC this MAC belongs too.

Since PE2 could not bind MAC-1 with correct AC, when it receives data traffic destined to MAC-1, it can not find correct AC where data MUST be forwarded.

1.2. Problem with Multicast route synchronization

[I-D.ietf-bess-evpn-igmp-mld-proxy] defines mechanism to synchronize multicast routes between multihome peer. In above case if Receiver behind S1 send IGMP membership request, CE could hash it to either of the PE. When Multicast route is originated, it does not contain any AC information. Once it reaches to remote PE, it does not have any information about which subnet this IGMP membership request belong to.

1.3. Potential Security concern caused by misconfiguration

In case of single subnet per bridge domain, there is potential case of security issue. For example if PE1 , BD1 is configured with Vlan-1 where as multihome peer PE2 has configured Vlan-2. Now each of the IGMP membership request on PE1 would be synchronized to PE2. and PE2 would process multicast routes and start forwarding multicast traffic on Vlan-2, which was not intended.

2. Terminology

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] .

AC: Attachment Circuit.

ARP: Address Resolution Protocol.

BD: Broadcast Domain. As per [RFC7432], an EVI consists of a single or multiple BDs. In case of VLAN-bundle and VLAN-based service models (see [RFC7432]), a BD is equivalent to an EVI. In case of VLAN-aware bundle service model, an EVI contains multiple BDs. Also, in this document, BD and subnet are equivalent terms.

BD Route Target: refers to the Broadcast Domain assigned Route Target [RFC4364]. In case of VLAN-aware bundle service model, all the BD instances in the MAC-VRF share the same Route Target.

BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per [RFC7432].

DGW: Data Center Gateway.

Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].

Ethernet NVO tunnel: refers to Network Virtualization Overlay tunnels with Ethernet payload. Examples of this type of tunnels are VXLAN or GENEVE.

EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].

EVPN: Ethernet Virtual Private Networks, as per [RFC7432].

GRE: Generic Routing Encapsulation.

GW IP: Gateway IP Address.

IPL: IP Prefix Length.

IP NVO tunnel: it refers to Network Virtualization Overlay tunnels with IP payload (no MAC header in the payload)

IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by EVPN and IP-VPN address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on an NVE/PE, as per [RFC7432]. A MAC-VRF is also an instantiation of an EVI in an NVE/PE.

ML: MAC address length.

ND: Neighbor Discovery Protocol.

NVE: Network Virtualization Edge.

GENEVE: Generic Network Virtualization Encapsulation, [GENEVE].

NVO: Network Virtualization Overlays.

RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].

RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].

SBD: Supplementary Broadcast Domain. A BD that does not have any ACs, only IRB interfaces, and it is used to provide connectivity among all the IP-VRFs of the tenant. The SBD is only required in IP-VRF- to-IP-VRF use-cases (see Section 4.4.).

SN: Subnet.

TS: Tenant System.

VA: Virtual Appliance.

VNI: Virtual Network Identifier. As in [RFC8365], the term is used as a representation of a 24-bit NVO instance identifier, with the understanding that VNI will refer to a VXLAN Network Identifier in VXLAN, or Virtual Network Identifier in GENEVE, etc. unless it is stated otherwise.

VTEP: VXLAN Termination End Point, as in [RFC7348].

VXLAN: Virtual Extensible LAN, as in [RFC7348].

This document also assumes familiarity with the terminology of [RFC7432],[RFC8365], [RFC7365].

3. Requirements

1. Service interface MUST be able to support multiple subnets designated by Vlan under single bridge domain.

2. Service interface MUST be applicable to Multihomed peers only

3. New Service interface handling procedure MUST make sure to have backward compatibility with implementation procedures defined in [RFC7432]

4. New Service interface MUST be extendible to multicast routes defined in [I-D.ietf-bess-evpn-igmp-mld-proxy] too.

4. Solution Description

```
                                 H3
                                  |
                             +----+-----+
                             |          |
                             |   PE3    | EVI-1
                             |          |
           +-----------------+----------+----------------------+
           |                                                   |
           |                                                   |
           |                                                   |
           |                  IP MPLS core                     |
           |                                                   |
           |                                                   |
           |                                                   |
           |                                                   |
           +------+----------------------------------------+--+
                  |                                        |
   +--------------+----+                      +------------+------+
   |                   |                      |                   |
   |       PE1         |                      |           PE2     |
   |                   |                      |                   |
   |      +-----+      |                      |      +-----+      |
   |      | IRB |      |                      |      | IRB |      |
   |   +--+-----+--+   |                      |   +--+-----+--+   |
   |   |  BD & EVI |   |                      |   |  BD & EVI |   |
   |   +--+--+--+--+   |                      |   +-----------+   |
   |   |S1|S2|S3|S4|   |                      |   |S1|S2|S3|S4|   |
   +---+--+-X+--+--+---+                      +---+--+-X+--+--+---+
               X                                    X
                  X                              X
                     X                        X  ESI-100
                        X                  X     EVI-1
                           X            X        BD-1
                              X      X
                                 XX
                              +-------+
                              |  CE   |
                              +-+--+--+
                                |  |
                               H1  H2
                            MAC-1  MAC-2
                           Vlan-1  Vlan-2
                           (S,G)   (S,G)  ------> Multicast receiver
```

Figure 2: AC aware bundling procedures

Consider the above topology, where AC aware bundling service interface is supported. Host H1 on Vlan-1 has MAC address as MAC-1 and Host H2 on Vlan 2 has MAC address as MAC-2.

4.1. Control Plane Operation

4.1.1. MAC/IP Address Advertisement

4.1.1.1. Local Unicast MAC learning

1. [RFC7432] section 9.1 describes different mechanism to learn Unicast MAC address locally. PEs where AC aware bundling is supported, MAC address is learnt along with Vlan associated with AC.

2. MAC/IP route construction follows mechanism defined in [RFC7432] section 9.2.1. Along with RT-2 it must attach Attachment Circuit ID Extended Community (Section 6.1).

3. From Figure-2 PE1 learns MAC-1 on S1. It MUST construct MAC route with procedure defined in [RFC7432] section 9.2.1. It MUST attach Attachment Circuit ID Extended Community (Section 6.1).

4.1.1.2. Remote Unicast MAC learning

1. Presence of Attachment Circuit ID Extended Community (Section 6.1) MUST be ignored by non multihoming PEs. Remote PE (Non Multihome PE) MUST process MAC route as defined in [RFC7432]

2. Multihoming peer MUST process Attachment Circuit ID Extended Community (Section 6.1) to attach remote MAC address to appropriate AC.

3. From Figure-2 PE3 receives MAC route for MAC-1. It MUST not ignore AC information in Attachment Circuit ID Extended Community (Section 6.1) which was received with RT-2.

4. PE2 receives MAC route for MAC-1. It MUST get Attachment Circuit ID from Attachment Circuit ID Extended Community (Section 6.1) in RT-2 and associate MAC address with specific subnet.

4.1.2. Multicast route Advertisement

4.1.2.1. Local multicast state

When a local multihomed bridge port in given BD receives IGMP membership request and ES is operating in All-active or Single-Active redundancy mode, it MUST synchronize multicast state by originating multicast route defined in section 7 of [I-D.ietf-bess-evpn-igmp-mld-proxy]. When Service interface is AC aware it MUST attach Attachment Circuit ID Extended Community (Section 6.1) along with multicast route. For example in Figure-2 when H2 sends IGMP membership request for (S,G) , CE hashed it to one of the PE. Lets say PE1 received IGMP membership request, now PE1 MUST originate multicast route to synchronize multicast state with PE2. Multicast route MUST contain Attachment Circuit ID Extended Community (Section 6.1) along with multicast route.

If PE1 had already originated multicast route for (S,G) from subnet S2. Now if host H1 also sends IGMP membership request for (S,G) on subnet S1, PE1 MUST originate route update with Attachment Circuit ID Extended Community (Section 6.1).

4.1.2.2. Remote multicast state

If multihomed PE receives remote multicast route on Bridge Domain for given ES, route MUST be programmed to correct subnet. Subnet information MUST be get from Attachment Circuit ID Extended Community. For example PE2 receives multicast route on Bridge Domain BD-1 for ES ESI-100, From Attachment Circuit ID Extended Community (Section 6.1) it receives AC information and associates multicast route (S,G) to subnet S2.

When PE2 receives route update with Attachment Circuit ID Extended Community added for subnet S1, port associated with subnet S1 MUST be added for multicast route.

4.2. Data Plane Operation

4.2.1. Unicast Forwarding

1. Packet received from CE must follow same procedure as defined in [RFC7432] section 13.1

2. Unknown Unicast packets from a Remote PE MUST follow procedure as per [RFC7432] section 13.2.1.

3. Known unicast Received on a Remote PE MUST follow procedure as per [RFC7432] section 13.2.2. So in Figure-2 if PE3 receives known unicast packet for destination MAC MAC-1, it MUST follow procedure defined in [RFC7432] section 13.2.2.

4. If destination MAC lookup is performed on known unicast packet, destination MAC lookup MUST provide Vlan and Port tuple. For example if PE2 receives unicast packet which is destined to MAC-1 (packet might be coming from IRB or remote PE with EVPN tunnel),

destination MAC lookup on PE2 MUST provide outgoing port along with associated MAC address. In this case traffic MUST be forwarded to S1 with Vlan 1.

4.2.2. Multicast Forwarding

1. Multicast traffic from CE and remote PE MUST follow procedure defined in [RFC7432]

2. Multicast traffic received from IRB interface or EVPN tunnel, route lookup would be performed based on IGMP snooping state and traffic would be forwarded to appropriate AC.

5. Mis-configuration of VLAN ranges across multihoming peer

If there is mis-configuration of Vlan or Vlan range across multihoming peer, same MAC address would be learnt with different Vlan in Bridge Domain. In this case Error message MUST be thrown for operator to make configuration changes. errored MAC route MUST be ignored.

6. BGP Encoding

This document defines one new BGP Extended Community for EVPN.

6.1. Attachment Circuit ID Extended Community

A new EVPN BGP Extended Community called Attachment Circuit ID is introduced here. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of TBD. It is advertised along with EVPN MAC/IP Advertisement Route (Route Type 2) per [RFC7432] for AC-Aware Bundling Service Interface.

The Attachment Circuit ID Extended Community is encoded as an 8-octet value 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=0x06     | Sub-Type=TBD  |      Reserved (16 bits)       |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |               Attachment Circuit ID (32 bits)                 |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Attachment Circuit ID Extended Community

This extended community is used to carry the Attachment Circuit ID associated with the received MAC address and it is advertised along with EVPN MAC/IP Advertisement route. The receiving PE who is a member of an All-Active multi-homing group uses this information to not only synchronize the MAC address but also the associated AC over which the MAC addresses is received.

7. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

8. IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of TBD for EVPN Attachment Circuit Extended Community needs to be allocated by IANA.

9. Acknowledgement

10. References

10.1. Normative References

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-00 (work in progress), March 2017.

[I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-10 (work in progress), August 2018.

10.2. Informative 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>.

[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, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

[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>.

Authors' Addresses

Ali Sajassi
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sajassi@cisco.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: mankamis@cisco.com

Samir Thoria
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sthoria@cisco.com

Patrice Brissette
Cisco Systems

Email: pbrisset@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
UNITED STATES

Email: jorge.rabadan@nokia.com

John Drake
Juniper Networks

Email: jdrake@juniper.net

BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: May 6, 2021

J. Rabadan, Ed.
J. Kotalwar
S. Sathappan
Nokia
Z. Zhang
W. Lin
Juniper
E. Rosen
Individual

November 2, 2020

Multicast Source Redundancy in EVPN Networks
draft-skr-bess-evpn-redundant-mcast-source-02

Abstract

EVPN supports intra and inter-subnet IP multicast forwarding. However, EVPN (or conventional IP multicast techniques for that matter) do not have a solution for the case where: a) a given multicast group carries more than one flow (i.e., more than one source), and b) it is desired that each receiver gets only one of the several flows. Existing multicast techniques assume there are no redundant sources sending the same flow to the same IP multicast group, and, in case there were redundant sources, the receiver's application would deal with the received duplicated packets. This document extends the existing EVPN specifications and assumes that IP Multicast source redundancy may exist. It also assumes that, in case two or more sources send the same IP Multicast flows into the tenant domain, the EVPN PEs need to avoid that the receivers get packet duplication by following the described procedures.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Intra and Inter-subnet IP Multicast forwarding are supported in EVPN networks. [I-D.ietf-bess-evpn-igmp-mld-proxy] describes the procedures required to optimize the delivery of IP Multicast flows when Sources and Receivers are connected to the same EVPN BD

(Broadcast Domain), whereas [I-D.ietf-bess-evpn-irb-mcast] specifies the procedures to support Inter-subnet IP Multicast in a tenant network. Inter-subnet IP Multicast means that IP Multicast Source and Receivers of the same multicast flow are connected to different BDs of the same tenant.

[I-D.ietf-bess-evpn-igmp-mld-proxy], [I-D.ietf-bess-evpn-irb-mcast] or conventional IP multicast techniques do not have a solution for the case where a given multicast group carries more than one flow (i.e., more than one source) and it is desired that each receiver gets only one of the several flows. Multicast techniques assume there are no redundant sources sending the same flows to the same IP multicast group, and, in case there were redundant sources, the receiver’s application would deal with the received duplicated packets.

As a workaround in conventional IP multicast (PIM or MVPN networks), if all the redundant sources are given the same IP address, each receiver will get only one flow. The reason is that, in conventional IP multicast, (S,G) state is always created by the RP (Rendezvous Point), and sometimes by the Last Hop Router (LHR). The (S,G) state always binds the (S,G) flow to a source-specific tree, rooted at the source IP address. If multiple sources have the same IP address, one may end up with multiple (S,G) trees. However, the way the trees are constructed ensures that any given LHR or RP is on at most one of them. The use of an anycast address assigned to multiple sources may be useful for warm standby redundancy solutions. However, on one hand, it’s not really helpful for hot standby redundancy solutions and on the other hand, configuring the same IP address (in particular IPv4 address) in multiple sources may bring issues if the sources need to be reached by IP unicast traffic or if the sources are attached to the same Broadcast Domain.

In addition, in the scenario where several G-sources are attached via EVPN/OISM, there is not necessarily any (S,G) state created for the redundant sources. The LHRs may have only (*,G) state, and there may not be an RP (creating (S,G) state) either. Therefore, this document extends the above two specifications and assumes that IP Multicast source redundancy may exist. It also assumes that, in case two or more sources send the same IP Multicast flows into the tenant domain, the EVPN PEs need to avoid that the receivers get packet duplication.

The solution provides support for Warm Standby (WS) and Hot Standby (HS) redundancy. WS is defined as the redundancy scenario in which the upstream PEs attached to the redundant sources of the same tenant, make sure that only one source of the same flow can send multicast to the interested downstream PEs at the same time. In HS the upstream PEs forward the redundant multicast flows to the downstream PEs, and the downstream PEs make sure only one flow is forwarded to the interested attached receivers.

1.1. Terminology

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.

- PIM: Protocol Independent Multicast.

- MVPN: Multicast Virtual Private Networks.

- OISM: Optimized Inter-Subnet Multicast, as in [I-D.ietf-bess-evpn-irb-mcast].

- Broadcast Domain (BD): an emulated ethernet, such that two systems on the same BD will receive each other's link-local broadcasts. In this document, BD also refers to the instantiation of a Broadcast Domain on an EVPN PE. An EVPN PE can be attached to one or multiple BDs of the same tenant.

- Designated Forwarder (DF): as defined in [RFC7432], an ethernet segment may be multi-homed (attached to more than one PE). An ethernet segment may also contain multiple BDs, of one or more EVIs. For each such EVI, one of the PEs attached to the segment becomes that EVI's DF for that segment. Since a BD may belong to only one EVI, we can speak unambiguously of the BD's DF for a given segment.

- Upstream PE: in this document an Upstream PE is referred to as the EVPN PE that is connected to the IP Multicast source or closest to it. It receives the IP Multicast flows on local ACs (Attachment Circuits).

- Downstream PE: in this document a Downstream PE is referred to as the EVPN PE that is connected to the IP Multicast receivers and gets the IP Multicast flows from remote EVPN PEs.

- G-traffic: any frame with an IP payload whose IP Destination Address (IP DA) is a multicast group G.

- G-source: any system sourcing IP multicast traffic to G.

- SFG: Single Flow Group, i.e., a multicast group address G which represents traffic that contains only a single flow. However, multiple sources - with the same or different IP - may be transmitting an SFG.

- o Redundant G-source: a host or router that transmits an SFG in a tenant network where there are more hosts or routers transmitting the same SFG. Redundant G-sources for the same SFG SHOULD have different IP addresses, although they MAY have the same IP address when in different BDs of the same tenant network. Redundant G-sources are assumed NOT to be "bursty" in this document (typical example are Broadcast TV G-sources or similar).

- o P-tunnel: Provider tunnel refers to the type of tree a given upstream EVPN PE uses to forward multicast traffic to downstream PEs. Examples of P-tunnels supported in this document are Ingress Replication (IR), Assisted Replication (AR), Bit Indexed Explicit Replication (BIER), multicast Label Distribution Protocol (mLDP) or Point to Multi-Point Resource Reservation protocol with Traffic Engineering extensions (P2MP RSVP-TE).

- o Inclusive Multicast Tree or Inclusive Provider Multicast Service Interface (I-PMSI): defined in [RFC6513], in this document it is applicable only to EVPN and refers to the default multicast tree for a given BD. All the EVPN PEs that are attached to a specific BD belong to the I-PMSI for the BD. The I-PMSI trees are signaled by EVPN Inclusive Multicast Ethernet Tag (IMET) routes.

- o Selective Multicast Tree or Selective Provider Multicast Service Interface (S-PMSI): defined in [RFC6513], in this document it is applicable only to EVPN and refers to the multicast tree to which only the interested PEs of a given BD belong to. There are two types of EVPN S-PMSIs:

  * EVPN S-PMSIs that require the advertisement of S-PMSI AD routes from the upstream PE, as in [EVPN-BUM]. The interested downstream PEs join the S-PMSI tree as in [EVPN-BUM].

  * EVPN S-PMSIs that don’t require the advertisement of S-PMSI AD routes. They use the forwarding information of the IMET routes, but upstream PEs send IP Multicast flows only to downstream PEs issuing Selective Multicast Ethernet Tag (SMET) routes for the flow. These S-PMSIs are only supported with the following P-tunnels: Ingress Replication (IR), Assisted Replication (AR) and BIER.

This document also assumes familiarity with the terminology of [RFC7432], [RFC4364], [RFC6513], [RFC6514], [I-D.ietf-bess-evpn-igmp-mld-proxy], [I-D.ietf-bess-evpn-irb-mcast], [EVPN-RT5] and [EVPN-BUM].

1.2. Background on IP Multicast Delivery in EVPN Networks

IP Multicast is all about forwarding a single copy of a packet from a source S to a group of receivers G along a multicast tree. That multicast tree can be created in an EVPN tenant domain where S and the receivers for G are connected to the same BD or different BD. In the former case, we refer to Intra-subnet IP Multicast forwarding, whereas the latter case will be referred to as Inter-subnet IP Multicast forwarding.

1.2.1. Intra-subnet IP Multicast Forwarding

When the source S1 and receivers interested in G1 are attached to the same BD, the EVPN network can deliver the IP Multicast traffic to the receivers in two different ways (Figure 1):

```
                     S1  +                        S1  +
           (a)       +   |              (b)       +   |
                     |   | (S1,G1)                |   | (S1,G1)
                 PE1 |   |                    PE1 |   |
                 +-----+ v                    +-----+ v
                 |+---+|                      |+---+|
                 ||BD1||                      ||BD1||
                 |+---+|                      |+---+|
                 +-----+                      +-----+
            +-------|-------+            +-------|
            |       |       |            |       |
            v       v       v            v       v
         +-----+ +-----+ +-----+      +-----+ +-----+ +-----+
         |+---+| |-----| |-----|      |+---+| |+---+| |+---+|
         ||BD1|| ||BD1|| ||BD1||      ||BD1|| ||BD1|| ||BD1||
         |+---+| |-----| |-----|      |+---+| |+---+| |+---+|
         +-----+ +-----+ +-----+      +-----+ +-----+ +-----+
         PE2|    PE3|    PE4|         PE2|    PE3|    PE4
          - | - - - | -     |          - | - - - | -
         |  |       |  |    |         |  |       |  |
            v       v       v            v       v
         |  R1      R2 |    R3        |  R1      R2 |    R3
          - - - G1- - -                - - - G1- - -
```

Figure 1: Intra-subnet IP Multicast

Model (a) illustrated in Figure 1 is referred to as "IP Multicast delivery as BUM traffic". This way of delivering IP Multicast traffic does not require any extensions to [RFC7432], however, it sends the IP Multicast flows to non-interested receivers, such as e.g., R3 in Figure 1. In this example, downstream PEs can snoop IGMP/MLD messages from the receivers so that layer-2 multicast state is created and, for instance, PE4 can avoid sending (S1,G1) to R3, since R3 is not interested in (S1,G1).

Model (b) in Figure 1 uses an S-PMSI to optimize the delivery of the (S1,G1) flow. For instance, assuming PE1 uses IR, PE1 sends (S1,G1) only to the downstream PEs that issued an SMET route for (S1,G1), that is, PE2 and PE3. In case PE1 uses any P-tunnel different than IR, AR or BIER, PE1 will advertise an S-PMSI A-D route for (S1,G1) and PE2/PE2 will join that tree.

Procedures for Model (b) are specified in [I-D.ietf-bess-evpn-igmp-mld-proxy].

1.2.2. Inter-subnet IP Multicast Forwarding

If the source and receivers are attached to different BDs of the same tenant domain, the EVPN network can also use Inclusive or Selective Trees as depicted in Figure 2, models (a) and (b) respectively.

```
                     S1  +                     S1  +
           (a)       +   |              (b)    +   |
                     |   | (S1,G1)             |   | (S1,G1)
                 PE1 |   |                 PE1 |   |
                 +-----+ v                 +-----+ v
                 |+---+|                   |+---+|
                 ||BD1||                   ||BD1||
                 |+---+|                   |+---+|
                 +-----+                   +-----+
            +-------|-------+         +-------|
            |       |       |         |       |
            v       v       v         v       v
         +-----+ +-----+ +-----+   +-----+ +-----+ +-----+
         |+---+| |+---+| |+---+|   |+---+| |+---+| |+---+|
         ||SBD|| ||SBD|| ||SBD||   ||SBD|| ||SBD|| ||SBD||
         |+-|-+| |+-|-+| |+---+|   |+-|-+| |+-|-+| |+---+|
         | VRF | | VRF | | VRF |   | VRF | | VRF | | VRF |
         |+-v-+| |+-v-+| |+---+|   |+-v-+| |+-v-+| |+---+|
         ||BD2|| ||BD3|| ||BD4||   ||BD2|| ||BD3|| ||BD4||
         |+-|-+| |+-|-+| |+---+|   |+-|-+| |+-|-+| |+---+|
         +--|--+ +--|--+ +-----+   +--|--+ +--|--+ +-----+
         PE2|    PE3|    PE4       PE2|    PE3|    PE4
          - | - - - | -             - | - - - | -
         |  |       |  |           |  |       |  |
            v       v                 v       v
         |  R1      R2 |    R3     |  R1      R2 |    R3
          - - - G1- - -             - - - G1- - -
```

Figure 2: Inter-subnet IP Multicast

[I-D.ietf-bess-evpn-irb-mcast] specifies the procedures to optimize the Inter-subnet Multicast forwarding in an EVPN network. The IP Multicast flows are always sent in the context of the source BD. As described in [I-D.ietf-bess-evpn-irb-mcast], if the downstream PE is not attached to the source BD, the IP Multicast flow is received on the SBD (Supplementary Broadcast Domain), as in the example in Figure 2.

[I-D.ietf-bess-evpn-irb-mcast] supports Inclusive or Selective Multicast Trees, and as explained in Section 1.2.1, the Selective Multicast Trees are setup in a different way, depending on the P-tunnel being used by the source BD. As an example, model (a) in Figure 2 illustrates the use of an Inclusive Multicast Tree for BD1 on PE1. Since the downstream PEs are not attached to BD1, they will all receive (S1,G1) in the context of the SBD and will locally route the flow to the local ACs. Model (b) uses a similar forwarding model, however PE1 sends the (S1,G1) flow in a Selective Multicast Tree. If the P-tunnel is IR, AR or BIER, PE1 does not need to advertise an S-PMSI A-D route.

[I-D.ietf-bess-evpn-irb-mcast] is a superset of the procedures in [I-D.ietf-bess-evpn-igmp-mld-proxy], in which sources and receivers can be in the same or different BD of the same tenant. [I-D.ietf-bess-evpn-irb-mcast] ensures every upstream PE attached to a source will learn of all other PEs (attached to the same Tenant Domain) that have interest in a particular set of flows. This is because the downstream PEs advertise SMET routes for a set of flows with the SBD’s Route Target and they are imported by all the Upstream PEs of the tenant. As a result of that, inter-subnet multicasting can be done within the Tenant Domain, without requiring any Rendezvous Points (RP), shared trees, UMH selection or any other complex aspects of conventional multicast routing techniques.

1.3. Multi-Homed IP Multicast Sources in EVPN

Contrary to conventional multicast routing technologies, multi-homing PEs attached to the same source can never create IP Multicast packet duplication if the PEs use a multi-homed Ethernet Segment (ES). Figure 3 illustrates this by showing two multi-homing PEs (PE1 and PE2) that are attached to the same source (S1). We assume that S1 is connected to an all-active ES by a layer-2 switch (SW1) with a Link Aggregation Group (LAG) to PE1 and PE2.

```
                                     S1
                                     |
                                     v
                                  +-----+
                                  | SW1 |
                                  +-----+
                            +----  |   |
                     (S1,G1)| +----+   +----+
         IGMP               | | all-active  |
         J(S1,G1)     PE1   v |    ES-1     |    PE2
         +---->   +-----------|---+     +---|-----------+
                  | +---+   +---+ |     | +---+         |
          R1  <-----|BD2|   |BD1| |     | |BD1|         |
                  | +---+---+---+ |     | +---+---+     |
             +----|     |VRF|  |  |     |     |VRF|     |----+
             |    | +---+---+  |  |     | +---+---+     |    |
             |    | |SBD|      |  |     | |SBD|         |    |
             |    | +---+      |  |     | +---+         |    |
             |    +------------|--+     +---------------+    |
             |                 |                             |
             |                 |                             |
             |                 |                             |
             |  EVPN           |               ^             |
             |  OISM           v    PE3        | SMET        |
             |              +---------------+  | (*,G1)      |
             |              | +---+         |  |             |
             |              | |SBD|         |                |
             |              | +---+---+     |                |
             +--------------|     |VRF|     |----------------+
                            | +---+---+---+ |
                            | |BD2|   |BD3| |
                            | +-|-+   +-|-+ |
                            +---|-------|---+
                            ^   |       |   ^
                   IGMP     |   v       v   | IGMP
                    J(*,G1) |  R2       R3  | J(S1,G1)
```

Figure 3: All-active Multi-homing and OISM

When receiving the (S1,G1) flow from S1, SW1 will choose only one link to send the flow, as per [RFC7432]. Assuming PE1 is the receiving PE on BD1, the IP Multicast flow will be forwarded as soon as BD1 creates multicast state for (S1,G1) or (*,G1). In the example of Figure 3, receivers R1, R2 and R3 are interested in the multicast flow to G1. R1 will receive (S1,G1) directly via the IRB interface as per [I-D.ietf-bess-evpn-irb-mcast]. Upon receiving IGMP reports from R2 and R3, PE3 will issue an SMET (*,G1) route that will create state in PE1's BD1. PE1 will therefore forward the IP Multicast flow to PE3’s SBD and PE3 will forward to R2 and R3, as per [I-D.ietf-bess-evpn-irb-mcast] procedures.

When IP Multicast source multi-homing is required, EVPN multi-homed Ethernet Segments MUST be used. EVPN multi-homing guarantees that only one Upstream PE will forward a given multicast flow at the time, avoiding packet duplication at the Downstream PEs. In addition, the SMET route for a given flow creates state in all the multi-homing Upstream PEs. Therefore, in case of failure on the Upstream PE forwarding the flow, the backup Upstream PE can forward the flow immediately.

This document assumes that multi-homing PEs attached to the same source always use multi-homed Ethernet Segments.

1.4. The Need for Redundant IP Multicast Sources in EVPN

While multi-homing PEs to the same IP Multicast G-source provides certain level of resiliency, multicast applications are often critical in the Operator’s network and greater level of redundancy is required. This document assumes that:

a. Redundant G-sources for an SFG may exist in the EVPN tenant network. A Redundant G-source is a host or a router that sends an SFG in a tenant network where there is another host or router sending traffic to the same SFG.

b. Those redundant G-sources may be in the same BD or different BDs of the tenant. There must not be restrictions imposed on the location of the receiver systems either.

c. The redundant G-sources can be single-homed to only one EVPN PE or multi-homed to multiple EVPN PEs.

d. The EVPN PEs must avoid duplication of the same SFG on the receiver systems.

2. Solution Overview

An SFG is represented as (*,G) if any source that issues multicast traffic to G is a redundant G-source. Alternatively, this document allows an SFG to be represented as (S,G), where S is a prefix of any length. In this case, a source is considered a redundant G-source for the SFG if it is contained in the prefix. This document allows variable length prefixes in the Sources advertised in S-PMSI A-D routes only for the particular application of redundant G-sources.

There are two redundant G-source solutions described in this document:

- Warm Standby (WS) Solution

- Hot Standby (HS) Solution

The WS solution is considered an upstream-PE-based solution (since downstream PEs do not participate in the procedures), in which all the upstream PEs attached to redundant G-sources for an SFG represented by (*,G) or (S,G) will elect a "Single Forwarder" (SF) among themselves. Once a SF is elected, the upstream PEs add an Reverse Path Forwarding (RPF) check to the (*,G) or (S,G) state for the SFG:

- A non-SF upstream PE discards any (*,G)/(S,G) packets received over a local AC.

- The SF accepts and forwards any (*,G)/(S,G) packets it receives over a single local AC (for the SFG). In case (*,G)/(S,G) packets for the SFG are received over multiple local ACs, they will be discarded in all the local ACs but one. The procedure to choose the local AC that accepts packets is a local implementation matter.

A failure on the SF will result in the election of a new SF. The Election requires BGP extensions on the existing EVPN routes. These extensions and associated procedures are described in Section 3 and Section 4 respectively.

In the HS solution the downstream PEs are the ones avoiding the SFG duplication. The upstream PEs are aware of the locally attached G-sources and add a unique Ethernet Segment Identifier label (ESI-label) per SFG to the SFG packets forwarded to downstream PEs. The downstream PEs pull the SFG from all the upstream PEs attached to the redundant G-sources and avoid duplication on the receiver systems by adding an RPF check to the (*,G) state for the SFG:

- A downstream PE discards any (*,G) packets it receives from the "wrong G-source".

- The wrong G-source is identified in the data path by an ESI-label that is different than the ESI-label used for the selected G-source.

- Note that the ESI-label is used here for "ingress filtering" (at the egress/downstream PE) as opposed to the [RFC7432] "egress filtering" (at the egress/downstream PE) used in the split-horizon procedures. In [RFC7432] the ESI-label indicates what egress ACs must be skipped when forwarding BUM traffic to the egress. In this document, the ESI-label indicates what ingress traffic must be discarded at the downstream PE.

The use of ESI-labels for SFGs forwarded by upstream PEs require some control plane and data plane extensions in the procedures used by [RFC7432] for multi-homing. Upon failure of the selected G-source, the downstream PE will switch over to a different selected G-source, and will therefore change the RPF check for the (*,G) state. The extensions and associated procedures are described in Section 3 and Section 5 respectively.

An operator should use the HS solution if they require a fast fail-over time and the additional bandwidth consumption is acceptable (SFG packets are received multiple times on the downstream PEs). Otherwise the operator should use the WS solution, at the expense of a slower fail-over time in case of a G-source or upstream PE failure. Besides bandwidth efficiency, another advantage of the WS solution is that only the upstream PEs attached to the redundant G-sources for the same SFG need to be upgraded to support the new procedures.

This document does not impose the support of both solutions on a system. If one solution is supported, the support of the other solution is OPTIONAL.

3. BGP EVPN Extensions

This document makes use of the following BGP EVPN extensions:

1. SFG flag in the Multicast Flags Extended Community

   The Single Flow Group (SFG) flag is a new bit requested to IANA out of the registry Multicast Flags Extended Community Flag Values. This new flag is set for S-PMSI A-D routes that carry a (*,G)/(S,G) SFG in the NLRI.

2. ESI Label Extended Community is used in S-PMSI A-D routes

   The HS solution requires the advertisement of one or more ESI Label Extended Communities [RFC7432] that encode the Ethernet Segment Identifier(s) associated to an S-PMSI A-D (*,G)/(S,G) route that advertises the presence of an SFG. Only the ESI Label value in the extended community is relevant to the procedures in this document. The Flags field in the extended community will be advertised as 0x00 and ignored on reception. [RFC7432] specifies that the ESI Label Extended Community is advertised along with the A-D per ES route. This documents extends the use of this extended community so that it can be advertised multiple times (with different ESI values) along with the S-PMSI A-D route.

4. Warm Standby (WS) Solution for Redundant G-Sources

The general procedure is described as follows:

1. Configuration of the upstream PEs

   Upstream PEs (possibly attached to redundant G-sources) need to be configured to know which groups are carrying only flows from redundant G-sources, that is, the SFGs in the tenant domain. They will also be configured to know which local BDs may be attached to a redundant G-source. The SFGs can be configured for any source, E.g., SFG for "*", or for a prefix that contains multiple sources that will issue the same SFG, i.e., "10.0.0.0/30". In the latter case sources 10.0.0.1 and 10.0.0.2 are considered as Redundant G-sources, whereas 10.0.0.10 is not considered a redundant G-source for the same SFG.

   As an example:

   * PE1 is configured to know that G1 is an SFG for any source and redundant G-sources for G1 may be attached to BD1 or BD2.

   * Or PE1 can also be configured to know that G1 is an SFG for the sources contained in 10.0.0.0/30, and those redundant G-sources may be attached to BD1 or BD2.

2. Signaling the location of a G-source for a given SFG

   Upon receiving G-traffic for a configured SFG on a BD, an upstream PE configured to follow this procedure, e.g., PE1:

   * Originates an S-PMSI A-D (*,G)/(S,G) route for the SFG. An (*,G) route is advertised if the SFG is configured for any source, and an (S,G) route is advertised (where the Source can have any length) if the SFG is configured for a prefix.

   * The S-PMSI A-D route is imported by all the PEs attached to the tenant domain. In order to do that, the route will use the SBD-RT (Supplementary Broadcast Domain Route-Target) in addition to the BD-RT of the BD over which the G-traffic is received. The route SHOULD also carry a DF Election Extended Community (EC) and a flag indicating that it conveys an SFG. The DF Election EC and its use is specified in [RFC8584].

- The above S-PMSI A-D route MAY be advertised with or without PMSI Tunnel Attribute (PTA):

  + With no PTA if an I-PMSI or S-PMSI A-D with IR/AR/BIER are to be used.

  + With PTA in any other case.

- The S-PMSI A-D route is triggered by the first packet of the SFG and withdrawn when the flow is not received anymore. Detecting when the G-source is no longer active is a local implementation matter. The use of a timer is RECOMMENDED. The timer is started when the traffic to G1 is not received. Upon expiration of the timer, the PE will withdraw the route

3. Single Forwarder (SF) Election

   If the PE with a local G-source receives one or more S-PMSI A-D routes for the same SFG from a remote PE, it will run a Single Forwarder (SF) Election based on the information encoded in the DF Election EC. Two S-PMSI A-D routes are considered for the same SFG if they are advertised for the same tenant, and their Multicast Source Length, Multicast Source, Multicast Group Length and Multicast Group fields match.

   1. A given DF Alg can only be used if all the PEs running the DF Alg have consistent input. For example, in an OISM network, if the redundant G-sources for an SFG are attached to BDs with different Ethernet Tags, the Default DF Election Alg MUST NOT be used.

   2. In case the there is a mismatch in the DF Election Alg or capabilities advertised by two PEs competing for the SF, the lowest PE IP address (given by the Originator Address in the S- PMSI A-D route) will be used as a tie-breaker.

4. RPF check on the PEs attached to a redundant G-source

   All the PEs with a local G-source for the SFG will add an RPF check to the (*,G)/(S,G) state for the SFG. That RPF check depends on the SF Election result:

   1. The non-SF PEs discard any (*,G)/(S,G) packets for the SFG received over a local AC.

   2. The SF accepts any (*,G)/(S,G) packets for the SFG it receives over one (and only one) local AC.

The solution above provides redundancy for SFGs and it does not require an upgrade of the downstream PEs (PEs where there is certainty that no redundant G-sources are connected). Other G-sources for non-SFGs may exist in the same tenant domain. This document does not change the existing procedures for non-SFG G-sources.

The redundant G-sources can be single-homed or multi-homed to a BD in the tenant domain. Multi-homing does not change the above procedures.

Section 4.1 and Section 4.2 show two examples of the WS solution.

4.1. WS Example in an OISM Network

Figure 4 illustrates an example in which S1 and S2 are redundant G-sources for the SFG (*,G1).

```
                        S1 (Single               S2
                        |   Forwarder)           |
                 (S1,G1)|                 (S2,G1)|
                        |                        |
               PE1      |               PE2      |
               +--------v---+           +--------v---+
        S-PMSI |      +---+ |           |      +---+ | S-PMSI
        (*,G1) |  +---|BD1| |           |  +---|BD2| | (*,G1)
       Pref200 |  |VRF+---+ |           |  |VRF+---+ | Pref100
         |SFG  |+---+  | |  |           |+---+  |    |  SFG|
         | +----|SBD|--+ |  |-----------||SBD|--+    |---+ |
         v |   |+---+    |  |           |+---+       |   | v
           |   +---------|--+           +------------+   |
    SMET   |             |                               | SMET
    (*,G1) |             |   (S1,G1)                     | (*,G1)
           |    +--------+------------------+            |
       ^   |    |                           |            |   ^
       |   |    |                EVPN       |            |   |
       |   |    |                OISM       |            |   |
       |   |    |                           |            |   |
       PE3 |    |           PE4             |            | PE5
       +--------v---+       +------------+  |   +------------+
       |      +---+ |       |      +---+ |  |   |      +---+ |
       |  +---|SBD| |-------|  +---|SBD| |--|---|  +---|SBD| |
       |  |VRF+---+ |       |  |VRF+---+ |  |   |  |VRF+---+ |
       |+---+  |    |       |+---+  |    |  |   |+---+  |    |
       ||BD3|--+    |       ||BD4|--+    |  +--->|BD1|--+    |
       |+---+       |       |+---+       |      |+---+       |
       +------------+       +------------+      +------------+
         |  ^                                     |  ^
         |  | IGMP                                |  | IGMP
         R1 | J(*,G1)                             R3 | J(*,G1)
```

Figure 4: WS Solution for Redundant G-Sources

The WS solution works as follows:

1. Configuration of the upstream PEs, PE1 and PE2

    PE1 and PE2 are configured to know that G1 is an SFG for any source and redundant G-sources for G1 may be attached to BD1 or BD2, respectively.

2. Signaling the location of S1 and S2 for (*,G1)

    Upon receiving (S1,G1) traffic on a local AC, PE1 and PE2 originate S-PMSI A-D (*,G1) routes with the SBD-RT, DF Election

Extended Community (EC) and a flag indicating that it conveys an SFG.

3. Single Forwarder (SF) Election

   Based on the DF Election EC content, PE1 and PE2 elect an SF for (*,G1). Assuming both PEs agree on e.g., Preference based Election as the algorithm to use [DF-PREF], and PE1 has a higher preference, PE1 becomes the SF for (*,G1).

4. RPF check on the PEs attached to a redundant G-source

   A. The non-SF, PE2, discards any (*,G1) packets received over a local AC.

   B. The SF, PE1 accepts (*,G1) packets it receives over one (and only one) local AC.

The end result is that, upon receiving reports for (*,G1) or (S,G1), the downstream PEs (PE3 and PE5) will issue SMET routes and will pull the multicast SFG from PE1, and PE1 only. Upon a failure on S1, the AC connected to S1 or PE1 itself will trigger the S-PMSI A-D (*,G1) withdrawal from PE1 and PE2 will be promoted to SF.

4.2. WS Example in a Single-BD Tenant Network

Figure 5 illustrates an example in which S1 and S2 are redundant G-sources for the SFG (*,G1), however, now all the G-sources and receivers are connected to the same BD1 and there is no SBD.

```
                        S1 (Single               S2
                        |   Forwarder)           |
                 (S1,G1)|                 (S2,G1)|
                        |                        |
               PE1      |              PE2      |
               +--------v---+           +--------v---+
       S-PMSI  |      +---+ |           |      +---+ | S-PMSI
       (*,G1)  |      |BD1| |           |      |BD1| | (*,G1)
       Pref200 |      +---+ |           |      +---+ | Pref100
        |SFG   |         |  |           |            |  SFG|
        |  +---|         |  |-----------|            |---+ |
        v  |   |         |  |           |            |   | v
           |   +---------|--+           +------------+   |
    SMET   |             |                               | SMET
    (*,G1) |             |     (S1,G1)                   | (*,G1)
           |    +--------+------------------------+      |
       ^   |    |                                 |      |   ^
       |   |    |                EVPN             |      |   |
       |   |    |                                 |      |   |
       |   |    |                                 |      |   |
       PE3 |    |           PE4                   |      | PE5
       +--------v---+       +------------+      +-|----------+
       |      +---+ |       |      +---+ |      | |    +---+ |
       |      |BD1| |-------|      |BD1| |------| +--->|BD1| |
       |      +---+ |       |      +---+ |      |      +---+ |
       |            |       |            |      |            |
       |            |       |            |      |            |
       +------------+       +------------+      +------------+
         |  ^                                     |  ^
         |  | IGMP                                |  | IGMP
         R1 | J(*,G1)                             R3 | J(*,G1)
```

Figure 5: WS Solution for Redundant G-Sources in the same BD

The same procedure as in Section 4.1 is valid here, being this a sub-case of the one in Section 4.1. Upon receiving traffic for the SFG G1, PE1 and PE2 advertise the S-PMSI A-D routes with BD1-RT only, since there is no SBD.

5. Hot Standby (HS) Solution for Redundant G-Sources

If fast-failover is required upon the failure of a G-source or PE attached to the G-source and the extra bandwidth consumption in the tenant network is not an issue, the HS solution should be used. The procedure is as follows:

1. Configuration of the PEs

As in the WS case, the upstream PEs where redundant G-sources may exist need to be configured to know which groups (for any source or a prefix containing the intended sources) are carrying only flows from redundant G-sources, that is, the SFGs in the tenant domain.

In addition (and this is not done in WS mode), the individual redundant G-sources for an SFG need to be associated with an Ethernet Segment (ES) on the upstream PEs. This is irrespective of the redundant G-source being multi-homed or single-homed. Even for single-homed redundant G-sources the HS procedure relies on the ESI labels for the RPF check on downstream PEs. The term "S-ESI" is used in this document to refer to an ESI associated to a redundant G-source.

Contrary to what is specified in the WS method (that is transparent to the downstream PEs), the support of the HS procedure is required not only on the upstream PEs but also on all downstream PEs connected to the receivers in the tenant network. The downstream PEs do not need to be configured to know the connected SFGs or their ESIs, since they get that information from the upstream PEs. The downstream PEs will locally select an ESI for a given SFG, and will program an RPF check to the (*,G)/(S,G) state for the SFG that will discard (*,G)/(S,G) packets from the rest of the ESIs. The selection of the ESI for the SFG is based on local policy.

2. Signaling the location of a G-source for a given SFG and its association to the local ESIs

   Based on the configuration in step 1, an upstream PE configured to follow the HS procedures:

   A. Advertises an S-PMSI A-D (*,G)/(S,G) route per each configured SFG. These routes need to be imported by all the PEs of the tenant domain, therefore they will carry the BD-RT and SBD-RT (if the SBD exists). The route also carries the ESI Label Extended Communities needed to convey all the S-ESIs associated to the SFG in the PE.

   B. The S-PMSI A-D route will convey a PTA in the same cases as in the WS procedure.

   C. The S-PMSI A-D (*,G)/(S,G) route is triggered by the configuration of the SFG and not by the reception of G-traffic.

3. Distribution of DCB (Domain-wide Common Block) ESI-labels and G-source ES routes

   An upstream PE advertises the corresponding ES, A-D per EVI and A-D per ES routes for the local S-ESIs.

   A. ES routes are used for regular DF Election for the S-ES. This document does not introduce any change in the procedures related to the ES routes.

   B. The A-D per EVI and A-D per ES routes MUST include the SBD-RT since they have to be imported by all the PEs in the tenant domain.

   C. The A-D per ES routes convey the S-ESI labels that the downstream PEs use to add the RPF check for the (*,G)/(S,G) associated to the SFGs. This RPF check requires that all the packets for a given G-source are received with the same S-ESI label value on the downstream PEs. For example, if two redundant G-sources are multi-homed to PE1 and PE2 via S-ES-1 and S-ES-2, PE1 and PE2 MUST allocate the same ESI label "Lx" for S-ES-1 and they MUST allocate the same ESI label "Ly" for S-ES-2. In addition, Lx and Ly MUST be different. These ESI labels are Domain-wide Common Block (DCB) labels and follow the allocation procedures in [I-D.zzhang-bess-mvpn-evpn-aggregation-label].

4. Processing of A-D per ES/EVI routes and RPF check on the downstream PEs

   The A-D per ES/EVI routes are received and imported in all the PEs in the tenant domain. The processing of the A-D per ES/EVI routes on a given PE depends on its configuration:

   A. The PEs attached to the same BD of the BD-RT that is included in the A-D per ES/EVI routes will process the routes as in [RFC7432] and [RFC8584]. If the receiving PE is attached to the same ES as indicated in the route, [RFC7432] split-horizon procedures will be followed and the DF Election candidate list may be modified as in [RFC8584] if the ES supports the AC-DF capability.

   B. The PEs that are not attached to the BD-RT but are attached to the SBD of the received SBD-RT, will import the A-D per ES/EVI routes and use them for redundant G-source mass withdrawal, as explained later.

C. Upon importing A-D per ES routes corresponding to different S-ESes, a PE MUST select a primary S-ES and add an RPF check to the (*,G)/(S,G) state in the BD or SBD. This RPF check will discard all ingress packets to (*,G)/(S,G) that are not received with the ESI-label of the primary S-ES. The selection of the primary S-ES is a matter of local policy.

5. G-traffic forwarding for redundant G-sources and fault detection

Assuming there is (*,G) or (S,G) state for the SFG with OIF (Ouput Interface) list entries associated to remote EVPN PEs, upon receiving G-traffic on a S-ES, the upstream PE will add a S-ESI label at the bottom of the stack before forwarding the traffic to the remote EVPN PEs. This label is allocated from a DCB as described in step 3. If P2MP or BIER PMSIs are used, this is not adding any new data path procedures on the upstream PEs (except that the ESI-label is allocated from a DCB). However, if IR/AR are used, this document extends the [RFC7432] procedures by pushing the S-ESI labels not only on packets sent to the PEs that shared the ES but also to the rest of the PEs in the tenant domain. This allows the downstream PEs to receive all the multicast packets from the redundant G-sources with a S-ESI label (irrespective of the PMSI type and the local ESes), and discard any packet that conveys a S-ESI label different from the primary S-ESI label (that is, the label associated to the selected primary S-ES), as discussed in step 4.

If the last A-D per EVI or the last A-D per ES route for the primary S-ES is withdrawn, the downstream PE will immediately select a new primary S-ES and will change the RPF check. Note that if the S-ES is re-used for multiple tenant domains by the upstream PEs, the withdrawal of all the A-D per-ES routes for a S-ES provides a mass withdrawal capability that makes a downstream PE to change the RPF check in all the tenant domains using the same S-ES.

The withdrawal of the last S-PMSI A-D route for a given (*,G)/(S,G) that represents a SFG SHOULD make the downstream PE remove the S-ESI label based RPF check on (*,G)/(S,G).

5.1. Use of BFD in the HS Solution

In addition to using the state of the A-D per EVI, A-D per ES or S-PMSI A-D routes to modify the RPF check on (*,G)/(S,G) as discussed in Section 5, Bidirectional Forwarding Detection (BFD) protocol MAY be used to find the status of the multipoint tunnels used to forward the SFG from the redundant G-sources.

The BGP-BFD Attribute is advertised along with the S-PMSI A-D or IMET routes (depending on whether I-PMSI or S-PMSI trees are used) and the procedures described in [EVPN-BFD] are used to bootstrap multipoint BFD sessions on the downstream PEs.

5.2. HS Example in an OISM Network

Figure 6 illustrates the HS model in an OISM network. Consider S1 and S2 are redundant G-sources for the SFG (*,G1) in BD1 (any source using G1 is assumed to transmit an SFG). S1 and S2 are (all-active) multi-homed to upstream PEs, PE1 and PE2. The receivers are attached to downstream PEs, PE3 and PE5, in BD3 and BD1, respectively. S1 and S2 are assumed to be connected by a LAG to the multi-homing PEs, and the multicast traffic can use the link to either upstream PE. The diagram illustrates how S1 sends the G-traffic to PE1 and PE1 forwards to the remote interested downstream PEs, whereas S2 sends to PE2 and PE2 forwards further. In this HS model, the interested downstream PEs will get duplicate G-traffic from the two G-sources for the same SFG. While the diagram shows that the two flows are forwarded by different upstream PEs, the all-active multi-homing procedures may cause that the two flows come from the same upstream PE. Therefore, finding out the upstream PE for the flow is not enough for the downstream PEs to program the required RPF check to avoid duplicate packets on the receiver.

```
                        S1(ESI-1)                S2(ESI-2)
                        |                        |
                        | +----------------------+
                 (S1,G1)| |               (S2,G1)|
                        +----------------------+ |
               PE1      | |             PE2    | |
               +--------v---+           +--------v---+
               |      +---+ |           |      +---+ |  S-PMSI
    S-PMSI     |  +---|BD1| |           |  +---|BD1| |  (*,G1)
    (*,G1)     |  |VRF+---+ |           |  |VRF+---+ |   SFG
     SFG       |+---+  | |  |           |+---+  | |  |   ESI1,2
    ESI1,2 +---||SBD|--+ |  |-----------||SBD|--+ |  |---+  |
       |   |   |+---+    |  |   EVPN    |+---+    |  |   |  v
       v   |   +---------|--+   OISM    +---------|--+   |
           |             |                        |      |
           |             |   (S1,G1)              |      |
    SMET   |   +---------+------------------+     |      | SMET
    (*,G1) |   |                            |     |      | (*,G1)
       ^   |   | +----------------------------+---+      |   ^
       |   |   | |             (S2,G1)      | |          |   |
       |   |   | |                          | |          |   |
       PE3 |   | |          PE4             | |          | PE5
       +-------v-v--+       +------------+  | | +------------+
       |      +---+ |       |      +---+ |  | | |      +---+ |
       |  +---|SBD| +-------|  +---|SBD| |--|-|-|  +---|SBD| |
       |  |VRF+---+ |       |  |VRF+---+ |  | | |  |VRF+---+ |
       |+---+  |    |       |+---+  |    |  | | |+---+  |    |
       ||BD3|--+    |       ||BD4|--+    |  | +->|BD1|--+    |
       |+---+       |       |+---+       |  +--->+---+       |
       +------------+       +------------+      +------------+
         |  ^                                     |  ^
         |  | IGMP                                |  | IGMP
         R1 | J(*,G1)                             R3 | J(*,G1)
```

Figure 6: HS Solution for Multi-homed Redundant G-Sources in OISM

In this scenario, the HS solution works as follows:

1. Configuration of the upstream PEs, PE1 and PE2

   PE1 and PE2 are configured to know that G1 is an SFG for any source (a source prefix length could have been configured instead) and the redundant G-sources for G1 use S-ESIs ESI-1 and ESI-2 respectively. Both ESes are configured in both PEs and the ESI value can be configured or auto-derived. The ESI-label values are allocated from a DCB [I-D.zzhang-bess-mvpn-evpn-aggregation-label] and are configured either locally or by a centralized controller. We assume ESI-1 is configured to use ESI-label-1 and ESI-2 to use ESI-label-2.

The downstream PEs, PE3, PE4 and PE5 are configured to support HS mode and select the G-source with e.g., lowest ESI value.

2. PE1 and PE2 advertise S-PMSI A-D (*,G1) and ES/A-D per ES/EVI routes

   Based on the configuration of step 1, PE1 and PE2 advertise an S-PMSI A-D (*,G1) route each. The route from each of the two PEs will include TWO ESI Label Extended Communities with ESI-1 and ESI-2 respectively, as well as BD1-RT plus SBD-RT and a flag that indicates that (*,G1) is an SFG.

   In addition, PE1 and PE2 advertise ES and A-D per ES/EVI routes for ESI-1 and ESI-2. The A-D per ES and per EVI routes will include the SBD-RT so that they can be imported by the downstream PEs that are not attached to BD1, e.g., PE3 and PE4. The A-D per ES routes will convey ESI-label-1 for ESI-1 (on both PEs) and ESI-label-2 for ESI-2 (also on both PEs).

3. Processing of A-D per ES/EVI routes and RPF check

   PE1 and PE2 received each other’s ES and A-D per ES/EVI routes. Regular [RFC7432] [RFC8584] procedures will be followed for DF Election and programming of the ESI-labels for egress split-horizon filtering. PE3/PE4 import the A-D per ES/EVI routes in the SBD. Since PE3 has created a (*,G1) state based on local interest, PE3 will add an RPF check to (*,G1) so that packets coming with ESI-label-2 are discarded (lowest ESI value is assumed to give the primary S-ES).

4. G-traffic forwarding and fault detection

   PE1 receives G-traffic (S1,G1) on ES-1 that is forwarded within the context of BD1. Irrespective of the tunnel type, PE1 pushes ESI-label-1 at the bottom of the stack and the traffic gets to PE3 and PE5 with the mentioned ESI-label (PE4 has no local interested receivers). The G-traffic with ESI-label-1 passes the RPF check and it is forwarded to R1. In the same way, PE2 sends (S2,G1) with ESI-label-2, but this G-traffic does not pass the RPF check and gets discarded at PE3/PE5.

   If the link from S1 to PE1 fails, S1 will forward the (S1,G1) traffic to PE2 instead. PE1 withdraws the ES and A-D routes for ESI-1. Now both flows will be originated by PE2, however the RPF checks don’t change in PE3/PE5.

If subsequently, the link from S1 to PE2 fails, PE2 also withdraws the ES and A-D routes for ESI-1. Since PE3 and PE5 have no longer A-D per ES/EVI routes for ESI-1, they immediately change the RPF check so that packets with ESI-label-2 are now accepted.

Figure 7 illustrates a scenario where S1 and S2 are single-homed to PE1 and PE2 respectively. This scenario is a sub-case of the one in Figure 6. Now ES-1 only exists in PE1, hence only PE1 advertises the A-D per ES/EVI routes for ESI-1. Similarly, ES-2 only exists in PE2 and PE2 is the only PE advertising A-D routes for ESI-2. The same procedures as in Figure 6 applies to this use-case.

```
                        S1(ESI-1)                S2(ESI-2)
                        |                        |
                 (S1,G1)|                 (S2,G1)|
                        |                        |
               PE1      |               PE2      |
               +--------v---+           +--------v---+
               |      +---+ |           |      +---+ |  S-PMSI
    S-PMSI     |  +---|BD1| |           |  +---|BD2| |  (*,G1)
    (*,G1)     |  |VRF+---+ |           |  |VRF+---+ |   SFG
     SFG       |+---+  | |  |           |+---+  | |  |   ESI2
     ESI1  +---||SBD|--+ |  |-----------||SBD|--+ |  |---+  |
       |   |   |+---+    |  |   EVPN    |+---+    |  |   |  v
       v   |   +---------|--+   OISM    +---------|--+   |
           |             |                        |      |
           |             |   (S1,G1)              |      |
    SMET   |   +---------+------------------+     |      | SMET
    (*,G1) |   |                            |     |      | (*,G1)
       ^   |   | +--------------------------------+----+ |   ^
       |   |   | |             (S2,G1)      |          | |   |
       |   |   | |                          |          | |   |
       PE3 |   | |          PE4             |          | | PE5
       +-------v-v--+       +------------+  |   +------v-----+
       |      +---+ |       |      +---+ |  |   |      +---+ |
       |  +---|SBD| |-------|  +---|SBD| |--|---|  +---|SBD| |
       |  |VRF+---+ |       |  |VRF+---+ |  |   |  |VRF+---+ |
       |+---+  |    |       |+---+  |    |  |   |+---+  |    |
       ||BD3|--+    |       ||BD4|--+    |  +--->|BD1|--+    |
       |+---+       |       |+---+       |      |+---+       |
       +------------+       +------------+      +------------+
         |  ^                                     |  ^
         |  | IGMP                                |  | IGMP
         R1 | J(*,G1)                             R3 | J(*,G1)
```

Figure 7: HS Solution for single-homed Redundant G-Sources in OISM

5.3. HS Example in a Single-BD Tenant Network

Irrespective of the redundant G-sources being multi-homed or single-homed, if the tenant network has only one BD, e.g., BD1, the procedures of Section 5.2 still apply, only that routes do not include any SBD-RT and all the procedures apply to BD1 only.

6. Security Considerations

The same Security Considerations described in [I-D.ietf-bess-evpn-irb-mcast] are valid for this document.

From a security perspective, out of the two methods described in this document, the WS method is considered lighter in terms of control plane and therefore its impact is low on the processing capabilities of the PEs. The HS method adds more burden on the control plane of all the PEs of the tenant with sources and receivers.

7. IANA Considerations

IANA is requested to allocate a Bit in the Multicast Flags Extended Community to indicate that a given (*,G) or (S,G) in an S-PMSI A-D route is associated with an SFG.

8. References

8.1. 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>.

[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>.

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Patel, K., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-05 (work in progress), April 2020.

[I-D.ietf-bess-evpn-irb-mcast]
Lin, W., Zhang, Z., Drake, J., Rosen, E., Rabadan, J., and A. Sajassi, "EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding", draft-ietf-bess-evpn-irb-mcast-05 (work in progress), October 2020.

[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>.

[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.zzhang-bess-mvpn-evpn-aggregation-label]
Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-zzhang-bess-mvpn-evpn-aggregation-label-01 (work in progress), April 2018.

8.2. Informative References

[EVPN-RT5]
Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", internet-draft ietf-bess-evpn-prefix-advertisement-11.txt, May 2018.

[EVPN-BUM]
Zhang, Z., Lin, W., Rabadan, J., and K. Patel, "Updates on EVPN BUM Procedures", internet-draft ietf-bess-evpn-bum-procedure-updates-06, June 2019.

[DF-PREF] Rabadan, J., Sathappan, S., Przygienda, T., Lin, W., Drake, J., Sajassi, A., and S. Mohanty, "Preference-based EVPN DF Election", internet-draft ietf-bess-evpn-pref-df-04.txt, June 2019.

[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>.

[EVPN-BFD]
Govindan, V., Mallik, M., Sajassi, A., and G. Mirsky, "Fault Management for EVPN networks", internet-draft ietf-bess-evpn-bfd-01.txt, October 2020.

Appendix A. Acknowledgments

The authors would like to thank Mankamana Mishra and Ali Sajassi for their review and valuable comments.

Appendix B. Contributors

Authors' Addresses

Jorge Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

Jayant Kotalwar
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: jayant.kotalwar@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA

Email: senthil.sathappan@nokia.com

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Wen Lin
Juniper Networks

Email: wlin@juniper.net

Eric C. Rosen
Individual

Email: erosen52@gmail.com

tsvwg
Internet-Draft
Intended status: Standards Track
Expires: May 5, 2021

Z. Zhang
R. Bonica
K. Kompella
Juniper Networks

November 01, 2020

Generic Transport Functions
draft-zzhang-tsvwg-generic-transport-functions-00

Abstract

Some functionalities (e.g. fragmentation/reassembly and Encapsulating Security Payload) provided by IPv6 can be viewed as independent of IPv6 or even IP entirely. This document proposes to provide those functionalities at different layers (e.g., MPLS, BIER or even Ethernet) independent of IP.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Consider an operator providing Ethernet services such as pseudowires, VPLS or EVPN. The Ethernet frames that a Provider Edge (PE) device receives from a Customer Edge (CE) device may have a larger size than the PE-PE path MTU (pMTU) in the provider network. This could be because

1. the provider network is built upon virtual connections (e.g. pseudowires) provided by another infrastructure provider, or

2. the customer network uses jumbo frames while the provider network does not, or

3. the provider-side overhead for transporting customers packets across the network pushes past the pMTU.

In any case, the provider simply cannot require its customers to change their MTU.

To get those large frames across the provider network, currently the only workaround is to encapsulate the frames in IP (with or without GRE) and then fragment the IP packets. Even if MPLS is used for service delimiting, IP is used for transporation (MPLS over IP/GRE). This may not be desirable in certain deployment scenarios, where MPLS is the preferred transport or IP encapsulation overhead is deemed excessive.

IPv6 fragmentation and reassembly are based on the IPv6 Fragmentation header below [RFC8200]:

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Next Header  |   Reserved    |      Fragment Offset    |Res|M|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Identification                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: IPv6 Fragmentation Header

This document proposes reusing this header in non-IP contexts, since the fragmentation/reassembly function is actually independent of IPv6 except the following aspects:

- The fragment header is identified as such by the "previous" header.

- The "Next Header" value is from the "Internet Protocol Numbers" registry.

- The "Identification" value is unique in the (source, destination) context provided by the IPv6 header

The "Identification" field, in conjunction with the IPv6 source and destination identifies fragments of the original packet, for the purpose of reassembly.

Therefore, the fragmentation/reassembly function can be applied at other layers as long as a) the fragment header is identified as such; and b) the context for packet identification is provided. Examples of such layers include MPLS, BIER, and Ethernet (if IEEE determines it is so desired).

For the layers where the IETF is concerned, the "Next Header" value will still be from the "Internet Protocol Numbers" registry when the function is applied at non-IP layers.

For the same consideration, the IP Encapsulating Security Payload (ESP) [RFC4303] could also be applied at other layers if ESP is desired there. For example, if for whatever reason the Ethernet service provider wants to provide ESP between its PEs, it could do so without requiring IP encapsulation if ESP is applied at non-IP layers.

The possibility of applying some other IP functions (e.g. Authentication Header [RFC4302]) is for further study.

2. Specifications

2.1. Generic Fragmentation Header

For generic fragmentation/reassembly functionality independent of IP, the following Generic Fragmentation Header (GFH) is defined:

```
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  Next Header  | Header Length |      Fragment Offset    |R|S|M|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                         Identification                        |
   |                           (variable)                          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: Generic Fragmentation Header

The "Next Header", "Fragment Offset" and "M" flag bit fields are as in the IPv6 Fragmentation Header.

Header Length: the number of octets of the entire header.

R: The "R" flag bit is reserved. It MUST be 0 on transmitting and ignored on receiving.

Identification: at least 4-octet long.

S: If the "S" flag bit is clear, the context for the Identification field is provided by the outer header, and only the source-identifying information in the outer header is used. If the "S" flag bit is set, the variable Identification field encodes both source-identifying information (e.g. the IP address of the node adding the GFH) and an identification number unique within that source.

The outer header MUST identify that a Generic Fragmentation Header follows and MAY carry source-identifying information.

If the outer header is BIER, a TBD value for the "proto" field in the BIER header identifies that a GFH follows. If the "S" flag bit is clear, the "BFIR-id" field in the BIER header provides the context for the "Identification" field.

If the outer header is MPLS, the "S" flag bit MAY be clear if the the label preceeding the GFH identifies the sending BFR in addition to indicating that a GFH follows (see Section 2.2).

2.2. MPLS Signaling

When GFH is used with MPLS, the preceeding label needs to indicate that a GFH follows, and optionally identify the node that does the fragmentation. The label can be signaled via BGP or IGP as sepcified below.

2.2.1. BGP Signaling

This document defines a new transitive BGP "GFH Labels" attribute, very similar to the "PE Distinguisher Labels" attribute defined in [RFC6514] (and the text below is adapted from Section 8 of [RFC6514]):

```
     +---------------------------------+
     |     Node Address                |
     +---------------------------------+
     |     Label (3 octets)            |
     +---------------------------------+
     .......
     +---------------------------------+
     |     Node Address                |
     +---------------------------------+
     |     Label (3 octets)            |
     +---------------------------------+
```

The Label field contains an MPLS label encoded as 3 octets, where the high-order 20 bits contain the label value. The Node Address MAY be 0, meaning that the following label only indicates a GFH follows when the label is used in the label stack of a data packet.

The Node Address MAY also be a unicast address, indicating that the following label when used in the label stack of a data packet will both indicate that a GFH follows and identify the sending node.

If a node supports GFH with MPLS, it attaches the attribute in the BGP routes for its local addresses. A border router SHOULD remove the attribute if no node beyond the border will use GFH with MPLS to send traffic to the corresponding addresses.

A router that supports the attribute considers this attribute to be malformed if the Node Address field does not contain a unicast address or 0. The attribute is also considered to be malformed if: (a) the Node Address field is expected to be an IPv4 address, and the length of the attribute is not a multiple of 7 or (b) the Node Address field is expected to be an IPv6 address, and the length of the attribute is not a multiple of 19. The Address Family Indicator (AFI) of the BGP route that the attribute is attached to provides the information on whether the Node Address field contains an IPv4 or IPv6 address. Each of the Node Addresses in the attribute MUST be of the same address family as the route that is carrying the attribute.

2.2.2. IGP Signaling

This document defines an OSPFv2 "GFH Labels" sub-TLV of OSPFv2 Extended Prefix TLV [RFC7684], with the value part being the same as BGP "GFH Labels" attribute above. If an OSPFv2 router surports GFH with MPLS, it includes the GFH Labels sub-TLV in the Extended Prefix TLV that is attached to its local addresses advertised in its OSPFv2 Extended Prefix Opaque LSA.

Similary, This document defines an OSPFv3 "GFH Labels" sub-TLV of OSPFv3 Intra/Inter-Area-Prefix TLVs [RFC8362], with the value part being the same as BGP "GFH Labels" attribute above. If an OSPFv3 router surports GFH with MPLS, it includes the GFH Labels sub-TLV in the Intra-Area-Prefix TLV for its local addresses.

This document also defines an ISIS "GFH Labels" sub-TLV of ISIS prefix-reachability TLV [RFC5120] [RFC5305] [RFC5308], with the value part being the same as BGP "GFH Labels" attribute above. If an ISIS router surports GFH with MPLS, it includes the sub-TLV to the prefix-reachability TLV for its local addresses.

For both OSPF and ISIS, when advertising a prefix from one area/level to another, if there is a "GFH Labels TLV" attached in the source area/level, the TLV SHOULD be attached in the target area/level and the prefix SHOULD NOT be summarized.

2.3. Generic ESP/Authentication Header

To be specified in future revisions.

3. Security Considerations

To be provided.

4. IANA Considerations

This document makes the following IANA requests:

- A new BGP Attribute type for "GFH Labels" from the BGP Path Attributes registry

- A new OSPFv2 sub-TLV type for "GFH Labels" from the OSPFv2 Extended Prefix TLV Sub-TLVs registry

- A new OSPFv3 sub-TLV type for "GFH Labels" from the OSPFv3 Extended-LSA sub-TLV registry

- A new BIER Next Protocol Identifier value for GFH from BIER Next Protocol Identifiers registry

5. Acknowledgements

6. References

6.1. Normative References

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <https://www.rfc-editor.org/info/rfc4303>.

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <https://www.rfc-editor.org/info/rfc5120>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, DOI 10.17487/RFC5308, October 2008, <https://www.rfc-editor.org/info/rfc5308>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <https://www.rfc-editor.org/info/rfc7684>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC8362] Lindem, A., Roy, A., Goethals, D., Reddy Vallem, V., and F. Baker, "OSPFv3 Link State Advertisement (LSA) Extensibility", RFC 8362, DOI 10.17487/RFC8362, April 2018, <https://www.rfc-editor.org/info/rfc8362>.

6.2. Informative References

[RFC4302] Kent, S., "IP Authentication Header", RFC 4302, DOI 10.17487/RFC4302, December 2005, <https://www.rfc-editor.org/info/rfc4302>.

[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>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks
1133 Innovation Way
Sunnyvale 94089
USA

Phone: +1 408 745 2000
Email: zzhang@juniper.net

Ron Bonica
Juniper Networks
1133 Innovation Way
Sunnyvale 94089
USA

Phone: +1 408 745 2000
Email: rbonica@juniper.net

Kireeti Kompella
Juniper Networks
1133 Innovation Way
Sunnyvale 94089
USA

Phone: +1 408 745 2000
Email: kireeti@juniper.net