

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 6, 2021

H. Bidgoli, Ed.
Nokia
V. Voyer
Bell Canada
A. Stone
Nokia
R. Parekh
Cisco System
S. Krier
A. Venkateswaran
Cisco System, Inc.
June 04, 2021

Advertising p2mp policies in BGP
draft-hb-idr-sr-p2mp-policy-02

Abstract

SR P2MP policies are set of policies that enable architecture for P2MP service delivery.

A P2MP policy consists of candidate paths that connects the Root of the Tree to a set of Leaves. The P2MP policy is composed of replication segments. A replication segment is a forwarding instruction for a candidate path which is downloaded to the Root, transit nodes and the leaves.

This document specifies a new BGP SAFI with a new NLRI in order to advertise P2MP policy from a controller to a set of nodes.

This document introduces two new route types within this NLRI, one for P2MP policy and its candidate paths that need to be programmed on the Root node and another for the replication segment and forwarding instructions that needs to be programmed on the Root, and optionally on Transit and Leaf nodes.

It should be noted that this document does not specify how the Root and the Leaves are discovered on the controller, it only describes how the P2MP Policy and Replication Segments are programmed from the controller to the nodes.

Status of This Memo

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

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

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

This Internet-Draft will expire on December 6, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions used in this document	4
3. P2MP Policy and Replication Segment Encoding	4
3.1. P2MP Policy SAFI and NLRI	4
3.1.1. P2MP Policy Route - Route Type TBD1	5
3.1.2. Replication segment Route - Route type TBD 2	6
3.2. Tunnel Encapsulation Attribute	7
3.2.1. SR P2MP policy encoding	7
3.2.2. Replication segment encoding	8
3.3. P2MP Policy Sub-TLVs	9
3.3.1. preference Sub-TLV	9
3.3.2. leaf-list Sub-TLV	9
3.3.3. path-instance Sub-TLV	10
3.3.3.1. active instance-id Sub-TLV	10
3.3.3.2. instance-id Sub-TLV	11
3.4. Replication segment Sub-TLVs	11
3.4.1. Replication SID (Binding SID)	11
3.4.2. Down stream nodes Sub-TLV	12
3.4.3. Segment list Sub-TLV	13

3.4.4. Weight sub-tlv	13
3.4.5. Protection sub-tlv	13
3.4.6. Segment Sub-TLV	14
4. P2MP Policy Operation	15
4.1. Configuration and advertisement of P2MP Policies	15
4.2. Reception of an P2MP Policy NLRI	15
4.3. Global Optimization for P2MP LSPs	16
5. IANA Consideration	16
6. Security Considerations	16
7. Acknowledgments	16
8. References	16
8.1. Normative References	16
8.2. Informative References	17
Authors' Addresses	17

1. Introduction

The draft [draft-ietf-pim-sr-p2mp-policy] defines a variant of the SR Policy [draft-ietf-spring-segment-routing-policy] for constructing a P2MP segment to support multicast service delivery.

A Point-to-Multipoint (P2MP) Policy contains a set of candidate paths and identifies a Root node and a set of Leaf nodes in a Segment Routing Domain. The draft also defines a Replication segment, which corresponds to the state of a P2MP segment on a particular node. The Replication segment is the forwarding instruction for a P2MP LSP at the Root, Transit and Leaf nodes.

For a P2MP segment, a controller may be used to compute a tree from a Root node to a set of Leaf nodes, optionally via a set of replication nodes. A packet is replicated at the root node and optionally on Replication nodes towards each Leaf node.

We define two types of a P2MP segment: Ingress Replication (aka Spray) and Downstream Replication (aka TreeSID).

A Point-to-Multipoint service delivery could be via Ingress Replication (aka Spray in some SR context), i.e., the root unicasts individual copies of traffic to each leaf. The corresponding P2MP segment consists of replication segments only for the root and the leaves.

A Point-to-Multipoint service delivery could also be via Downstream Replication (aka TreeSID in some SR context), i.e., the root and some downstream replication nodes replicate the traffic along the way as it traverses closer to the leaves.

It should be noted that two replication nodes can be connected directly, or they can be connected via unicast SR segment or a segment list.

The leaves and the root of a p2mp policy can be discovered via the multicast protocols or procedures like NG-MVPN [RFC6513] or manually configured on the PCC (CLI) or the PCE.

Based on the discovered root and leaves, the controller builds a P2MP policy and advertise it to the head-end router (i.e. the root of the P2MP Tree). The advertisement uses BGP extensions defined in this document. The controller also calculates the tree path and builds the replication segments on each segment of the tree, Root, Transit and Leaf nodes and downloads the forwarding instructions to the nodes via BGP extensions defined in this document.

SR p2mp policy is a variant of the SR policy and as such it reuses the concept of a candidate path. This draft reuses some of the concepts and TLVs mentioned in [draft-ietf-idr-segment-routing-te-policy]

A candidate path within the P2MP policy can contain multiple path-instances. A path-instance can be viewed as a P2MP LSP. For candidate path global optimization purposes, two or more path-instances can be used to execute make before break procedures.

Each path-instance is a P2MP LSP as such each path-instance needs a set of replication segments to construct its forwarding instructions.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. P2MP Policy and Replication Segment Encoding

3.1. P2MP Policy SAFI and NLRI

This document defines a new BGP NLRI, called the P2MP-POLICY NLRI.

A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA). The following is the format of the P2MP-POLICY NLRI:

	route type		1 octet
	length		1 octet
	route type specific (variable)		

- o The Route type field defines the encoding of the rest of the P2MP-POLICY NLRI.
- o The length field indicates the length in octets of the route type specific data, excluding route type and length
- o This document defines the following route types:
 - * P2MP Policy route: TBD1
 - * Replication Segment: TBD2

The NLRI containing the SR P2MP Policy is carried in a BGP UPDATE message [RFC4271] using BGP multiprotocol extensions [RFC4760] with an AFI of 1 or 2 (IPv4 or IPv6) and with a SAFI of "TBD" (assigned by IANA from the "Subsequent Address Family Identifiers (SAFI) Parameters" registry).

All other recommendations of [draft-ietf-idr-segment-routing-te-policy] section SR Policy SAFI and NLRI, should be taken into account for P2MP policy.

3.1.1. P2MP Policy Route - Route Type TBD1

~	Root-ID	~	4 or 16 octets (ipv4/ipv6)
	Tree-ID		4 octets
	Distinguisher		4 octets

- o Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree, based on AFI.
- o Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end (root)router.
- o Distinguisher: 4-octets value uniquely identifying the policy in the context of <Tree-ID, Originating Router's IP> tuple. The

distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.

3.1.2. Replication segment Route - Route type TBD 2

There can be two type of replication segment, shared and non-shared. A shared replication segment can carry multiple MVPN services or it can be used for Facility Fast reroute protecting multiple P2MP trees. A non-shared tree is used when the label field of the PMSI Tunnel Attribute (PTA) is set to 0 as per [draft-ietf-bess-mvpn-evpn-sr-p2mp]. The following route type can be encoded as per [draft-ietf-idr-segment-routing-te-policy] for shared and non-shared replication segment.

+-----+ ~	Root-ID	+-----+ ~	4 or 16 octets (ipv4/ipv6)
+-----+ 	Tree-ID	+-----+ 	4 octets
+-----+ 	instance-ID	+-----+ 	2 octets
+-----+ 	Distinguisher	+-----+ 	4 octets
+-----+ 	Node-ID	+-----+ 	2 octets
+-----+		+-----+	

- o Root-ID: IPv4/IPv6 address of the head-end (Root) of the p2mp tree based on AFI.
- o Tree-ID: a unique 4 octets identifier of the p2mp tree on the head- end router (Root)
- o instance-id, identifies the path-instance with in the p2mp-policy. Each candidate path can have one, two or more path-instance. Path-instance is used for global optimization of the candidate path via make before break procedures. Instance-ID can be used
- o Distinguisher: 4-octets value uniquely identifying the policy in the context of <Root-ID, Tree-ID> tuple. The distinguisher has no semantic value and is solely used by the SR P2MP Policy originator to make unique (from an NLRI perspective) multiple occurrences of the same SR P2MP Policy.
- o Node-ID: Node's IPv4/IPv6 address

3.2. Tunnel Encapsulation Attribute

The content of this new NLRI is encoded in the tunnel Encapsulation Attribute originally defined in [ietf-idr-tunnel-encaps] using two new Tunnel-Type TLV (codepoint is TBD, assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry) one for P2MP Policy and another for Replication segment.

3.2.1. SR P2MP policy encoding

SR P2MP Policy SAFI NLRI: <route-type p2mp-policy>
Attributes:

 Tunnel Encaps Attribute (23)
 Tunnel Type: (TBD, P2MP-Policy)
 Preference
 Policy Name
 Policy Candidate Path Name
 leaf-list (optional)
 remote-end point
 remote-end point
 ...
 path-instance
 active-instance-id
 instance-id
 instance-id
 ...

- o Relevant only at the Root.
- o SR P2MP-POLICY NLRI and P2MP Policy route type.
- o Tunnel Encapsulation Attribute is defined in [ietf-idr-tunnel-encaps]
- o Tunnel-Type is set to P2MP-Policy Tunnel-Type TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).
- o Policy Name, Policy Candidate Path Name are defined in [draft-ietf-idr-segment-routing-te-policy]
- o Preference, leaf-list, remote-end point and path- instance, instance-ids are defined in this document.
- o Additional sub-TLVs may be defined in the future.

3.2.2. Replication segment encoding

replication segment SAFI NLRI: <route-type non-shared/shared
tree replication-segment>

Attributes:

Tunnel Encaps Attribute (23)

Tunnel Type: (TBD Replication-Segment)

replication-sid (equivalent to Binding Sid)

SRv6 replication-sid (equivalent to SRv6 Binding SID)

downstream-nodes (can be protection enabled via a flag)

segment-list

weight (optional)

g is enabled for downstream-nodes) protection (optional, must be present when protection fla

segment

segment

...

segment-list

weight (optional)

g is enabled for downstream-nodes) protection (optional, must be present when protection fla

segment

segment

...

segment-list (protection segment list)

weight sub-tlv) protection (protecting the first segment list, can't have

segment

segment

...

...

...

- o SR P2MP-POLICY NLRI and non-shared tree Replication segment route type or shared tree Replication segment route type.
- o Tunnel Encapsulation Attribute is defined in [ietf-idr-tunnel-encaps].
- o Tunnel-Type is set to Replication Segment Tunnel Type, TBD (assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry).
- o tree-identifier, replication-sid (binding sid), SRv6 replication-sid, downstream-nodes and segment-list are defined in this document.
- o Additional sub-TLVs may be defined in the future.

3.3. P2MP Policy Sub-TLVs

EACH P2MP policy NLRI represents a candidate path for a P2MP policy. A P2MP policy can have multiple candidate paths and would need multiple P2MP policy NLRI to download all the candidate paths.

3.3.1. preference Sub-TLV

As defined in preference Sub-TLV section in [draft-ietf-idr-segment-routing-te-policy] the candidate path with highest preference is the active candidate path.

3.3.2. leaf-list Sub-TLV

The leaf list sub-tlv identifies a set of leaves for the tree. Each leaf is a remote endpoint as defined in [ietf-idr-tunnel-encaps] The leaf-list sub-tlv is optional. The PCE can choose to download the leaf list every time it is configured or learns a new leaf. If the PCE chooses to download this optional sub-tlv it should download the entire set of the end-points every time the endpoint list has been modified. The leaf list has informational value only hence why it is optional and it is not required for the root PE to operate. However, it must be noted that in some cases the end-points list can become very large with 100s of leaves.

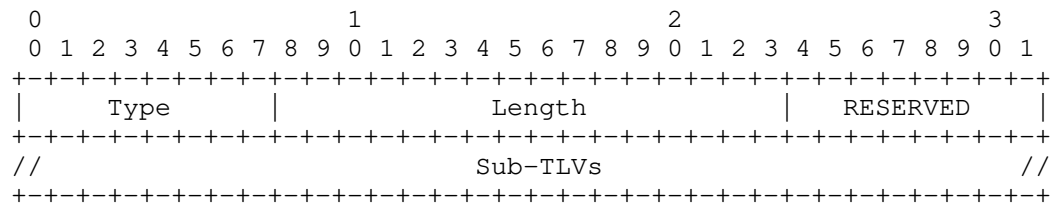
0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type										Length										RESERVED																			
//										sub-TLVs										//																			

- o Type: TBD, 1 octet
- o Length: 2 octets, the total length (not including the Type and Length fields) of the sub-TLVs encoded within the leaf-list sub-TLV.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o sub-TLVs: One or more remote endpoint sub-TLVs. Note the remote endpoint object is defined in [ietf-idr-tunnel-encaps]

3.3.3. path-instance Sub-TLV

The path instance sub-tlv contains a set of instance-ids (P2MP LSPs). These LSPs can be used for MBB procedure under a candidate path. Each LSP Instance-id has a unique id (4 octets) with in the <root node, P2MP policy>, in other word it is unique per <root node,tree-id>. The PCE SHOULD always download all instance-ids to the node. The active instance is identified via the active instance-id sub-tlv.

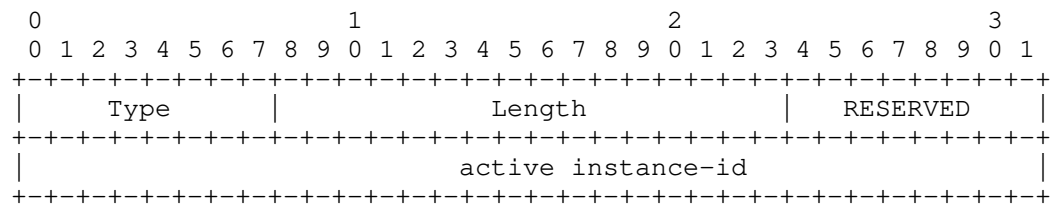
The P2MP LSP and its replication segments should be configured from root to the leaves first before the PCE switches that active instance-id to this new instance.



- o Type: TBD, 1 octet
- o Length: 2 octets, the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt
- o sub-TLVs: * active instance-id * one or more instance-id

3.3.3.1. active instance-id Sub-TLV

The Active instance-id is used to identify the P2MP LSP which should be active amongst the collection of instances.

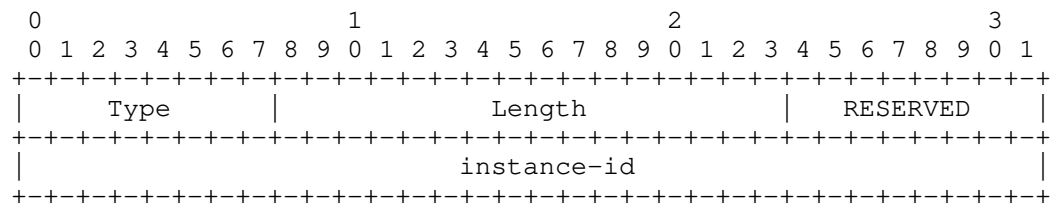


- o Type: TBD.

- o Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o active instant-id: The identifier of the active instance-id

3.3.3.2. instance-id Sub-TLV

Multiple Instance-ids can be programmed for a candidate path.



- o Type: TBD
- o Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the Segment List sub-TLV.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o instan-id: a 32 bit unique identifier. The instance-id is unique with in the context of the <root node, p2mp policy>

3.4. Replication segment Sub-TLVs

3.4.1. Replication SID (Binding SID)

The Replication SID is form of a Binding SID as it is defined in [draft-ietf-idr-segment-routing-te-policy]. The definition of replication sid with in P2MP Policy is defined in [draft-ietf-spring-sr-replication-segment]. On the transit and leaf node the replication SID can be used to identify the replication segment and the forwarding information at the node. However on the head-end node (Root), the replication segment acts as a Binding SID to direct the traffic into the P2MP Tree. It should be noted that two replication SIDs can be directly connected or connected via a unicast SR segment list, in this case the replication sid needs to be at the bottom of sid.

The sr-te-policy binding sid and SRv6 binding sid sub-tlvs are used for replication sid. This draft defines a new flag for replication sid at transit and leaf node

```

  0 1 2 3 4 5 6 7
+---+---+---+---+---+---+
|S|I|R|         |
+---+---+---+---+---+---+

```

R-FLAG: is Replication SID. Replication SID can be used to define the forwarding information of the transit or leaf nodes.

3.4.2. Down stream nodes Sub-TLV

The down-stream nodes sub-tlv is the list of down stream nodes that the arriving packet needs to be replicated to. As an example an arriving packet that needs to be replicated to downstream node A and node B will have two down stream node TLVs. Each down stream node sub-tlv could have a single segment list or multiple segment list. Multiple segment list can be used for ECMP or fast reroute. In the case of the fast reroute the downstream node flag needs to set the P bit to explicitly indicate the this downstream node is protected and the protection sub-tlv needs to be included with every segment list.

```

      0              1              2              3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|      Type      |      Length      |      Flags      |P| RESERVED |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                                     downstream node id                                     ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
//                                     sub-TLVs                                     //
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

- o Type: TBD.
- o Length: the total length (not including the Type and Length fields) of the sub-TLVs encoded within the down-stream nodes sub-TLV.
- o RESERVED: 1 octet of reserved bits. SHOULD be unset on transmission and MUST be ignored on receipt.
- o flags p: this down stream node has protected segment list.
- o downstream node id: an id to uniquely identify the downstream node for this sub-tlv, as an example the loopback IPv4/IPv6 of the node.

- o sub-TLVs: One or more segment list sub-TLVs. As an example there can be Two segment list for ECMP or FRR.

3.4.3. Segment list Sub-TLV

The segment list Sub-TLV is defined in [ietf-spring-segment-routing-policy]. The segment-list Sub-TLV contains one or more segment Sub-TLVs. Two replication segments can be directly connected via a replication sid or can be connected via a unicast segment list and a replication sid. In the later case the replication sid needs to be at the bottom of the unicast segment list.

3.4.4. Weight sub-tlv

The Weight sub-TLV is optional and is as defined in [draft-ietf-idr-segment-routing-te-policy]. With in the downstream node sub-tlv, there can be one or more segment list used for ECMP. In this case the weight sub-tlv can provide weighted ECMP.

3.4.5. Protection sub-tlv

Protection sub-tlv is optional, if FRR is desired for the downstream node this sub-tlv can be used to identify the protection segment list. To identify protection segment list this sub-tlv provides a segment list identifier. If protection is desired under the endpoint all the segment lists should have this sub-tlv. A protection segment list can not have a weight sub-tlv and it can not participate in ECMP. That said a segment list that is being protected can have a weight sub-tlv and participate in ECMP.

In general protection segment list is used only if replication segments are directly connected and there is no unicast segment list connecting two replication segment. If there is a unicast replication segment connecting the two replication sid, then the unicast protection mechanism can be exercise and there is no need for this protection sub-tlv, hence why this sub-tlv is optional.

0										1										2										3										
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1									
Type										Length										Flags										P	RESERVED									
segment list id										protection segment list id																														

- o Type : tbd, 1 octet.

- o Length: 8
- o Flag: 1 octet, the P bit is set when this segment list is protected by another segment list for the downstream node
- o segment list id: the segment list id
- o protection segment list id: the segment list id that is being used as protection.

3.4.6. Segment Sub-TLV

The segment sub-Tlv is identified in [draft-ietf-idr-segment-routing-te-policy]. As it was mentioned before two replication segments can be connected directly to each other or via a segment list. If they are connected directly to each other then the segment list can be constructed via:

- o If the replication segment is steered via IPv4 or IPv6 nexthops or interface then the segment type E or G can be used with the new R flag set.
- o If the replication segment is steered via a SR Unicast node or adjacency SID then segment type A can be used with the new R flag set. Unicast SR segment types can also be configured for steering.

If they are connected via SR domain then the segment list can contain multiple different types of SIDs, such as Node, Adjacency or Binding SIdE. In this case the replication sid is at the bottom of the stack and of type A with the R flag set. The SR node/adjacency or binding sids steer the packet through a SR domain until it reaches another replication segment. where the bottom of the stack replication sid identifies the forwarding information on that replication segment.

A replication segment can use the same type of segment types defined in [draft-ietf-idr-segment-routing-te-policy]. To identify a replication segment explicitly a new flag is defined.

```

0 1 2 3 4 5 6 7
+---+---+---+---+
|V|A|R|         |
+---+---+---+---+

```

Where R-Flag is set for a segment Sub-TLV that identifies a Replication Segment. It should be noted that in a segment list only the last segment can have the R flag set. Multiple replication segments can not be stacked on top of each other. That said there

can be special cases for Link Protection where a bypass tunnel is build via a shared replication segment. As an example when the PCE downloads a bypass tunnel for link protection that is only constructed via shared replication segments to protect a group of non-shared replication segments.

4. P2MP Policy Operation

Inline with [draft-ietf-idr-segment-routing-te-policy] the consumer of an P2MP Policy is not the BGP process. The BGP process is used for distributing the P2MP policy NLRI and its route-types but its installation and use is outside the scope of BGP. The detail for P2MP Policy can be found in [draft-ietf-pim-sr-p2mp-policy]

4.1. Configuration and advertisement of P2MP Policies

The controller usually is connected to the receivers via a route reflector. As such one or more route-target SHOULD be attached to the advertisement of P2MP Policy NLRI and its route-type. Each route target identifies one head-end (root nodes) for P2MP Policy route or one or more head-end, transit and leaf nodes for the Non- Shared/ Shared Tree Replication Segment route, for the advertised P2MP Policy.

4.2. Reception of an P2MP Policy NLRI

When a BGP speaker receives an P2MP Policy NLRI the following rules apply:

- o The P2MP Policy update MUST have either the NO_ADVERTISE community or at least one route-target extended community in IPv4-address format. If a router supporting this document receives an P2MP Policy update with no route-target extended communities and no NO_ADVERTISE community, the update MUST NOT be processed. Furthermore, it SHOULD be considered to be malformed, and the "treat-as-withdraw" strategy of [RFC7606] is applied.
- o If one or more route-targets are present, then at least one route-target MUST match one of the BGP Identifiers of the receiver in order for the update to be considered usable. The BGP Identifier is defined in [RFC4271] as a 4 octet IPv4 address. Therefore the route- target extended community MUST be of the same format.
- o If one or more route-targets are present and no one matches any of the local BGP Identifiers, then, while the P2MP Policy NLRI is acceptable, it is not usable on the receiver node.

4.3. Global Optimization for P2MP LSPs

When a P2MP LSP needs to be optimized for any reason (i.e. it is taking on an FRR Path or new routers are added to the network) a global optimization is possible. Note that optimization works per candidate path. Each candidate path is capable of global optimization. To do so each candidate path contains two or more path- instances. Each path instance is a P2MP LSP, each P2MP LSP is identified via a path-instance-id (equivalent to an lsp-id [RFC3209]). After calculating an optimized P2MP LSP path the PCE will program the candidate path with a 2nd path instance and its set of replication segments for this path-instance on the root, transit and leaf nodes. After the optimized LSP replication segments are downloaded a MBB procedure is performed and the previous instance of the path instance is deleted and removed from head-end node and its corresponding replication segments from head-end, transit and leaves.

5. IANA Consideration

- o A new SAFI is defined: the SR P2MP Policy SAFI, (Codepoint tbd assigned by IANA)
- o 2 new Route type field defines the encoding of the rest of the P2MP- POLICY SAFI
 - * P2MP Policy Route
 - * Replication Segment
- o Two new Tunnel type to be assigned by IANA

6. Security Considerations

TBD

7. Acknowledgments

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>>.

8.2. Informative References

[draft-ietf-bess-mvpn-evpn-sr-p2mp]

.

[draft-ietf-idr-segment-routing-te-policy]

.

[draft-ietf-pim-sr-p2mp-policy]

"D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang,
"draft-ietf-pim-sr-p2mp-policy"", October 2019.

[draft-ietf-spring-segment-routing-policy]

.

[draft-ietf-spring-sr-replication-segment]

"D. Yoyer, C. Filsfils, R.Prekh, H.bidgoli, Z. Zhang,
"draft-ietf-pim-sr-p2mp-policy "", July 2020.

[ietf-idr-tunnel-encaps]

.

[ietf-spring-segment-routing-policy]

.

[RFC4271] .

[RFC4760] .

[RFC6513] .

[RFC7606] .

Authors' Addresses

Hooman Bidgoli (editor)
Nokia
Ottawa
Canada

Email: hooman.bidgoli@nokia.com

Daniel Voyer
Bell Canada
Montreal
Canada

Email: daniel.yover@bell.ca

Andrew Stone
Nokia
Ottawa
Canada

Email: andrew.stone@nokia.com

Rishabh Parekh
Cisco System
San Jose
USA

Email: riparekh@cisco.com

Serge Krier
Cisco System, Inc.
Rixensart
Belgium

Email: sekrier@cisco.com

Arvind Venkateswaran
Cisco System, Inc.
Ottawa
Canada

Email: arvvenka@cisco.com

BESS
Internet-Draft
Intended status: Standards Track
Expires: October 13, 2022

Z. Zhang
Juniper Networks
R. Raszuk
NTT Network Innovations
D. Pacella
Verizon
A. Gulko
Edward Jones Wealth Management
April 11, 2022

Controller Based BGP Multicast Signaling
draft-ietf-bess-bgp-multicast-controller-09

Abstract

This document specifies a way that one or more centralized controllers can use BGP to set up multicast distribution trees (identified by either IP source/destination address pair, mLDP FEC, or SR-P2MP Tree-ID) in a network. Since the controllers calculate the trees, they can use sophisticated algorithms and constraints to achieve traffic engineering. The controllers directly signal dynamic replication state to tree nodes, leading to very simple multicast control plane on the tree nodes, as if they were using static routes. This can be used for both underlay and overlay multicast trees, including replacing BGP-MVPN signaling.

Requirements Language

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

Status of This Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 13, 2022.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Overview	3
1.1. Introduction	3
1.2. Resilience	4
1.3. Signaling	5
1.4. Label Allocation	6
1.4.1. Using a Common per-tree Label for All Routers	7
1.4.2. Upstream-assignment from Controller's Local Label Space	8
1.5. Determining Root/Leaves	9
1.5.1. PIM-SSM/Bidir or mLDP	9
1.5.2. PIM ASM	9
1.6. Multiple Domains	10
1.7. SR-P2MP	11
2. Alternative to BGP-MVPN	11
3. Specification	13
3.1. Enhancements to TEA	13
3.1.1. Any-Encapsulation Tunnel	13
3.1.2. Load-balancing Tunnel	14
3.1.3. Segment List Tunnel	14
3.1.4. Receiving MPLS Label Stack	14
3.1.5. RPF Sub-TLV	15
3.1.6. Tree Label Stack sub-TLV	15
3.1.7. Backup Tunnel sub-TLV	16
3.2. Context Label TLV in BGP-LS Node Attribute	17
3.3. Replicate State Route Type	17
3.4. SR P2MP Signaling	18

3.4.1.	Replication State Route for SR P2MP	18
3.4.2.	BGP Community Container for SR P2MP Policy	19
3.4.3.	Tunnel Encapsulation Attribute	20
3.5.	Replication State Route with Label Stack for Tree Identification	21
4.	Procedures	21
5.	Security Considerations	22
6.	IANA Considerations	22
7.	Acknowledgements	23
8.	References	23
8.1.	Normative References	23
8.2.	Informative References	24
	Authors' Addresses	25

1. Overview

1.1. Introduction

[I-D.ietf-bess-bgp-multicast] describes a way to use BGP as a replacement signaling for PIM [RFC7761] or mLDP [RFC6388]. The BGP-based multicast signaling described there provides a mechanism for setting up both (s,g)/(*,g) multicast trees (as PIM does, but optionally with labels) and labeled (MPLS) multicast tunnels (as mLDP does). Each router on a tree performs essentially the same procedures as it would perform if using PIM or mLDP, but all the inter-router signaling is done using BGP.

These procedures allow the routers to set up a separate tree for each individual multicast (x,g) flow where the 'x' could be either 's' or '*', but they also allow the routers to set up trees that are used for more than one flow. In the latter case, the trees are often referred to as "multicast tunnels" or "multipoint tunnels", and specifically in this document they are mLDP tunnels (except that they are set up with BGP signaling). While it actually does not have to be restricted to mLDP tunnels, mLDP FEC is conveniently borrowed to identify the tunnel. In the rest of the document, the term tree and tunnel are used interchangeably.

The trees/tunnels are set up using the "receiver-initiated join" technique of PIM/mLDP, hop by hop from downstream routers towards the root. The BGP messages of MCAST-TREE SAFI are either sent hop by hop between downstream routers and their upstream neighbors, or can be reflected by Route Reflectors (RRs).

As an alternative to each hop independently determining its upstream router and signaling upstream towards the root (following PIM/mLDP model), the entire tree can be calculated by a centralized controller, and the signaling can be entirely done from the

controller using the same MCAST-TREE SAFI. For that, some additional procedures and optimizations are specified in this document.

[I-D.ietf-bess-bgp-multicast] uses S-PMSI, Leaf, and Source Active Auto-Discovery (A-D) routes because the main procedures and concepts are borrowed from the BGP-MVPN [RFC6514]. While the same Leaf A-D routes can be used to signal replication state to tree nodes from controllers, this document introduces a new route type "Replication State" for the same functionality, so that familiarity with the BGP-MVPN concepts is not required.

While it is outside the scope of this document, signaling from the controllers could be done via other means as well, like Netconf or any other SDN methods.

1.2. Resilience

Each router could establish direct BGP sessions with one or more controllers, or it could establish BGP sessions with RRs who in turn peer with controllers. For the same tree/tunnel, each controller may independently calculate the tree/tunnel and signal the routers on the tree/tunnel using MCAST-TREE Replication State routes. How the calculation is done are outside the scope of this document.

On each router, BGP route selection rules will lead to one controller's route for the tree/tunnel being selected as the active route and used for setting up forwarding state. As long as all the routers on a tree/tunnel consistently pick the same controller's routes for the tree/tunnel, the setup should be consistent. If the tree/tunnel is labeled, different labels will be used from different controllers so there is no traffic loop issue even if the routers do not consistently select the same controller's routes. In the unlabeled case, to ensure the consistency the selection SHOULD be solely based on the identifier of the controller.

Another consistency issue is when a bidirectional tree/tunnel needs to be re-routed. Because this is no longer triggered hop-by-hop from downstream to upstream, it is possible that the upstream change happens before the downstream, causing traffic loop. In the unlabeled case, there is no good solution (other than that the controller issues upstream change only after it gets acknowledgement from downstream). In the labeled case, as long as a new label is used there should be no problem.

Besides the traffic loop issue, there could be transient traffic loss before both the upstream and downstream's forwarding state are updated. This could be mitigated if the upstream keep sending traffic on the old path (in addition to the new path) and the

downstream keep accepting traffic on the old path (but not on the new path) for some time. It is a local matter when for the downstream to switch to the new path - it could be data driven (e.g., after traffic arrives on the new path) or timer driven.

For each tree, multiple disjoint instances could be calculated and signaled for live-live protection. Different labels are used for different instances, so that the leaves can differentiate incoming traffic on different instances. As far as transit routers are concerned, the instances are just independent. Note that the two instances are not expected to share common transit routers (it is otherwise outside the scope of this document/revision).

1.3. Signaling

When a router receives a Replication State route, the re-advertisement is blocked if a configured import RT matches the RT of the route, which indicates that this router is the target and consumer of the route hence it should not be re-advertised further. The routes includes the forwarding information in the form of Tunnel Encapsulation Attributes (TEA) [RFC9012], with enhancements specified in this document.

Suppose that for a particular tree, there are two downstream routers D1 and D2 for a particular upstream router U. A controller C sends one Replication State route to U, with the Tree Node's IP Address field (see Section 3.3) set to U's IP address and the TEA specifying both the two downstreams and its upstream (see Section 3.1.5). In this case, the Originating Router's Address field of the Replication State route is set to the controller's address. Note that for a TEA attached to a unicast NLRI, only one of the tunnels in a TEA is used for forwarding a particular packet, while all the tunnels in a TEA are used to reach multiple endpoints when it is attached to a multicast NLRI.

It could be that U may need to replicate to many downstream routers, say D1 through D1000. In that case, it may not be possible to encode all those branches in a single TEA, or may not be optimal to update a large TEA when a branch is added/removed. In that case, C may send multiple Replication State routes, each with a different Originating Router's Address field and a different TEA that encodes a subset of the branches. This provides a flexible way to optimize the encoding of large number of branches and incremental updates of branches.

Notice that, in case of labeled trees, the (x,g), mLDLP FEC, or SR-P2MP tree identification (Section 1.7) signaling is actually not needed to transit routers but only needed to tunnel root/leaves. However, for consistency among the root/leaf/transit nodes, and for

consistency with the hop-by-hop signaling, the same signaling (with tree identification encoded in the NLRI) is used to all routers.

Nonetheless, a new NLRI route type of the MCAST-TREE SAFI is defined to encode label/SID instead of tree identification in the NLRI, for scenarios where there is really no need to signal tree identification, e.g. as described in Section 2. On a tunnel root, the tree's binding SID can be encoded in the NLRI.

For a tree node to acknowledge to the controller that it has received the signaling and installed corresponding forwarding state, it advertises a corresponding Replication State route, with the Originating Router's IP Address set to itself and with a Route Target to match the controller. For comparison, the tree signaling Replication State route from the controller has the Originating Router's IP Address set to the controller and the Route Target matching the tree node. The two Replication State routes (for controller to signal to a tree node and for a tree node to acknowledge back) differ only in those two aspects.

With the acknowledgement Replication State routes, the controller knows if tree setup is complete. The information can be used for many purposes, e.g. the controller may instruct the ingress to start forwarding traffic onto a tree only after it knows that the tree setup has completed.

1.4. Label Allocation

In the case of labeled multicast signaled hop by hop towards the root, whether it's (x,g) multicast or "mLDP" tunnel, labels are assigned by a downstream router and advertised to its upstream router (from traffic direction point of view). In the case of controller based signaling, routers do not originate tree join routes anymore, so the controllers have to assign labels on behalf of routers, and there are three options for label assignment:

- o From each router's SRLB that the controller learns
- o From the common SRGB that the controller learns
- o From the controller's local label space

Assignment from each router's SRLB is no different from each router assigning labels from its own local label space in the hop-by-hop signaling case. The assignments for one router is independent of assignments for another router, even for the same tree.

Assignment from the controller's local label space is upstream-assigned [RFC5331]. It is used if the controller does not learn the common SRGB or each router's SRLB. Assignment from the SRGB [RFC8402] is only meaningful if all SRGBs are the same and a single common label is used for all the routers on a tree in case of unidirectional tree/tunnel (Section 1.4.1). Otherwise, assignment from SRLB is preferred.

The choice of which of the options to use depends on many factors. An operator may want to use a single common label per tree for ease of monitoring and debugging, but that requires explicit RPF checking and either common SRGB or upstream assigned labels, which may not be supported due to either the software or hardware limitations (e.g. label imposition/disposition limits). In an SR network, assignment from the common SRGB if it's required to use a single common label per unidirectional tree, or otherwise assignment from SRLB is a good choice because it does not require support for context label spaces.

1.4.1. Using a Common per-tree Label for All Routers

MPLS labels only have local significance. For an LSP that goes through a series of routers, each router allocates a label independently and it swaps the incoming label (that it advertised to its upstream) to an outgoing label (that it received from its downstream) when it forwards a labeled packet. Even if the incoming and outgoing labels happen to be the same on a particular router, that is just incidental.

With Segment Routing, it is becoming a common practice that all routers use the same SRGB so that a SID maps to the same label on all routers. This makes it easier for operators to monitor and debug their network. The same concept applies to multicast trees as well - a common per-tree label can be used for a router to receive traffic from its upstream neighbor and replicate traffic to all its downstream neighbor.

However, a common per-tree label can only be used for unidirectional trees. Additionally, unless the entire tree is updated for every tree node to use a new common per-tree label with any change in the tree (no matter how small and local the change is), it requires each router to do explicit RPF check, so that only packets from its expected upstream neighbor are accepted. Otherwise, traffic loop may form during topology changes, because the forwarding state update is no longer ordered.

Traditionally, p2mp mpls forwarding does not require explicit RPF check as a downstream router advertises a label only to its upstream router and all traffic with that incoming label is presumed to be

from the upstream router and accepted. When a downstream router switches to a different upstream router a different label will be advertised, so it can determine if traffic is from its expected upstream neighbor purely based on the label. Now with a single common label used for all routers on a tree to send and receive traffic with, a router can no longer determine if the traffic is from its expected neighbor just based on that common tree label. Therefore, explicit RPF check is needed. Instead of interface based RPF checking as in PIM case, neighbor based RPF checking is used - a label identifying the upstream neighbor precedes the common tree label and the receiving router checks if that preceding neighbor label matches its expected upstream neighbor. Notice that this is similar to what's described in Section "9.1.1 Discarding Packets from Wrong PE" of RFC 6513 (an egress PE discards traffic sent from a wrong ingress PE). The only difference is one is used for label based forwarding and the other is used for (s,g) based forwarding. [note: for bidirectional trees, we may be able to use two labels per tree - one for upstream traffic and one for downstream traffic. This needs further verification].

Both the common per-tree label and the neighbor label are allocated either from the common SRGB or from the controller's local label space. In the latter case, an additional label identifying the controller's label space is needed, as described in the following section.

1.4.2. Upstream-assignment from Controller's Local Label Space

In this case in the multicast packet's label stack the tree label and upstream neighbor label (if used in case of single common-label per tree) are preceded by a downstream-assigned "context label". The context label identifies a context-specific label space (the controller's local label space), and the upstream-assigned label that follows it is looked up in that space.

This specification requires that, in case of upstream-assignment from a controller's local label space, each router D to assign, corresponding to each controller C, a context label that identifies the upstream-assigned label space used by that controller. This label, call it Lc-D, is communicated by D to C via BGP-LS [RFC 7752].

Suppose a controller is setting up unidirectional tree T. It assigns that tree the label Lt, and assigns label Lu to identify router U which is the upstream of router D on tree T. C needs to tell U: "to send a packet on the given tree/tunnel, one of the things you have to do is push Lt onto the packet's label stack, then push Lu, then push Lc-D onto the packet's label stack, then unicast the packet to D".

Controller C also needs to inform router D of the correspondence between <Lc-D, Lu, Lt> and tree T.

To achieve that, when C sends a Replication State route, for each tunnel in the TEA, it may include a label stack Sub-TLV [RFC9012], with the outer label being the context label Lc-D (received by the controller from the corresponding downstream), the next label being the upstream neighbor label Lu, and the inner label being the label Lt assigned by the controller for the tree. The router receiving the route will use the label stacks to send traffic to its downstreams.

For C to signal the expected label stack for D to receive traffic with, we overload a tunnel TLV in the TEA of the Replication State route sent to D - if the tunnel TLV has a RPF sub-TLV (Section 3.1.5), then it indicates that this is actually for receiving traffic from the upstream.

1.5. Determining Root/Leaves

For the controller to calculate a tree, it needs to determine the root and leaves of the tree. This may be based on provisioning (static or dynamically programmed), or based on BGP signaling as described in the following two sections.

In both of the following cases, the BGP updates are targeted at the controller, via an address specific Route Target with Global Administration Field set to the controller's address and the Local Administration Field set to 0.

1.5.1. PIM-SSM/Bidir or mLDP

In this case, the PIM Last Hop Routers (LHRs) with interested receivers or mLDP tunnel leaves encode a Leaf A-D route ([I-D.ietf-bess-bgp-multicast]) with the Upstream Router's IP Address field set to the controller's address and the Originating Router's IP Address set to the address of the LHR or the P2MP tunnel leaf. The encoded PIM SSM source or mLDP FEC provides root information and the Originating Router's IP Address provides leaf information.

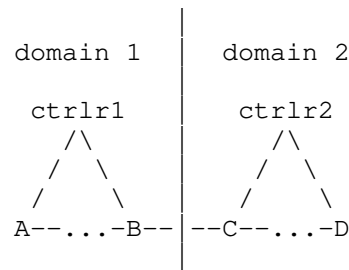
1.5.2. PIM ASM

In this case, the First Hop Routers (FHRs) originate Source Active routes which provides root information, and the LHRs originate Leaf A-D routes, encoded as in the PIM-SSM case except that it is (*,G) instead of (S,G). The Leaf A-D routes provide leaf information.

1.6. Multiple Domains

An end to end multicast tree may span multiple routing domains, and the setup of the tree in each domain may be done differently as specified in [I-D.ietf-bess-bgp-multicast]. This section discusses a few aspects specific to controller signaling.

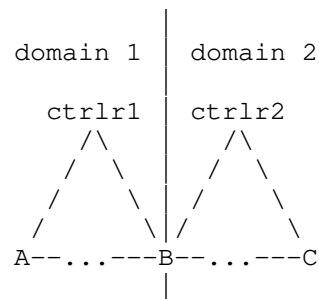
Consider two adjacent domains each with its own controller in the following configuration where router B is an upstream node of C for a multicast tree:



In the case of native (un-labeled) IP multicast, nothing special is needed. Controller 1 signals B to send traffic out of B-C link while Controller 2 signals C to accept traffic on the B-C link.

In the case of labeled IP multicast or mLDP tunnel, the controllers may be able to coordinate their actions such that Controller 1 signals B to send traffic out of B-C link with label X while Controller 2 signals C to accept traffic with the same label X on the B-C link. If the coordination is not possible, then C needs to use hop-by-hop BGP signaling to signal towards B, as specified in [I-D.ietf-bess-bgp-multicast].

The configuration could also be as following, where router B borders both domain 1 and domain 2 and is controlled by both controllers:



As discussed in Section 1.2, when B receives signaling from both Controller 1 and Controller 2, only one of the routes would be selected as the best route and used for programming the forwarding state of the corresponding segment. For B to stitch the two segments together, it is expected for B to know by provisioning that it is a border router so that B will look for the other segment (represented by the signaling from the other controller) and stitch the two together.

1.7. SR-P2MP

[I-D.ietf-pim-sr-p2mp-policy] describes an architecture to construct a Point-to-Multipoint (P2MP) tree to deliver Multi-point services in a Segment Routing domain. An SR P2MP tree is constructed by stitching together a set of Replication Segments that are specified in [I-D.ietf-spring-sr-replication-segment]. An SR Point-to-Multipoint (SR P2MP) Policy is used to define and instantiate a P2MP tree which is computed by a controller.

An SR P2MP tree is no different from an mLDP tunnel in MPLS forwarding plane. The difference is in control plane - instead of hop-by-hop mLDP signaling from leaves towards the root, to set up SR P2MP trees controllers program forwarding state (referred to as Replication Segments) to the root, leaves, and intermediate replication points using Netconf, PCEP, BGP or any other reasonable signaling/programming methods.

Procedures in this document can be used for controllers to set up SR P2MP trees with just an additional SR P2MP tree type and corresponding tree identification in the Replication State route.

If/once the SR Replication Segment is extended to bi-redirectional, and SR MP2MP is introduced, the same procedures in this document would apply to SR MP2MP as well.

2. Alternative to BGP-MVPN

Multicast with BGP signaling from controllers can be an alternative to BGP-MVPN [RFC6514]. It is an attractive option especially when the controller can easily determine the source and leaf information.

With BGP-MVPN, distributed signaling is used for the following:

- o Egress PEs advertise C-multicast (Type-6/7) Auto-Discovery (A-D) routes to join C-multicast trees at the overlay (PE-PE).

- o In case of ASM, ingress PEs advertise Source Active (Type-5) A-D routes to signal sources so that egress PEs can establish Shortest Path Trees (SPT).
- o PEs advertise I/S-PMSI (Type-1/2/3) A-D routes to signal the binding of overlay/customer traffic to underlay/provider tunnels. For some types of tunnels, Leaf (Type-4) A-D routes are advertised by egress PEs in response to I/S-PMSI A-D routes to join the tunnels.

Based on the above signaled information, an ingress PE builds forwarding state to forward traffic arriving on the PE-CE interface to the provider tunnel (and local interfaces if there are local downstream receivers), and an egress PE builds forwarding state to forward traffic arriving on a provider tunnel to local interfaces with downstream receivers.

Notice that multicast with BGP signaling from controllers essentially programs "static" forwarding state onto multicast tree nodes. As long as a controller can determine how a C-multicast flow should be forwarded on ingress/egress PEs, it can signal to the ingress/egress PEs using the procedures in this document to set up forwarding state, removing the need of the above-mentioned distributed signaling and processing.

For the controller to learn the egress PEs for a C-multicast tree (so that it can set up or find a corresponding provider tunnel), the egress PEs advertise MCAST-TREE Leaf A-D routes (Section 1.5.1) towards the controller to signal its desire to join C-multicast trees, each with an appropriate RD and an extended community derived from the Route Target for the VPN ([I-D.zzhang-idr-rt-derived-community]) so that the controller knows which VPN it is for. The controller then advertises corresponding MCAST-TREE Replication State routes to set up C-multicast forwarding state on ingress and egress PEs. To encode the provider tunnel information in the MCAST-TREE Replication State route for an ingress PE, the TEA can explicitly list all replication branches of the tunnel, or just the binding SID for the provider tunnel in the form of Segment List tunnel type, if the tunnel has a binding SID.

The Replication State route may also have a PMSI Tunnel Attribute (PTA) attached to specify the provider tunnel while the TEA specifies the local PE-CE interfaces where traffic need to be sent out. This not only allows provider tunnel without a binding SID (e.g., in a non-SR network) to be specified without explicitly listing its replication branches, but also allows the service controller for MVPN overlay state to be independent of provider tunnel setup (which could

be from a different transport controller or even without a controller).

However, notice that if the service controller and transport controller are different, then the service controller needs to signal the transport controller the tree information: identification, set of leaves, and applicable constraints. While this can be achieved (see Section 1.5.1), it is easier for the service and transport controller to be the same.

Depending on local policy, a PE may add PE-CE interfaces to its replication state based on local signaling (e.g., IGMP/PIM) instead of completely relying on signaling from controllers.

If dynamic switching between inclusive and selective tunnels based on data rate is needed, the ingress PE can advertise/withdraw S-PMSI routes targeted only at the controllers, without PMSI Tunnel Attribute attached. The controller then updates relevant MCAST-TREE Replication State routes to update C-multicast forwarding states on PEs to switch to a new tunnel.

3. Specification

3.1. Enhancements to TEA

A TEA may encode a list of tunnels. A TEA attached to an MCAST-TREE NLRI encodes replication information for a <tree, node > that is identified by the NLRI. Each tunnel in the TEA identifies a branch - either an upstream branch towards the tree root (Section 3.1.5) or a downstream branch towards some leaves. A tunnel in the TEA could have an outer encapsulation (e.g. MPLS label stack) or it could just be a one-hop direct connection for native IP multicast forwarding without any outer encapsulation.

This document specifies three new Tunnel Types and four new sub-TLVs. The type codes will be assigned by IANA from the "BGP Tunnel Encapsulation Attribute Tunnel Types".

3.1.1. Any-Encapsulation Tunnel

When a multicast packet needs to be sent from an upstream node to a downstream node, it may not matter how it is sent - natively when the two nodes are directly connected or tunneled otherwise. In case of tunneling, it may not matter what kind of tunnel is used - MPLS, GRE, IPinIP, or whatever.

To support this, an "Any-Encapsulation" tunnel type of value 20 is defined. This tunnel MAY have a Tunnel Egress Endpoint and other

Sub-TLVs. The Tunnel Egress Endpoint Sub-TLV specifies an IP address, which could be any of the following:

- o An interface's local address - when a packet needs to be sent out of the corresponding interface natively. On a LAN multicast MAC address MUST be used.
- o A directly connected neighbor's interface address - when a packet needs to be unicast to the address natively.
- o An address that is not directly connected - when a packet needs to be tunneled to the address (any tunnel type/instance can be used).

3.1.2. Load-balancing Tunnel

Consider that a multicast packet needs to be sent to a downstream node, which could be reached via four paths P1~P4. If it does not matter which of path is taken, an "Any-Encapsulation" tunnel with the Tunnel Egress Endpoint Sub-TLV specifying the downstream node's loopback address works well. If the controller wants to specify that only P1~P2 should be used, then a "Load-balancing" tunnel needs to be used, listing P1 and P2 as member tunnels of the "Load-balancing" tunnel.

A load-balancing tunnel has one "Member Tunnels" Sub-TLV defined in this document. The Sub-TLV is a list of tunnels, each specifying a way to reach the downstream. A packet will be sent out of one of the tunnels listed in the Member Tunnels Sub-TLV of the load-balancing tunnel.

3.1.3. Segment List Tunnel

A Segment List tunnel has a Segment List sub-TLV. The encoding of the sub-TLV is as specified in Section 2.4.4 of [I-D.ietf-idr-segment-routing-te-policy]. An example use of a Segment List tunnel is provided in Section 3.4.3.

3.1.4. Receiving MPLS Label Stack

While [I-D.ietf-bess-bgp-multicast] uses S-PMSI A-D routes to signal forwarding information for MP2MP upstream traffic, when controller signaling is used, a single Replication State route is used for both upstream and downstream traffic. Since different upstream and downstream labels need to be used, a new "Receiving MPLS Label Stack" of type TBD is added as a tunnel sub-TLV in addition to the existing MPLS Label Stack sub-TLV. Other than type difference, the two are the encoded the same way.

The Receiving MPLS Label Stack sub-TLV is added to each downstream tunnel in the TEA of Replication State route for an MP2MP tunnel to specify the forwarding information for upstream traffic from the corresponding downstream node. A label stack instead of a single label is used because of the need for neighbor based RPF check, as further explained in the following section.

The Receiving MPLS Label Stack sub-TLV is also used for downstream traffic from the upstream for both P2MP and MP2MP, as specified below.

3.1.5. RPF Sub-TLV

The RPF sub-TLV is of type 124 allocated by IANA and has a one-octet length. The length is 0 currently, but if necessary in the future, sub-sub-TLVs could be placed in its value part. If the RPF sub-TLV appears in a tunnel, it indicates that the "tunnel" is for the upstream node instead of a downstream node.

In case of MPLS, the tunnel contains an Receiving MPLS Label Stack sub-TLV for downstream traffic from the upstream node, and in case of MP2MP it also contains a regular MPLS Label Stack sub-TLV for upstream traffic to the upstream node.

The inner most label in the Receiving MPLS Label Stack is the incoming label identifying the tree (for comparison the inner most label for a regular MPLS Label Stack is the outgoing label). If the Receiving MPLS Label Stack sub-TLV has more than one labels, the second inner most label in the stack identifies the expected upstream neighbor and explicit RPF checking needs to be set up for the tree label accordingly.

3.1.6. Tree Label Stack sub-TLV

The MPLS Label Stack sub-TLV can be used to specify the complete label stack used to send traffic, with the stack including both a transport label (stack) and label(s) that identify the (tree, neighbor) to the downstream node. There are cases where the controller only wants to specify the tree-identifying labels but leave the transport details to the router itself. For example, the router could locally determine a transport label (stack) and combine with the tree-identifying labels signaled from the controller to get the complete outgoing label stack.

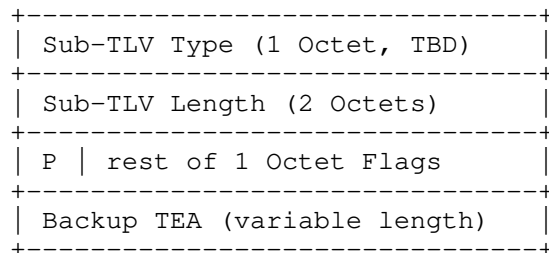
For that purpose, a new Tree Label Stack sub-TLV of type 125 is defined, with a one-octet length field. It MAY appear in an Any-Encapsulation tunnel. The value field contains a label stack with the same encoding as value part of the MPLS Label Stack sub-TLV, but

with a different type. A stack is specified because it may take up to three labels (see Section 1.4):

- o If different nodes use different labels (allocated from the common SRGB or the node's SRLB) for a (tree, neighbor) tuple, only a single label is in the stack. This is similar to current mLDP hop by hop signaling case.
- o If different nodes use the same tree label, then an additional neighbor-identifying label is needed in front of the tree label.
- o For the previous bullet, if the neighbor-identifying label is allocated from the controller's local label space, then an additional context label is needed in front of the neighbor label.

3.1.7. Backup Tunnel sub-TLV

The Backup Tunnel sub-TLV is used to specify the backup paths for an Any-Encapsulation or Segment List tunnel. The length is two-octet. The value part encodes a one-octet flags field and a variable length Tunnel Encapsulation Attribute. If the tunnel goes down, traffic that is normally sent out of the tunnel is fast rerouted to the tunnels listed in the encoded TEA.

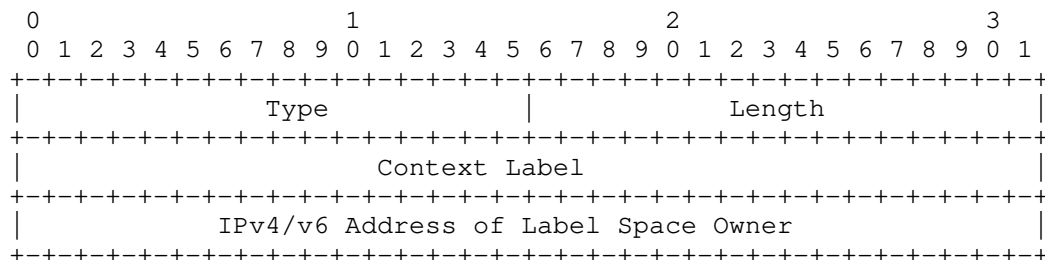


The backup tunnels can be going to the same or different nodes reached by the original tunnel.

If the tunnel carries a RPF sub-TLV and a Backup Tunnel sub-TLV, then both traffic arriving on the original tunnel and on the tunnels encoded in the Backup Tunnel sub-TLV's TEA can be accepted, if the Parallel (P-)bit in the flags field is set. If the P-bit is not set, then traffic arriving on the backup tunnel is accepted only if router has switched to receiving on the backup tunnel (this is the equivalent of PIM/mLDP MoFRR).

3.2. Context Label TLV in BGP-LS Node Attribute

For a router to signal the context label that it assigns for a controller (or any label allocator that assigns labels - from its local label space - that will be received by this router), a new BGP-LS Node Attribute TLV is defined:



The Length field implies the type of the address. Multiple Context Label TLVs may be included in a Node Attribute, one for each label space owner.

An as example, a controller with address 11.11.11.11 allocates label 200 from its own label space, and router A assigns label 100 to identify this controller's label space. The router includes the Context Label TLV (100, 11.11.11.11) in its BGP-LS Node Attribute and the controller instructs router B to send traffic to router A with a label stack (100, 200), and router A uses label 100 to determine the Label FIB in which to look up label 200.

3.3. Replicate State Route Type

The NLRI route type for signaling from controllers to tree nodes is "Replication State". The NLRI has the following format:

+-----+ Route Type - Replication State +-----+	
Length (1 octet) +-----+	
Tree Type (1 octet) +-----+	
Tree Type Specific Length (1 octet) +-----+	
~ Tree Identification (variable) ~ +-----+	
Tree Node's IP Address +-----+	
Originator's IP Address +-----+	

Replication State NLRI

Notice that Replication State is just a new route type with the same format of Leaf A-D route except some fields are renamed:

- o Tree Type in Replication State route matches the PMSI route type in Leaf A-D route
- o Tree Node's IP Address matches the Upstream Router's IP Address of the PMSI route key in Leaf A-D route

With this arrangement, IP multicast tree and mLDP tunnel can be signaled via Replication State routes from controllers, or via Leaf A-D routes either hop by hop or from controllers with maximum code reuse, while newer types of trees like SR-P2MP can be signaled via Replication State routes with maximum code reuse as well.

3.4. SR P2MP Signaling

An SR P2MP policy for an SR P2MP tree is identified by a (Root, Tree-id) tuple. It has a set of leaves and set of Candidate Paths (CPs). The policy is instantiated on the root of the tree, with corresponding Replication Segments - identified by (Root, Tree-id, Tree-Node-id) - instantiated on the tree nodes (root, leaves, and intermediate replication points).

3.4.1. Replication State Route for SR P2MP

For SR P2MP, forwarding on tree nodes state are represented as Replication Segments and are signaled from controllers to tree nodes via Replication State routes. A Replication State route for SR P2MP has a Tree Type 1 and the Tree Identification includes (Route

Distinguisher, Root ID, Tree ID), where the RD implicitly identifies the candidate path.

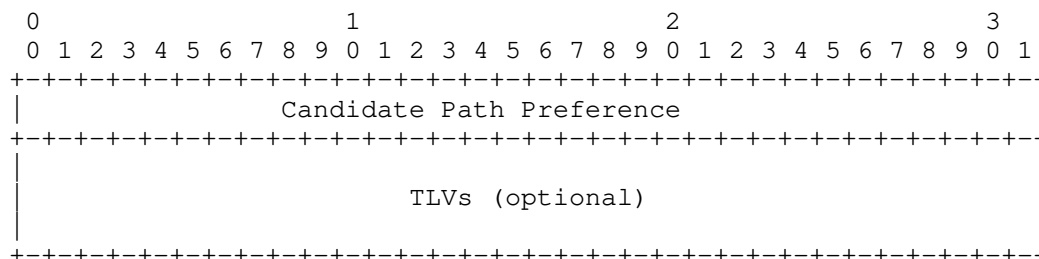
Route Type - Replication State
Length (1 octet)
Tree Type (1 - SR P2MP)
Tree Type Specific Length (1 octet)
RD (8 octets)
Root ID (4 or 16 octets)
Tree ID (4 octets)
Tree Node's IP Address
Originating Router's IP Address

Replication State route for SR Replication Segment

3.4.2. BGP Community Container for SR P2MP Policy

The Replication State route for Replication Segments signaled to the root is also used to signal (parts of) the SR P2MP Policy - the policy name, the set of leaves (optional, for informational purpose), preference of the CP and other information are all encoded in a newly defined BGP Community Container (BCC) [I-D.ietf-idr-wide-bgp-communities] called SR P2MP Policy BCC.

The SR P2MP Policy BCC has a BGP Community Container type to be assigned by IANA. It is composed of a fixed 4-octet Candidate Path Preference value, optionally followed by TLVs.



BGP Community Container for SR P2MP Policy

One optional TLV is to enclose the following optional Atoms TLVs that are already defined in [I-D.ietf-idr-wide-bgp-communities]:

- o An IPv4 or IPv6 Prefix list - for the set of leaves
- o A UTF-8 string - for the policy name

If more information for the policy are needed, more Atoms TLVs or SR P2MP Policy BCC specific TLVs can be defined.

The root receives one Replication State route for each Candidate Path of the policy. Only one of the routes need to, though more than one MAY include the above listed optional Atom TLVs in the SR P2MP Policy BCC.

Alternatively, an additional route type can be used to carry policy information instead. Details/decision to be specified in a future revision.

3.4.3. Tunnel Encapsulation Attribute

The TEA attached to a Replication State route for SR-P2MP encodes tunnels as specified in earlier sections. A tunnel could be an Any-Encapsulation tunnel with MPLS Label Stack sub-TLV or Receiving MPLS Label Stack sub-TLV (in case of SR-MPLS), a Segment List tunnel, or a Load-balancing tunnel.

For a Segment List tunnel in this context, the last segment in the segment list represents the SID of the tree. When it is without the RPF sub-TLV, the previous segments in the list steer traffic to the downstream node, and the segment before the last one MAY also be a binding SID for another P2MP tunnel, meaning that the replication branch represented by this "Segment List" is actually a P2MP tunnel to a set of downstream nodes.

3.5. Replication State Route with Label Stack for Tree Identification

As described in Section 1.3, tree label instead of tree identification could be encoded in the NLRI to identify the tree in the control plane as well as in the forwarding plane. For that a new Tree Type of 2 is used and the Replication State route has the following format:

Route Type - Replication State
Length (1 octet)
Tree Type 2 (Label as Tree ID)
Tree Type specific Length (1 octet)
RD (8 octets)
Label Stack (variable)
Tree Node's IP Address
Originating Router's IP Address

Replication State route for tree identification by label stack

As discussed in Section 1.4.2, a label stack may have to be used to identify a tree in the data plane so a label stack is encoded here. The number of labels is derived from the Tree Type Specific Length field. Each label stack entry is encoded as following:

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			
Label		0 0 0 0 0 0 0 0 0 0 0 0	

4. Procedures

Details to be added. The general idea is described in the introduction section.

5. Security Considerations

This document does not introduce new security risks.

6. IANA Considerations

IANA has assigned the following code points:

- o "Any-Encapsulation" tunnel type 78 from "BGP Tunnel Encapsulation Attribute Tunnel Types" registry
- o "RPF" sub-TLV type 124 and "Tree Label Stack" sub-TLV type 125 from "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry

This document makes the following additional IANA requests:

- o Assign "Segment List" and "Load-balancing" tunnel types from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry
- o Assign "Member Tunnels" and "Receiving MPLS Label Stack" sub-TLV types from the "BGP Tunnel Encapsulation Attribute Sub-TLVs" registry. The "Member Tunnels" sub-TLV has a two-octet value length (so the type should be in the 128-255 range), while the "Receiving MPLS Label Stack" sub-TLV has a one-octet value length.
- o Assign "Context Label TLV" type from the "BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs" registry.
- o Assign "Replication State" route type from the "BGP MCAST-TREE Route Types" registry.
- o Create a "Tree Type Registry for Replication State Route", with the following initial assignments:
 - * 1: SR-P2MP
 - * 2: P2MP Tree with Label as Identification
 - * 3: IP Multicast
 - * 0x43: mLDP
- o Assign a new BGP Community Container type "SR P2MP Policy", and to create an "SR P2MP Policy Community Container TLV Registry", with an initial entry for "TLV for Atoms".

7. Acknowledgements

The authors Eric Rosen for his questions, suggestions, and help finding solutions to some issues like the neighbor based explicit RPF checking. The authors also thank Lenny Giuliano, Sanoj Vivekanandan and IJsbrand Wijnands for their review and comments.

8. References

8.1. Normative References

- [I-D.ietf-bess-bgp-multicast]
Zhang, Z., Giuliano, L., Patel, K., Wijnands, I., Mishra, M., and A. Gulko, "BGP Based Multicast", draft-ietf-bess-bgp-multicast-04 (work in progress), January 2022.
- [I-D.ietf-idr-segment-routing-te-policy]
Previdi, S., Filsfils, C., Talaulikar, K., Mattes, P., Jain, D., and S. Lin, "Advertising Segment Routing Policies in BGP", draft-ietf-idr-segment-routing-te-policy-16 (work in progress), March 2022.
- [I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-06 (work in progress), January 2022.
- [I-D.ietf-pim-sr-p2mp-policy]
(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", draft-ietf-pim-sr-p2mp-policy-04 (work in progress), March 2022.
- [I-D.ietf-spring-sr-replication-segment]
(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "SR Replication Segment for Multi-point Service Delivery", draft-ietf-spring-sr-replication-segment-07 (work in progress), March 2022.
- [I-D.zzhang-idr-rt-derived-community]
Zhang, Z., Haas, J., and K. Patel, "Extended Communities Derived from Route Targets", draft-zzhang-idr-rt-derived-community-02 (work in progress), March 2022.

- [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>>.
- [RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <<https://www.rfc-editor.org/info/rfc7752>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <<https://www.rfc-editor.org/info/rfc9012>>.

8.2. Informative References

- [RFC6388] Wijnands, IJ., Ed., Minei, I., Ed., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, DOI 10.17487/RFC6388, November 2011, <<https://www.rfc-editor.org/info/rfc6388>>.
- [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>>.
- [RFC7060] Napierala, M., Rosen, E., and IJ. Wijnands, "Using LDP Multipoint Extensions on Targeted LDP Sessions", RFC 7060, DOI 10.17487/RFC7060, November 2013, <<https://www.rfc-editor.org/info/rfc7060>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.

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

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Robert Raszuk
NTT Network Innovations

EMail: robert@raszuk.net

Dante Pacella
Verizon

EMail: dante.j.pacella@verizon.com

Arkadiy Gulko
Edward Jones Wealth Management

EMail: arkadiy.gulko@edwardjones.com

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: 24 July 2022

P.B. Brissette, Ed.
A.S. Sajassi
LA.B. Burdet
Cisco
J.D. Drake
Juniper
J.R. Rabadan
Nokia
20 January 2022

Fast Recovery for EVPN DF Election
draft-ietf-bess-evpn-fast-df-recovery-03

Abstract

Ethernet Virtual Private Network (EVPN) solution provides Designated Forwarder election procedures for multi-homing Ethernet Segments. These procedures have been enhanced further by applying Highest Random Weight (HRW) Algorithm for Designated Forwarder election in order to avoid unnecessary DF status changes upon a failure. This draft improves these procedures by providing a fast Designated Forwarder (DF) election upon recovery of the failed link or node associated with the multi-homing Ethernet Segment. The solution is independent of number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each PEs in the multi-homing group.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119] and RFC 8174 [RFC8174].

Status of This Memo

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

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

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

This Internet-Draft will expire on 24 July 2022.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	2
1.1. Terminology	3
2. Challenges with Existing Solution	3
3. DF Election Synchronization Solution	4
3.1. Advantages	5
3.2. BGP Encoding	6
3.3. Note on NTP-based synchronization	6
3.4. Synchronization Scenarios	7
3.5. Backwards Compatibility	8
4. Security Considerations	8
5. IANA Considerations	8
6. Normative References	9
Appendix A. Contributors	9
Appendix B. Acknowledgements	10
Authors' Addresses	10

1. Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

EVPN solution [RFC7432] describes DF election procedures for multi-homing Ethernet Segments. These procedures are enhanced further in [RFC8584] by applying Highest Random Weight Algorithm for DF election in order to avoid DF status change unnecessarily upon a link or node failure associated with the multi-homing Ethernet Segment. This draft makes further improvement to DF election procedures in [RFC8584] by providing an option for a fast DF election upon recovery of the failed link or node associated with the multi-homing Ethernet Segment. This DF election is achieved independent of number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each PE in the multi-homing group. The solution is based on simple one-way signaling mechanism.

1.1. Terminology

Provider Edge (PE): A device that sits in the boundary of Provider and Customer networks and performs encaps/decap of data from L2 to L3 and vice-versa.

Designated Forwarder (DF): A PE that is currently forwarding (encapsulating/decapsulating) traffic for a given VLAN in and out of a site.

2. Challenges with Existing Solution

In EVPN technology, multiple PE devices have the ability to encaps and decap data belonging to the same VLAN. In certain situations, this may cause L2 duplicates and even loops if there is a momentary overlap of forwarding roles between two or more PE devices, leading to broadcast storms.

EVPN [RFC7432] currently uses timer based synchronization among PE devices in redundancy group that can result in duplications (and even loops) because of multiple DFs if the timer is too short or blackholing if the timer is too long.

Using ESI label Split Horizon filtering can prevent loops (but not duplicates), however if there are overlapping DFs in two different sites at the same time for the same VLAN, the site identifier will be different upon re-entry of the packet and hence the split horizon check will fail, leading to L2 loops.

The current state of art (Highest Random Weight) algorithm to avoid reshuffling of VLANs among PE devices in the redundancy group upon failure/recovery and thus reducing the impact of failure/recovery to VLANs not on the failed/recovered ports. This eliminates loops/duplicates in failure scenarios.[RFC8584] uses the well known HRW

However, upon PE insertion or port bring-up, HRW cannot help as a transfer of DF role need to happen to the newly inserted device/port while the old DF is still active.

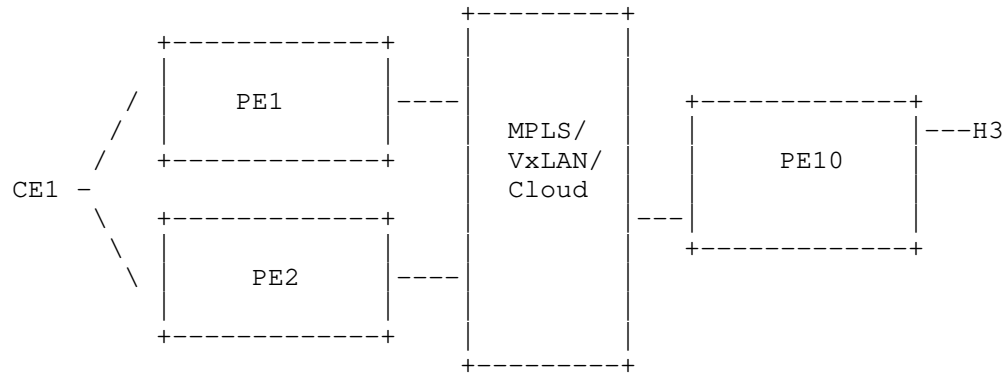


Figure 1: CE1 multi-homed to PE1 and PE2.

In the Figure 1, when PE2 is inserted or booted up, PE1 will transfer DF role of some VLANs to PE2 to achieve load balancing. However, because there is no handshake mechanism between PE1 and PE2, duplication of DF roles for a give VLAN is possible. Duplication of DF roles may eventually lead to L2 loops as well as duplication of traffic.

Current state of EVPN art relies on a blackholing timer for transferring the DF role to the newly inserted device. This can cause the following issues:

- * Loops/Duplicates if the timer value is too short
- * Prolonged Traffic Blackholing if the timer value is too long

3. DF Election Synchronization Solution

The solution relies on the concept of common clock alignment between partner PEs participating to a common Ethernet-Segment. The main idea is to have them all to perform/apply their carving state, resulting from DF election, at the well-known time.

The DF Election procedure, as described in [RFC7432] and as optionally signalled in [RFC8584], is applied. All PEs attached to a given Ethernet-Segment are clock-synchronized; using a networking protocol for clock synchronization (e.g. NTP, PTP, etc.). Newly inserted device PE or during failure recovery of a PE, that PE

communicates the current time to peering partners plus the remaining peering timer time left. This constitute an "end" or "absolute" time as seen from local PE. That absolute time is called "Service Carving Time" (SCT).

A new BGP Extended Community is advertised along with Ethernet-Segment route (RT-4) to communicate to other partners the Service Carving Time.

Upon reception of that new BGP Extended Community, partner PEs know exactly its carving time. The notion of skew is introduced to eliminate any potential duplicate traffic or loops. They add a skew (default = -10ms) to the Service Carving Time to enforce this. The previously inserted PE(s) must carve first, followed shortly(skew) by the newly insterted PE.

To summarize, all peering PEs carve almost simultaneously at the time announced by newly added/recovered PE. The newly inserted PE initiates the SCT, and carves immediately on peering timer expiry. The previously inserted PE(s) receiving Ethernet-Segment route (RT-4) with a SCT BGP extended community, carve shortly before Service Carving Time.

3.1. Advantages

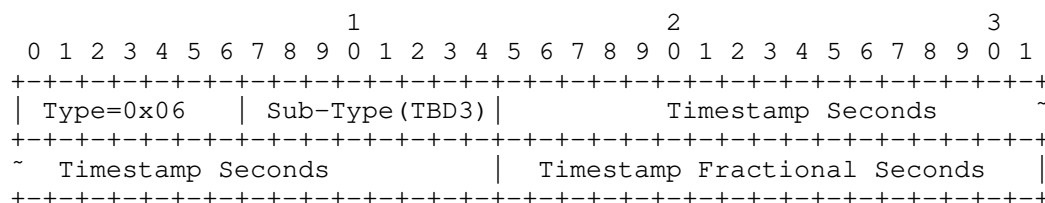
There are multiples advantages of using the approach. Here is a non-exhaustive list:

- A simple uni-directional signaling is all needed
- Backwards-compatible: PEs supporting only older [RFC7432] shall simply discard unrecognized new "Service Carving Timestamp" BGP Extended Community
- Multiple DF Election algorithms can be supported:
 - * [RFC7432] default ordered list ordinal algorithm (Modulo),
 - * [RFC8584] highest-random weight, etc.
- Independent of BGP transmission delay regarding Ethernet-Segment route (RT-4)
- Agnostic of the time synchronization mechanism used (e.g .NTP, PTP, etc.)

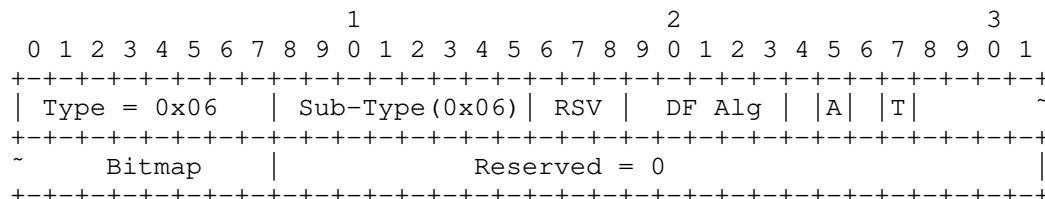
3.2. BGP Encoding

A new BGP extended community needs to be defined to communicate the Service Carving Timestamp for each Ethernet Segment.

A new transitive extended community where the Type field is 0x06, and the Sub-Type is [TBD3] is advertised along with Ethernet Segment route. Timestamp for expected Service carving is encoded as a 8-octet value as follows:



This document introduces a new flag called "T" (for Time Synchronization) to the bitmap field of the DF Election Extended Community defined in [RFC8584].



T: This flag is located in bit position 27 as shown above. When set to 1, it indicates the desire to use Time Synchronization capability with the rest of the PEs in the ES. This capability is used in conjunction with the agreed upon DF Type (DF Election Type). For example if all the PEs in the ES indicated that they have Time Synchronization capability and they want the DF type be of HRW, then HRW algorithm is used in conjunction with this capability.

3.3. Note on NTP-based synchronization

The 64-bit timestamp used by NTP protocol consists of a 32-bit part for seconds and a 32-bit part for fractional second. The timestamp exchanged uses the NTP epoch of January 1, 1900 [RFC5905]. The use of a 32-bit seconds and 16-bit fractional seconds yields adequate precision of 15 microseconds (2^{-16} s).

3.4. Synchronization Scenarios

Let's take Figure 1 as an example where initially PE2 had failed and PE1 had taken over. This example shows the problem with known mechanism.

Based on [RFC7432]:

- Initial state: PE1 is in steady-state, PE2 is recovering
- PE2 recovers at (absolute) time $t=99$
- PE2 advertises RT-4 (sent at $t=100$) to partner PE1
- PE2, it starts its 3sec peering timer as per RFC7432
- PE1 carves immediately on RT-4 reception, i.e. $t=100$ + minimal BGP propagation delay
- PE2 carves at time $t=103$

[RFC7432] aims of favouring traffic black hole over duplicate traffic. With above procedure, traffic black hole will occur as part of each PE recovery sequence. The peering timer value (default = 3 seconds) has a direct effect on the duration of the prolonged blackholing. A short (esp. zero) peering timer may, however, result in duplicate traffic or traffic loops.

Based on the Service Carving Time (SCT) approach:

- Initial state: PE1 is in steady-state, PE2 is recovering
- PE2 recovers at (absolute) time $t=99$
- PE2 advertises RT-4 (sent at $t=100$) with target SCT value $t=103$ to partner PE1
- PE2 starts its 3 second peering timer as per [RFC7432]
- Both PE1 and PE2 carves at (absolute) time $t=103$

In fact, PE1 should carve slightly before PE2 (skew). The previously inserted PE2 that is recovering performs both transitions DF to NDF and NDF to DF per VLANs at the peering timer expiry. Since the goal is to prevent duplicates, the original PE1, which received the SCT will apply:

- DF to NDF transition at $t=\text{SCT} - \text{skew}$ where both PEs are NDF

for 'skew' amount of time

- NDF to DF transition at $t=SCT$

It is this split-behaviour which ensures good transition of DF role with contained amount of loss.

Using SCT approach, the negative effect of the peering timer is mitigated. Furthermore, the BGP Ethernet-Segment route (RT-4) transmission delay (from PE2 to PE1) becomes a no-op. The usage of SCT approach remedies to the exposed problem with the usage of peering timer. The 3 seconds timer window is shorthen to few milliseconds.

3.5. Backwards Compatibility

Per redundancy group, for the DF election procedures to be globally convergent and unanimous, it is necessary that all the participating PEs agree on the DF Election algorithm to be used. It is, however, possible that some PEs continue to use the existing modulus based DF election and do not rely on the new SCT BGP extended community. PEs running an baseline DF election mechanism shall simply discard unrecognized new SCT BGP extended community.

A PE can indicate its willingness to support clock-synched carving by signaling the new 'T' DF Election Capability as well as including the new Service Carving Time BGP extended community along with the Ethernet-Segment Route (Type-4). In the case where one or more PEs attached to the Ethernet-Segment do not signal $T=1$, all PEs in the Ethernet-Segment may revert back to the RFC7432 timer approach.

4. Security Considerations

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

5. IANA Considerations

This document solicits the allocation of the following sub-type in the "EVPN Extended Community Sub-Types" registry setup by [RFC7153]:

TBD3	Service Carving Timestamp	This document
------	---------------------------	---------------

This document solicits the allocation of the following values in the "DF Election Capabilities" registry setup by [RFC8584]:

Bit ----	Name -----	Reference -----
3	Time Synchronization	This document

6. 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>>.
- [RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <<https://www.rfc-editor.org/info/rfc5905>>.
- [RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <<https://www.rfc-editor.org/info/rfc7153>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [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>>.
- [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>>.

Appendix A. Contributors

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

Gaurav Badoni Cisco Email: gbadoni@cisco.com

Dhananjaya Rao Cisco Email: dhrao@cisco.com

Appendix B. Acknowledgements

Authors would like to acknowledge helpful comments and contributions of Satya Mohanty and Bharath Vasudevan.

Authors' Addresses

Patrice Brissette (editor)
Cisco

Email: pbrisset@cisco.com

Ali Sajassi
Cisco

Email: sajassi@cisco.com

Luc Andre Burdet
Cisco

Email: lburdet@cisco.com

John Drake
Juniper

Email: jdrake@juniper.net

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com

BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2022

J. Rabadan, Ed.
J. Kotalwar
S. Sathappan
Nokia
Z. Zhang
W. Lin
Juniper
E. Rosen
Individual
July 12, 2021

Multicast Source Redundancy in EVPN Networks
draft-ietf-bess-evpn-redundant-mcast-source-01

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

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

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

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

This Internet-Draft will expire on January 13, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Terminology	4
1.2. Background on IP Multicast Delivery in EVPN Networks . .	6
1.2.1. Intra-subnet IP Multicast Forwarding	6
1.2.2. Inter-subnet IP Multicast Forwarding	7
1.3. Multi-Homed IP Multicast Sources in EVPN	9
1.4. The Need for Redundant IP Multicast Sources in EVPN . . .	11
2. Solution Overview	11
3. BGP EVPN Extensions	13
4. Warm Standby (WS) Solution for Redundant G-Sources	14
4.1. WS Example in an OISM Network	16
4.2. WS Example in a Single-BD Tenant Network	18
5. Hot Standby (HS) Solution for Redundant G-Sources	19
5.1. Extensions for the Advertisement of DCB Labels	22
5.2. Use of BFD in the HS Solution	23
5.3. HS Example in an OISM Network	24
5.4. HS Example in a Single-BD Tenant Network	28
6. Security Considerations	28
7. IANA Considerations	28
8. References	28
8.1. Normative References	28
8.2. Informative References	29
Appendix A. Acknowledgments	30
Appendix B. Contributors	30
Authors' Addresses	30

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.

- o PIM: Protocol Independent Multicast.
- o MVPN: Multicast Virtual Private Networks.
- o OISM: Optimized Inter-Subnet Multicast, as in [I-D.ietf-bess-evpn-irb-mcast].
- o 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.
- o 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.
- o 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).
- o 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.
- o G-traffic: any frame with an IP payload whose IP Destination Address (IP DA) is a multicast group G.
- o G-source: any system sourcing IP multicast traffic to G.

- o 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):

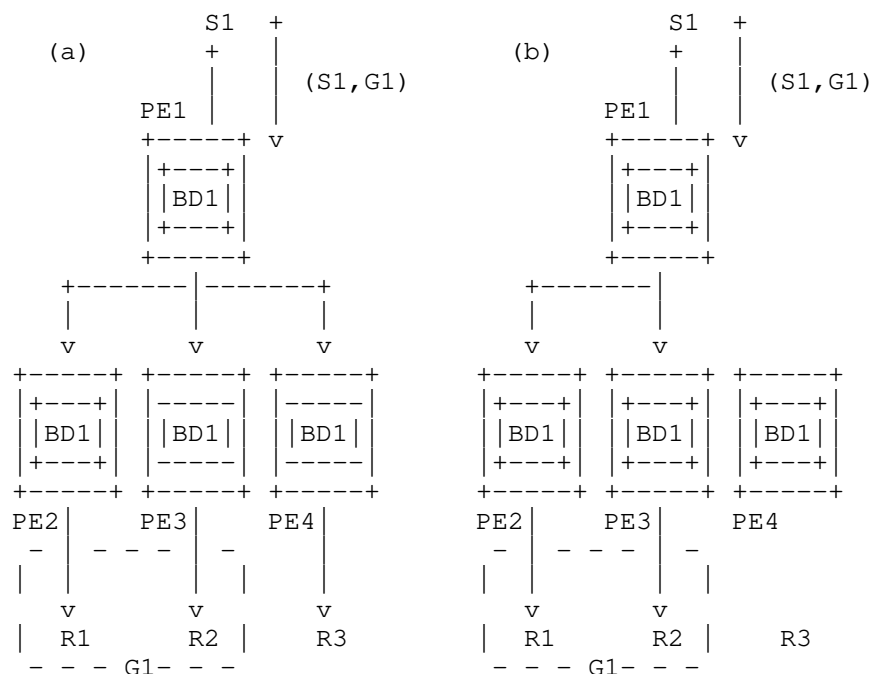


Figure 1: Intra-subnet IP Multicast

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.

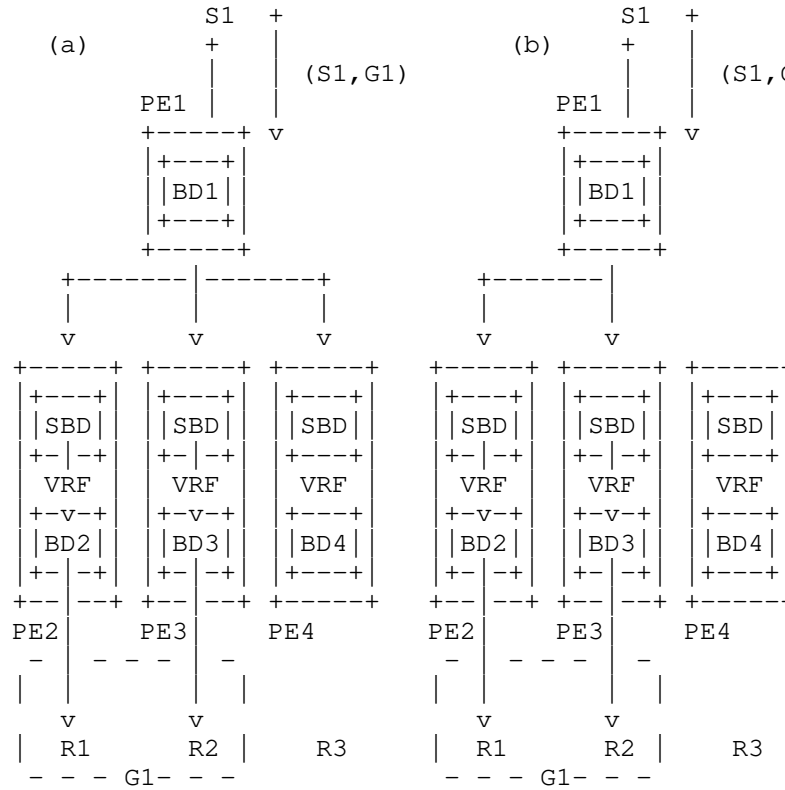


Figure 2: Inter-subnet IP Multicast

[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.

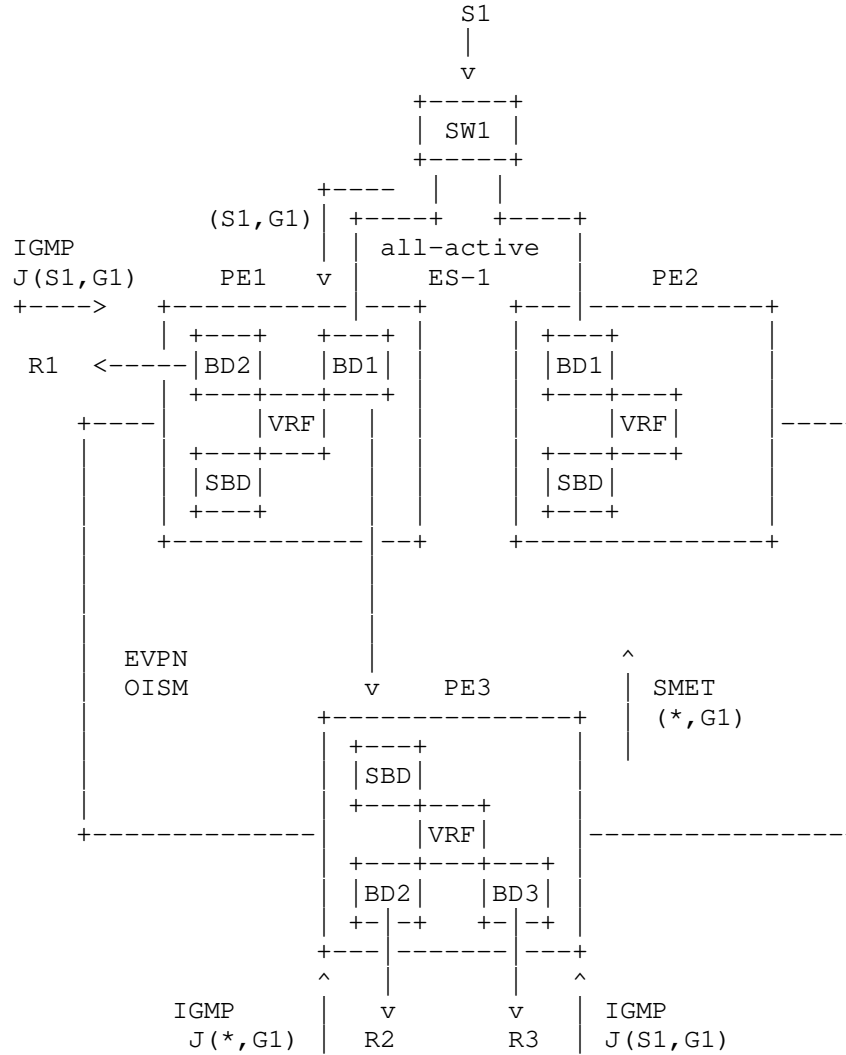


Figure 3: All-active Multi-homing and OISM

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:

- o Warm Standby (WS) Solution
- o 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:

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

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

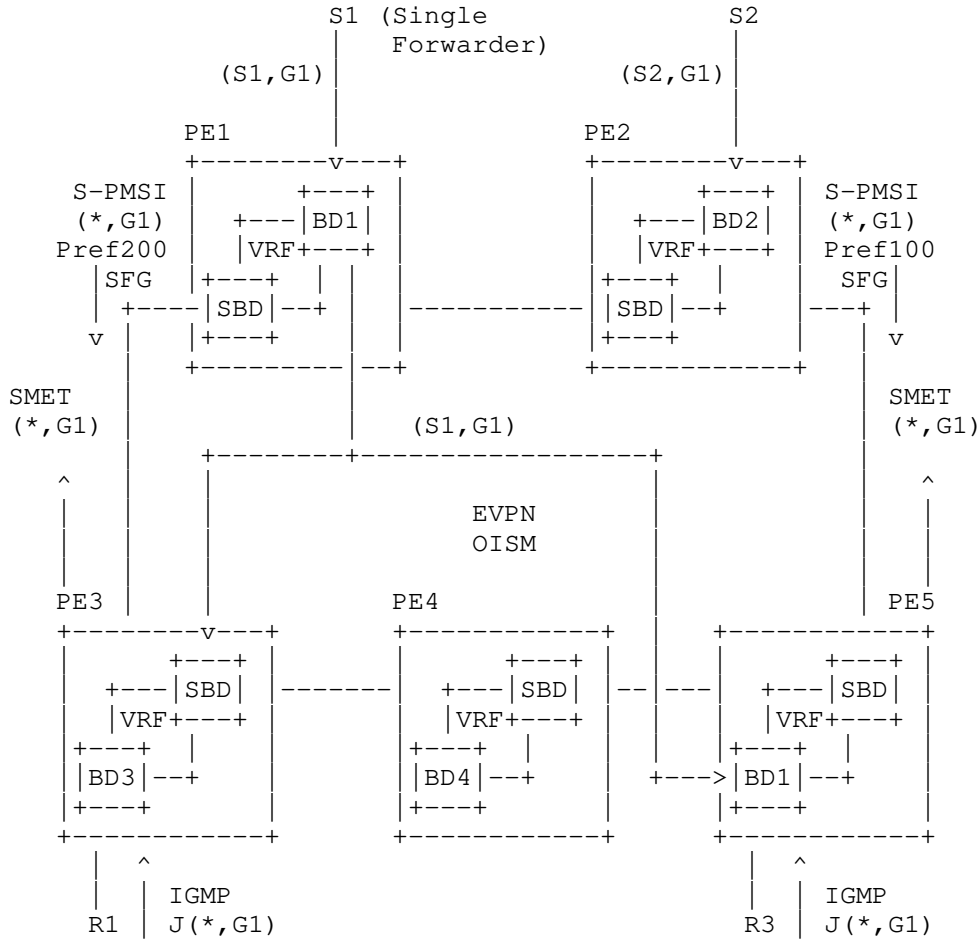


Figure 4: WS Solution for Redundant G-Sources

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.

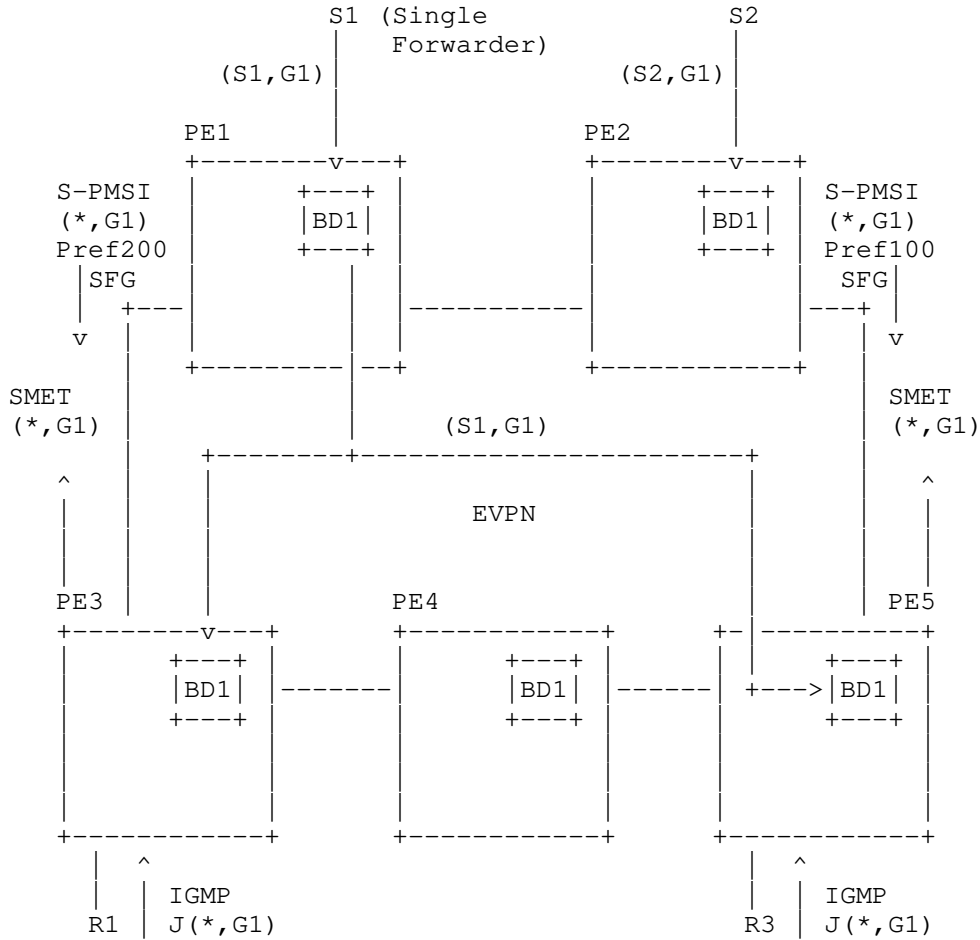


Figure 5: WS Solution for Redundant G-Sources in the same BD

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 (Output 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 as described in [I-D.zzhang-bess-mvpn-evpn-aggregation-label]). 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. Extensions for the Advertisement of DCB Labels

DCB Labels are specified in [I-D.zzhang-bess-mvpn-evpn-aggregation-label] and this document makes use of them for the procedures described in Section 5. [I-D.zzhang-bess-mvpn-evpn-aggregation-label] assumes that DCB labels can only be used along with MP2MP/P2MP/BIER tunnels and that, if the

PMSI label is signaled as a DCB label, then the ESI label used for multi-homing is also a DCB label. This document extends the use of the DCB allocation for ESI labels so that:

- a. DCB-allocated ESI labels MAY be used along with IR tunnels, and
- b. DCB-allocated ESI labels MAY be used by PEs that do not use DCB-allocated PMSI labels.

This control plane extension is indicated by adding the DCB-flag or the Context Label Space ID Extended Community to the A-D per ES route(s) advertised for the S-ES. The DCB-flag is encoded in the ESI Label Extended Community as follows:

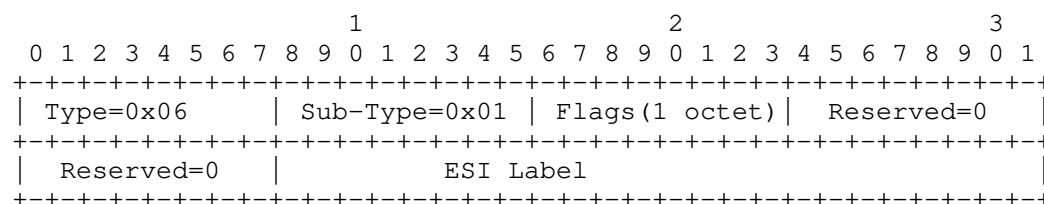


Figure 6: ESI Label Extended Community

This document defines the bit 5 in the Flags octet of the ESI Label Extended Community as the ESI-DCB-flag. When the ESI-DCB-flag is set, it indicates that the ESI label is a DCB label. Instead of the DCB-flag, the A-D per ES route can also carry a Context Label Space ID Extended Community, as described in [I-D.zzhang-bess-mvpn-evpn-aggregation-label], only for A-D per ES routes in this document, as opposed to x-PMSI/IMET routes.

5.2. 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.

5.3. HS Example in an OISM Network

Figure 7 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.

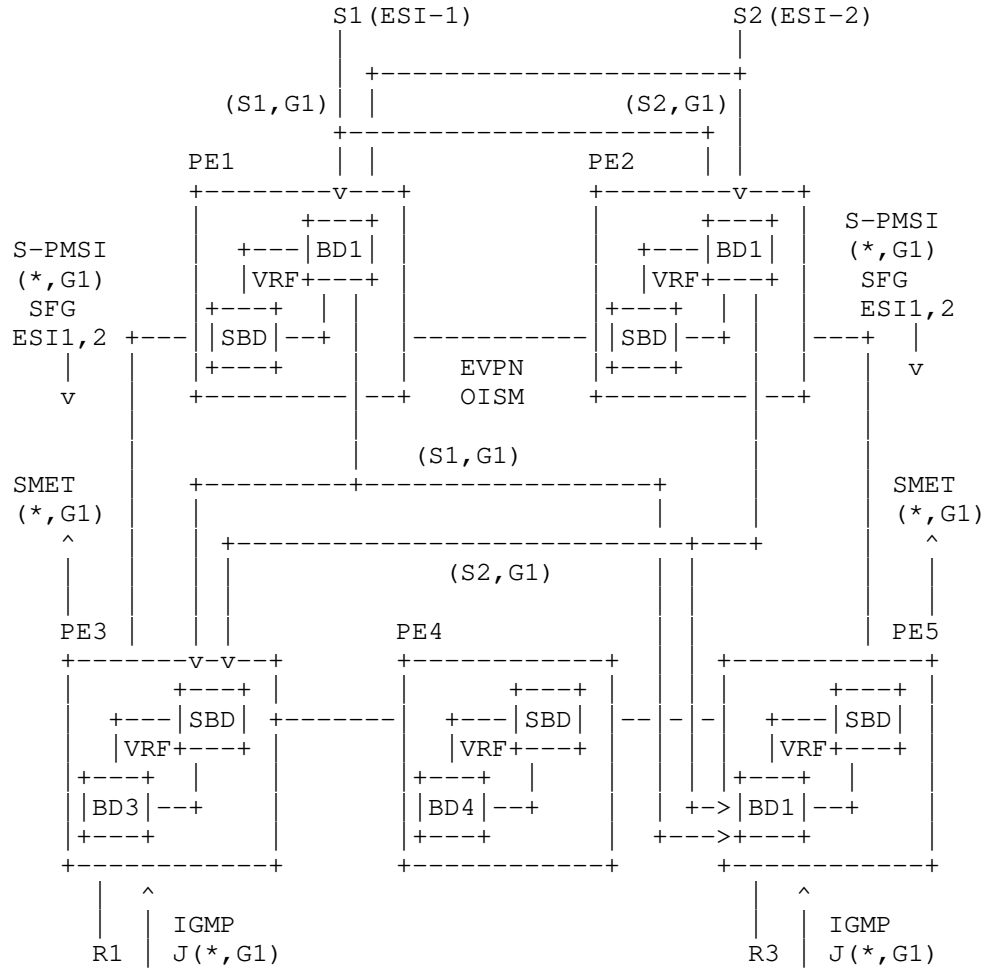


Figure 7: HS Solution for Multi-homed Redundant G-Sources in OISM

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 8 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 7. 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 7 applies to this use-case.

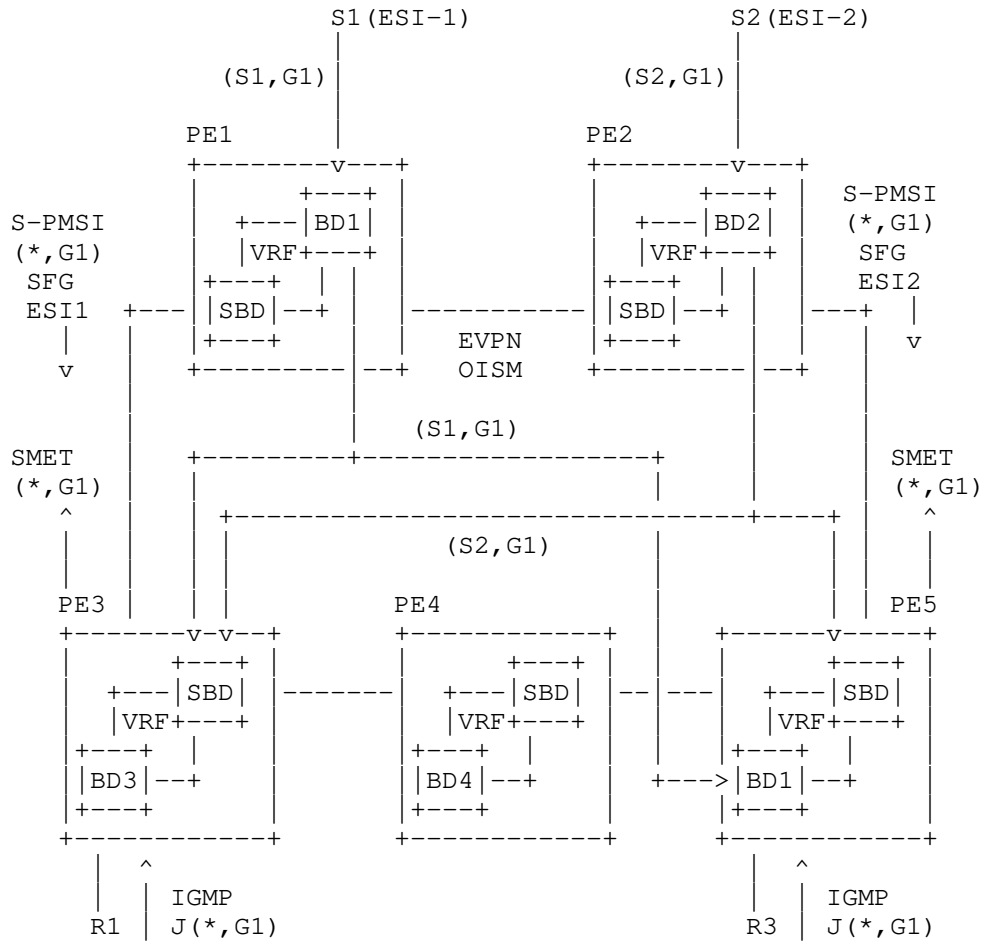


Figure 8: HS Solution for single-homed Redundant G-Sources in OISM

5.4. 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., Mishra, M., Patel, K., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-09 (work in progress), April 2021.

- [I-D.ietf-bess-evpn-irb-mcast]
Lin, W., Zhang, Z., Drake, J., Rosen, E. C., 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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 22 April 2022

R.P. Parekh
C. Filsfils
A.V. Venkateswaran
Cisco Systems, Inc.
H. Bidgoli
Nokia
D. Voyer
Bell Canada
Z. Zhang
Juniper Networks
19 October 2021

Multicast and Ethernet VPN with Segment Routing P2MP
draft-ietf-bess-mvpn-evpn-sr-p2mp-04

Abstract

A Point-to-Multipoint (P2MP) Tree in a Segment Routing domain carries traffic from a Root to a set of Leaves. This document describes extensions to BGP encodings and procedures for P2MP trees and Ingress Replication used in BGP/MPLS IP VPNs and Ethernet VPNs in a Segment Routing domain.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on 22 April 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. SR P2MP P-Tunnels	3
3. PMSI Tunnel Attribute for SR P2MP	4
3.1. MPLS Label	5
3.1.1. SR-MPLS	5
4. MVPN Auto-Discovery and Binding Procedures for P2MP Trees . .	5
4.1. Intra-AS I-PMSI	6
4.1.1. Originating Intra-AS I-PMSI routes	6
4.1.2. Receiving Intra-AS I-PMSI A-D routes	6
4.2. Using S-PMSIs for binding customer flows to P2MP Segments	7
4.2.1. Originating S-PMSI A-D routes	7
4.2.2. Receiving S-PMSI A-D routes	8
4.3. Inter-AS P-tunnels using P2MP Segments	9
4.3.1. Advertising Inter-AS I-PMSI routes into iBGP	9
4.3.2. Receiving Inter-AS I-PMSI A-D routes in iBGP	9
4.4. Leaf A-D routes for P2MP Segment Leaf Discovery	9
4.4.1. Originating Leaf A-D routes	9
4.4.2. Receiving Leaf A-D routes	10
5. MVPN with Ingress Replication over Segment Routing	10
5.1. SR-MPLS	10
5.2. SRv6	11
5.2.1. SRv6 Multicast Endpoint Behaviors	12
6. Dampening of MVPN routes	12
7. SR P2MP Trees for EVPN	13
8. IANA Considerations	13
9. Security Considerations	14
10. Acknowledgements	14
11. Contributors	14
12. References	15
12.1. Normative References	15
12.2. Informative References	16

Authors' Addresses	17
--------------------	----

1. Introduction

Multicast in MPLS/BGP IP VPNs [RFC6513] and BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs [RFC6514] specify procedures that allow a Service Provider to provide Multicast VPN (MVPN) service to its customers. Multicast traffic from a customer is tunneled across the service provider network over Provider Tunnels (P-Tunnels). P-Tunnels can be instantiated via different technologies. A service provider network that uses Segment Routing can use a Point-to-Multipoint (SR P2MP) tree [I-D.ietf-pim-sr-p2mp-policy] or P2MP Ingress Replication to instantiate P-Tunnels for MVPN. SR P2MP P-Tunnels can be realized both for SR-MPLS [RFC8660] and SRv6 [RFC8986][RFC8754].

In a Segment Routing network, a P2MP tree allows efficient delivery of traffic from a Root to set of Leaf nodes. A SR P2MP tree is defined by a SR P2MP Policy and instantiated via a PCE. A P2MP Policy consists of a Root, a Set of Leaf Nodes and a set of candidate paths with optional set of constraints and/or optimization objectives to be satisfied by the P2MP tree. A unique Identifier, called Tree-SID, is associated with a P2MP tree. This Tree-SID can be an MPLS label or an IPv6 address.

This document describes extensions to BGP Auto-Discovery procedures specified in RFC 6514 for SR P2MP P-Tunnels. Use of PIM for Auto-Discovery is outside scope of this document. Support for customer BIDIR-PIM is outside the scope of this document.

For BGP MPLS Ethernet VPN specified in [RFC7432] and extensions to this document, P-Tunnels are advertised for handling multi-destination traffic. These P-Tunnels can be realized by SR-MPLS or SRv6 P2MP trees. SRv6 P2MP trees can also be used to support Multicast in Network Virtualization over Layer 3 [RFC8293].

The reader is expected to be familiar with concepts and terminology of RFC 6513, RFC 6514 and SR P2MP drafts.

2. SR P2MP P-Tunnels

For MVPN or EVPN, Provider Edge(PE) routers steer customer traffic into a P-Tunnel that can be instantiated by a SR-MPLS or SRv6 P2MP. A SR P2MP tree is defined by a SR P2MP policy [I-D.ietf-pim-sr-p2mp-policy].

Given a SR P2MP policy, a PCE computes and instantiates the SR P2MP tree on the nodes that are part of the tree by stitching Replication segments [I-D.ietf-spring-sr-replication-segment] at Root, Leaf and intermediate replication nodes. Tree-SID is a unique identifier for the tree. A Replication segment of a SR P2MP tree can be initiated by various methods (BGP, PCEP, others) which are outside the scope of this document.

A PCE provides conceptual APIs, listed below, to define and modify SR P2MP policies SR P2MP Policy Section 4.1.1 (<https://tools.ietf.org/html/draft-ietf-pim-sr-p2mp-policy-00#section-4.1.1>). These APIs are invoked by a PCC, which is the root of P2MP tree, using various methods (BGP, PCEP, etc.) which are outside the scope of this document.

CreatePolicy: CreateSRP2MPPolicy<Root, Tree-ID>

DeletePolicy: DeleteSRP2MPPolicy<Root, Tree-ID>

UpdateLeafSet: SRP2MPPolicyLeafSetModify<Root, Tree-ID, {Leaf Set}>

The Root of a P2MP tree imposes the Tree-SID to steer the customer payload into the P2MP tree. Provider (P) routers replicate customer payload, using Replication segments, towards the Leaf nodes of the P2MP tree. Leaf nodes of the P2MP tree deliver the customer payload after disposing the Tree-SID.

An Ingress PE can deliver payload to egress PEs of the service using Ingress Replication. This payload is encapsulated in SR-MPLS or SRv6 and replicated to each egress PE.

3. PMSI Tunnel Attribute for SR P2MP

BGP PMSI Tunnel Attribute (PTA) is defined in RFC 6514 to identify the P-Tunnel that is used to instantiate a Provider Multicast Service Interface (PMSI). The PTA is carried in Intra-AS I-PMSI, Inter-AS I-PMSI, Selective PMSI, and Leaf Auto-Discovery routes.

A P2MP tree PTA is constructed as specified below.

- * Tunnel Type: The IANA assigned codepoint 0x0C for "SR-MPLS P2MP Tree" or codepoint
// TBD
- * Flags: See Section 4 for use of "Leaf Info Required bit".
- * MPLS Label: See Section 3.1

- * Tunnel Identifier: The SR P2MP P-Tunnel is identified by <Tree-ID, Root> where,
 - Tree-ID is a 32-bit unsigned value that identifies a unique P2MP tree at a Root.
 - Root is an IP address identifying the Root of a P2MP tree. This can be either an IPv4 or IPv6 address and can be inferred from the PTA length.

When a P-Tunnel is non-segmented, the PTA is created by PE router at the Root of a SR P2MP tree. For segmented P-Tunnels, each segment can be instantiated by a different technology. If a segment is instantiated using P2MP tree, the router at the root of a P2MP tree creates the PTA.

3.1. MPLS Label

[RFC6514] allows a PE to aggregate two or more MVPNs onto one P-Tunnel by advertising the same P-Tunnel in PTA of Auto-Discovery routes of different MVPNs. This section specifies how the "MPLS Label" field of PTA is filled to provide a context bound to a specific MVPN. Aggregating MVPNs on one SRv6 P2MP P-Tunnel will be addressed in future revision of this document. For EVPN considerations, see SR P2MP Trees for EVPN section.

3.1.1. SR-MPLS

When a SR P2MP P-Tunnel, shared across different MVPNs, is instantiated in a SR MPLS domain [RFC8660], "MPLS Field" of a PTA advertised in a Auto-Discovery route MUST contain an upstream-assigned MPLS label that the advertising PE has bound to the MVPN, or a label assigned from a global context such as "Domain- wide Common Block" (DCB) as specified in [I-D.ietf-bess-mvpn-evpn-aggregation-label].

When a customer payload is steered into a shared SR P2MP P-Tunnel, this MPLS label MUST be imposed before the MPLS label representing the Tree-SID.

4. MVPN Auto-Discovery and Binding Procedures for P2MP Trees

RFC 6514 defines procedures for discovering PEs participating in a given MVPN and binding customer multicast flows to specific P-Tunnels. This section specifies modifications to these procedures for SR P2MP tree P-Tunnels. In this section, the term "SR P2MP" refers to both SR-MPLS and SRv6.

4.1. Intra-AS I-PMSI

Intra-AS I-PMSI A-D routes are exchanged to discover PEs participating in a MVPN within an AS, or across different ASes when non-segmented P-Tunnels are used for inter-AS MVPNs.

4.1.1. Originating Intra-AS I-PMSI routes

RFC 6514 Section 9.1.1 (<https://tools.ietf.org/html/rfc6514#section-9.1.1>) describes procedures for originating Intra-AS I-PMSI A-D routes. For SR P2MP P-Tunnels, these procedures remain unchanged except as described in the following paragraphs.

When a PE originates an Intra-AS I-PMSI A-D route with a PTA having SR P2MP P-Tunnel Type, it MUST create a P2MP policy by invoking CreatePolicy API of the PCE. When the PCE instantiates the P2MP tree on the PE, the Tree-SID MUST be imposed for customer flow(s) steered into the P2MP tree. The Leaf nodes of P2MP tree are discovered using procedures described in Section 4.1.2.

For a PE in "Receiver Sites set", condition (c) is modified to include P2MP tree; such a PE MUST originate an Intra-AS I-PMSI A-D route when some PEs of the MVPN have VRFs that use SR P2MP tree but MUST NOT create a SR P2MP policy as described above.

When a PE withdraws an Intra-AS I-PMSI A-D route, advertised with a PTA having SR P2MP P-Tunnel Type, the Tree-SID imposition state at the PE MUST be removed.

A PE MAY aggregate two or more Intra-AS I-PMSIs from different MVPNs onto the same SR P2MP P-Tunnel. When a PE withdraws the last Intra-AS I-PMSI A-D route, advertised with a PTA identifying a SR P2MP P-Tunnel, it SHOULD remove the SR P2MP policy by invoking DeletePolicy API of the PCE.

4.1.2. Receiving Intra-AS I-PMSI A-D routes

Procedure for receiving Intra-AS I-PMSI A-D routes, as described in RFC 6514 Section 9.1.2 (<https://tools.ietf.org/html/rfc6514#section-9.1.2>), remain unchanged for SR P2MP P-Tunnels except as described in the following paragraphs.

When a PE that advertises a SR P2MP P-Tunnel in the PTA of its Intra-AS I-PMSI A-D route, imports an Intra-AS I-PMSI A-D route from some PE, it MUST add that PE as a Leaf node of the P2MP tree. The Originating IP Address of the Intra-AS i-PMSI A-D route is used as the Leaf Address when invoking UpdateLeafSet API of the PCE. This procedure MUST also be followed for all Intra-AS I-PMSI routes that are already imported when the PE advertises a SR P2MP P-Tunnel in PTA of its Intra-AS I-PMSI A-D route.

A PE that imports and processes an Intra-AS I-PMSI A-D route from another PE with PTA having SR P2MP P-Tunnel MUST program the Tree-SID of the P2MP tree identified in the PTA of the route for disposition. Note that an Intra-AS I-PMSI A-D route from another PE can be imported before the P2MP tree identified in the PTA of the route is instantiated by the PCE at the importing PE. In such case, the PE MUST correctly program Tree-SID for disposition. A PE in "Sender Sites set" MAY avoid programming the Tree-SID for disposition.

When an Intra-AS I-PMSI A-D route, advertised with a PTA having SR P2MP P-Tunnel Type is withdrawn, a PE MUST remove the disposition state of the Tree-SID associated with P2MP tree.

A PE MAY aggregate two or more Intra-AS I-PMSIs from different MVPNs onto the same SR P2MP P-Tunnel. When a remote PE withdraws an Intra-AS I-PMSI A-D route from a MVPN, and if this is the last MVPN sharing a SR P2MP P-Tunnel, a PE must remove the originating PE as a Leaf from the P2MP tree, by invoking UpdateLeafSet API.

4.2. Using S-PMSIs for binding customer flows to P2MP Segments

RFC 6514 specifies procedures for binding (C-S,C-G) customer flows to P-Tunnels using S-PMSI A-D routes. Wildcards in Multicast VPN Auto-Discovery Routes [RFC6625] specifies additional procedures to binding aggregate customer flows to P-Tunnels using "wildcard" S-PMSI A-D routes. This section describes modification to these procedures for SR P2MP P-Tunnels.

4.2.1. Originating S-PMSI A-D routes

RFC 6514 Section 12.1 (<https://tools.ietf.org/html/rfc6514#section-12.1>) describes procedures for originating S-PMSI A-D routes. For SR P2MP P-Tunnels, these procedures remain unchanged except as described in the following paragraphs.

When a PE originates S-PMSI A-D route with a PTA having SR P2MP P-Tunnel Type, it MUST set the "Leaf Info Required bit" in the PTA. The PE MUST create a SR P2MP policy by invoking CreatePolicy API of the PCE. When the PCE instantiates the P2MP tree on the PE, the Tree-SID MUST be imposed for customer flows steered into the SR P2MP P-Tunnel.

The Leaf nodes of P2MP tree are discovered by Leaf A-D routes using procedures described in Section 4.4.2. When a PE originates S-PMSI A-D route with a PTA having SR P2MP P-Tunnel Type, it is possible the PE might have imported Leaf A-D routes whose route keys match the S-PMSI A-D route. The PE MUST re-apply procedures of Section 4.4.2 to these Leaf A-D routes.

When a PE withdraws a S-PMSI A-D route, advertised with PTA having P2MP tree P-Tunnel type, the Tree-SID imposition state MUST be removed.

A PE MAY aggregate two or more S-PMSIs onto the same SR P2MP P-Tunnel. When a PE withdraws the last S-PMSI A-D route, advertised with a PTA identifying a specific SR P2MP P-Tunnel, it SHOULD remove the SR P2MP policy by invoking DeletePolicy API of the PCE.

4.2.2. Receiving S-PMSI A-D routes

RFC 6514 Section 12.3 (<https://tools.ietf.org/html/rfc6514#section-12.3>) describes procedures for receiving S-PMSI A-D routes. For SR P2MP P-Tunnels, these procedures remain unchanged except as described in the following paragraphs.

The procedure to join SR P2MP P-Tunnel of S-PMSI A-D route by using a Leaf A-D route is described in Section 4.4.1. If P2MP tree identified in PTA of S-PMSI A-D route is already instantiated by PCE, the PE MUST program Tree-SID for disposition. If the P2MP tree is instantiated later, the Tree-SID MUST be programmed for disposition at that time.

When a S-PMSI A-D route, whose SR P2MP P-Tunnel has been joined by a PE, is withdrawn, or when conditions (see RFC 6514 Section 12.3 (<https://tools.ietf.org/html/rfc6514#section-12.3>)) required to join that P-Tunnel are no longer satisfied, the PE MUST leave the P-Tunnel. The PE MUST withdraw the Leaf A-D route it had originated and remove the Tree-SID disposition state.

4.3. Inter-AS P-tunnels using P2MP Segments

A segmented inter-AS P-Tunnel consists of one or more intra-AS segments, one in each AS, connected by inter-AS segments between ASBRs of different ASes <https://tools.ietf.org/html/rfc6514#section-9.2>. These segments are constructed by PEs/ASBRs originating or re-advertising Inter-AS I-PMSI A-D routes. This section describes procedures for instantiating intra-AS segments using SR P2MP trees.

4.3.1. Advertising Inter-AS I-PMSI routes into iBGP

RFC 6514 Section 9.2.3.2 (<https://tools.ietf.org/html/rfc6514#section-9.2.3.2>) specifies procedures for advertising an Inter-AS I-PMSI A-D route to construct an intra-AS segment. The PTA of the route identifies the type and identifier of the P-Tunnel instantiating the intra-AS segment. The procedure for creating SR P2MP P-Tunnel for intra-AS segment are same as specified in Section 4.2.1 except that instead of S-PMSI A-D routes, the procedures apply to Inter-AS I-PMSI A-D routes.

4.3.2. Receiving Inter-AS I-PMSI A-D routes in iBGP

RFC 6514 Section 9.2.3.2 (<https://tools.ietf.org/html/rfc6514#section-9.2.3.2>) specifies procedures for processing an Inter-AS I-PMSI A-D route received via iBGP. If the PTA of the Inter-AS I-PMSI A-D route has SR P2MP P-Tunnel Type, the procedures are same as specified in Section 4.2.2 except that instead of S-PMSI A-D routes, the procedures apply to Inter-AS I-PMSI A-D routes. If the receiving router is an ASBR, the Tree-SID is stitched to the inter-AS segments to ASBRs in other ASes.

4.4. Leaf A-D routes for P2MP Segment Leaf Discovery

This section describes procedures for originating and processing Leaf A-D routes used for Leaf discovery of SR P2MP trees.

4.4.1. Originating Leaf A-D routes

The procedures for originating Leaf A-D route in response to receiving a S-PMSI or Inter-AS I-PMSI A-D route with PTA having SR P2MP P-Tunnel Type are same as specified in RFC 6514 Section 9.2.3.4.1 (<https://tools.ietf.org/html/rfc6514#section-9.2.3.4.1>).

4.4.2. Receiving Leaf A-D routes

Procedures for processing a received Leaf A-D route are specified in RFC 6514 Section 9.2.3.5 (<https://tools.ietf.org/html/rfc6514#section-9.2.3.5>). These procedures remain unchanged for discovering Leaf nodes of P2MP trees except for considerations described in following paragraphs. These procedures apply to Leaf A-D routes received in response to both S-PMSI and Inter-AS I-PMSI A-D routes, shortened to "A-D routes" in this section

A Root PE/ASBR MAY use the same SR P2MP P-Tunnel in PTA of two or more A-D routes. For such aggregated P2MP trees, the PE/ASBR may receive multiple Leaf A-D routes from a Leaf PE. The P2MP tree for which a Leaf A-D is received can be identified by examining the P2MP tunnel Identifier in the PTA of A-D route that matches "Route Key" field of the Leaf A-D route. When the PE receives the first Leaf A-D route from a Leaf PE, identified by the Originating Router's IP address field, it MUST add that PE as Leaf of the P2MP tree by invoking the UpdateLeafSet API of the PCE.

When a Leaf PE withdraws the last Leaf A-D route for a given SR P2MP P-Tunnel, the Root PE MUST remove the Leaf PE from the P2MP tree by invoking UpdateLeafSet API of PCE. Note that Root PE MAY remove the P2MP tree, via the DeletePolicyAPI, before the last Leaf A-D is withdrawn. In this case, the Root PE MAY decide to not invoke the UpdateLeafSet API.

5. MVPN with Ingress Replication over Segment Routing

A PE can provide MVPN service using Ingress Replication over Segment Routing. Customer payload is encapsulated in SR-MPLS or IPv6 (SRv6) at Ingress PE. The encapsulated payload is replicated and a unicast copy is sent to each egress PE.

Ingress Replication Tunnels in Multicast VPN [RFC7988] specifies procedures that can be used to provide MVPN service with Ingress Replication in a Segment Routing domain. A PE advertises Intra-AS, Inter-AS, Selective PMSI BGP Auto-Discovery routes with PTA for Ingress Replication. Egress PEs join as Leaf Nodes using Intra-AS I-PMSI or Leaf Auto-Discovery routes.

5.1. SR-MPLS

Procedures of RFC 7988 are sufficient to create a SR-MPLS Ingress Replication for MVPN service.

5.2. SRv6

Procedures of RFC 7988, along with modifications described in this Section, are sufficient to create a SRv6 Ingress Replication for MVPN service.

The PTA carried in Intra-AS, Inter-AS, Selective PMSI and Leaf Auto-Discovery routes is constructed as specified in RFC 7988 with modifications as below:

- * Tunnel Type: "Ingress Replication" as per RFC 6514.
- * MPLS Label: The high order 20 bits of this field carry the whole or a portion of the Function part of the SRv6 Multicast Service SID when ingress replication is used and the Transposition Scheme of encoding as defined in Section 4 of SRv6 BGP based Overlay Services (<https://datatracker.ietf.org/doc/html/draft-ietf-bess-srv6-services-07#section-4>) is used. Otherwise, it is set as defined in RFC 6514. When using the Transposition Scheme, the Transposition Length MUST be less than or equal to 20 and less than or equal to the Function Length.

Section 6 and 7 of RFC 7988 (<https://datatracker.ietf.org/doc/html/rfc7988#section-6>) describe considerations and procedures for allocating MPLS labels for IR P-Tunnel. For SRv6 Ingress Replication, these sections apply to SRv6 Multicast Service SID.

To join a SRv6 Ingress Replication P-Tunnel advertised in PTA of Intra-AS, Inter-AS, or Selective S-PMSI A-D routes, an egress PE constructs a Leaf A-D or Intra-AS I-PMSI route as described in RFC 7988 with modified PTA above. The egress PE attaches a BGP Prefix-SID attribute [RFC8669] in Leaf A-D or Intra-AS I-PMSI route with SRv6 L3 Service TLV [I-D.ietf-bess-srv6-services] to signal SRv6 Multicast Service SID. The SRv6 SID Information Sub-TLV carries the SRv6 Multicast Service SID in SRv6 SID Value field. The SRv6 Endpoint Behavior of the SRv6 SID Information Sub-TLV encodes one of End.DTM4, End.DTM6, or End.DTM46 codepoint value. The SRv6 SID Structure Sub-Sub-TLV encodes the structure of SRv6 Multicast Service SID. If Transposition scheme is used, the offset and length of SRv6 Multicast Endpoint function of SRv6 Multicast Service SID is set in Transposition Length and Transposition Offset fields of this sub-sub TLV. Otherwise, the Transposition Length and Offset fields MUST be set to zero.

The BGP Prefix SID attribute with SRv6 L3 Service TLV in Intra-AS I-PMSI or Leaf A-D route indicates to ingress PE that egress PE supports SRv6. The ingress PE MUST encapsulate payload in an outer IPv6 header with the SRv6 Multicast Service SID provided by the

egress PE as the destination address. If Transposition scheme is used, ingress PE MUST merge Function in MPLS field of PTA with SRv6 SID in SID Information TLV using the Transposition Offset and Length fields from SID structure sub-sub TLV to create SRv6 Multicast Service SID

5.2.1. SRv6 Multicast Endpoint Behaviors

The following behaviors can be associated with SRv6 Multicast Service SID.

5.2.1.1. End.DTM4: Decapsulation and Specific IPv4 Multicast Table Lookup

The "Endpoint with decapsulation and specific IPv4 Multicast table lookup" behavior ("End.DTM4" for short) is similar to End.DT4 behavior of RFC 8986 except the lookup is in IPv4 multicast table.

5.2.1.2. End.DTM6: Decapsulation and Specific IPv6 Multicast Table Lookup

The "Endpoint with decapsulation and specific IPv6 Multicast table lookup" behavior ("End.DTM6" for short) is similar to End.DT6 behavior of RFC 8986 except the lookup is in IPv6 multicast table.

5.2.1.3. End.DTM46: Decapsulation and Specific IP Multicast Table Lookup

The "Endpoint with decapsulation and specific IP Multicast table lookup" behavior ("End.DTM46" for short) is similar to End.DT4 and End.DT6 behaviors of RFC 8986 except the lookup is in IP multicast table.

6. Dampening of MVPN routes

When P2MP trees are used as P-Tunnels for S-PMSI A-D routes, change in group membership of receivers connected to PEs has direct impact on the Leaf node set of a P2MP tree. If group membership changes frequently for a large number of groups with a lot of receivers across sites connected to different PEs, it can have an impact on the interaction between PEs and the PCE.

Since Leaf A-D routes are used to discover Leaf PE of a P2MP tree, it is RECOMMENDED that PEs SHOULD damp Leaf A-D routes as described in Section 6.1 of RFC 7899 [RFC7899]. PEs MAY also implement procedures for damping other Auto-Discovery and BGP C-multicast routes as described in [RFC7899].

7. SR P2MP Trees for EVPN

BGP MPLS Ethernet VPN specified in RFC 7432 specifies Inclusive Multicast Ethernet Tag route to support Broadcast, Unknown Unicast and Multicast (BUM) traffic. This IMET route is the equivalent of MVPN Intra-AS I-PMSI route and is advertised with a PMSI Tunnel Attribute (PTA) as specified in RFC 6514 to advertise the inclusive P-Tunnels.

[I-D.ietf-bess-evpn-bum-procedure-updates] updates BUM procedures to support selective P-Tunnels and P-Tunnel segmentation in EVPN. That document specifies new route types that are advertised with PTA, including Selective PMSI (S-PMSI) Auto-Discovery route.

These inclusive/selective P-Tunnels can be realized by SR P2MP trees. As with other types of P2MP P-Tunnels, the ESI label used for split horizon MUST be either upstream assigned by PE advertising the IMET or S-PMSI route, or assigned from a global context such as "Domain-wide Common Block" (DCB) as specified in [I-D.ietf-bess-mvpn-evpn-aggregation-label].

[I-D.ietf-bess-evpn-irb-mcast] specifies procedures to support Inter-Subnet Multicast. [I-D.ietf-bess-evpn-mvpn-seamless-interop] specifies how MVPN SAFI routes can be used to support Inter-Subnet Multicast. The P-Tunnels advertised in PTA of either EVPN and MVPN routes as specified in these documents respectively can be realized by SR P2MP trees.

SRv6 P2MP trees can serve as an underlay multicast as described in RFC 8293 Section 3.4 (<https://tools.ietf.org/html/rfc8293#section-3.4>). A NVE encapsulates a tenant packet in an SRv6 header and deliver it over SRv6 P2MP trees to other NVEs.

The same procedures specified for MVPN are used to collect the leaf information of corresponding SR P2MP tree (either based on IMET route or Leaf A-D routes in response to x-PMSI routes), to pass the tree information to the PCE controller, and to get back tree forwarding state used for customer multicast traffic forwarding.

8. IANA Considerations

IANA has assigned the value 0x0C for "SR-MPLS P2MP Tree" in the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry <https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#pmsi-tunnel-types> [RFC 7538] in the "Border Gateway Protocol (BGP) Parameters" registry.

IANA is requested to assign codepoint for "SRv6 P2MP Tree" in the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry <https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#pmsi-tunnel-types> [RFC 7538] in the "Border Gateway Protocol (BGP) Parameters" registry. A proposed value is 0x0D.

This document requires registration of End.DT4M, End.DTM6 and End.DTM46 behaviors in "SRv6 Endpoint Behaviors" sub-registry of "Segment Routing Parameters" top-level registry.

Value	Hex	Endpoint behavior	Reference
TBD	TBD	End.DTM4	[This.ID]
TBD	TBD	End.DTM6	[This.ID]
TBD	TBD	End.DTM46	[This.ID]

Table 1: IETF - SRv6 Endpoint Behaviors

9. Security Considerations

The procedures in this document do not introduce any additional security considerations beyond those mentioned in [RFC6513] and [RFC6514]. For general security considerations applicable to P2MP trees, please refer to [I-D.ietf-pim-sr-p2mp-policy] .

10. Acknowledgements

The authors would like to acknowledge Luc Andre Burdett reviewing the document..

11. Contributors

Zafar Ali Cisco Systems, Inc. US

Email: zali@cisco.com

Ehsan Hemmati Cisco Systems, Inc. US

Email: ehemmati@cisco.com

Jayant Kotalwar Nokia Mountain View US

Email: jayant.kotalwar@nokia.com

Tanmoy Kundu Nokia Mountain View US

Email: tanmoy.kundu@nokia.com

Clayton Hassen Bell Canada Vancouver CA

Email: clayton.hassen@bell.ca

12. References

12.1. Normative References

[I-D.ietf-bess-srv6-services]

Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay Services", Work in Progress, Internet-Draft, draft-ietf-bess-srv6-services-07, 11 April 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-srv6-services-07.txt>>.

[I-D.ietf-pim-sr-p2mp-policy]

(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "Segment Routing Point-to-Multipoint Policy", Work in Progress, Internet-Draft, draft-ietf-pim-sr-p2mp-policy-03, 23 August 2021, <<https://www.ietf.org/archive/id/draft-ietf-pim-sr-p2mp-policy-03.txt>>.

[I-D.ietf-spring-sr-replication-segment]

(editor), D. V., Filsfils, C., Parekh, R., Bidgoli, H., and Z. Zhang, "SR Replication Segment for Multi-point Service Delivery", Work in Progress, Internet-Draft, draft-ietf-spring-sr-replication-segment-05, 20 August 2021, <<https://www.ietf.org/archive/id/draft-ietf-spring-sr-replication-segment-05.txt>>.

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

[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>>.
- [RFC7988] Rosen, E., Ed., Subramanian, K., and Z. Zhang, "Ingress Replication Tunnels in Multicast VPN", RFC 7988, DOI 10.17487/RFC7988, October 2016, <<https://www.rfc-editor.org/info/rfc7988>>.
- [RFC8660] Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <<https://www.rfc-editor.org/info/rfc8660>>.
- [RFC8669] Previdi, S., Filsfils, C., Lindem, A., Ed., Sreekantiah, A., and H. Gredler, "Segment Routing Prefix Segment Identifier Extensions for BGP", RFC 8669, DOI 10.17487/RFC8669, December 2019, <<https://www.rfc-editor.org/info/rfc8669>>.
- [RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <<https://www.rfc-editor.org/info/rfc8754>>.
- [RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <<https://www.rfc-editor.org/info/rfc8986>>.

12.2. Informative References

- [I-D.ietf-bess-evpn-bum-procedure-updates] Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates on EVPN BUM Procedures", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-bum-procedure-updates-11, 7 October 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-evpn-bum-procedure-updates-11.txt>>.

[I-D.ietf-bess-evpn-irb-mcast]

Lin, W., Zhang, Z., Drake, J., Rosen, E. C., Rabadan, J., and A. Sajassi, "EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-irb-mcast-06, 24 May 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-evpn-irb-mcast-06.txt>>.

[I-D.ietf-bess-evpn-mvpn-seamless-interop]

Sajassi, A., Thiruvengatasamy, K., Thoria, S., Gupta, A., and L. Jalil, "Seamless Multicast Interoperability between EVPN and MVPN PEs", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-mvpn-seamless-interop-02, 16 February 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-evpn-mvpn-seamless-interop-02.txt>>.

[I-D.ietf-bess-mvpn-evpn-aggregation-label]

Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", Work in Progress, Internet-Draft, draft-ietf-bess-mvpn-evpn-aggregation-label-06, 19 April 2021, <<https://www.ietf.org/archive/id/draft-ietf-bess-mvpn-evpn-aggregation-label-06.txt>>.

[RFC6625] Rosen, E., Ed., Rekhter, Y., Ed., Hendrickx, W., and R. Qiu, "Wildcardcards in Multicast VPN Auto-Discovery Routes", RFC 6625, DOI 10.17487/RFC6625, May 2012, <<https://www.rfc-editor.org/info/rfc6625>>.

[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>>.

[RFC7899] Morin, T., Ed., Litkowski, S., Patel, K., Zhang, Z., Kebler, R., and J. Haas, "Multicast VPN State Damping", RFC 7899, DOI 10.17487/RFC7899, June 2016, <<https://www.rfc-editor.org/info/rfc7899>>.

[RFC8293] Ghanwani, A., Dunbar, L., McBride, M., Bannai, V., and R. Krishnan, "A Framework for Multicast in Network Virtualization over Layer 3", RFC 8293, DOI 10.17487/RFC8293, January 2018, <<https://www.rfc-editor.org/info/rfc8293>>.

Authors' Addresses

Rishabh Parekh
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
United States of America

Email: riparekh@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Arvind Venkateswaran
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
United States of America

Email: arvvenka@cisco.com

Hooman Bidgoli
Nokia
Ottawa
Canada

Email: hooman.bidgoli@nokia.com

Daniel Voyer
Bell Canada
Montreal
Canada

Email: daniel.voyer@bell.ca

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: 1 July 2022

K. Vairavakkalai
M. Jeyanthan
Juniper Networks, Inc.
28 December 2021

BGP signalled MPLS-namespaces
draft-kaliraj-bess-bgp-sig-private-mpls-labels-04

Abstract

The MPLS-forwarding-layer in a core network is a shared resource. The MPLS FIB at nodes in this layer contains labels that are dynamically allocated and locally significant at that node.

For some usecases like upstream-label-allocation, it is useful to be able to create virtual private MPLS-forwarding-layers over this shared MPLS-forwarding-layer. This allows installing deterministic private label-values in the private-FIBs created at nodes participating in this private MPLS forwarding-layer, while preserving the "locally significant" nature of the underlying shared 'public' MPLS-forwarding-layer.

This specification describes the procedures to create such virtual private MPLS-forwarding layers (private MPLS-planes) using a new BGP family. And gives a few example use-cases on how this private forwarding-layers can be used.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on 1 July 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

1. Introduction	3
2. Motivation	3
3. Constructs and building blocks	4
3.1. Context Protocol Nexthop Address	4
3.2. MPLS context FIB	4
3.3. Context Label	5
3.4. Roles of nodes in a MPLS-plane	5
3.4.1. Edge-nodes (PLER)	5
3.4.2. Transit-nodes (PLSR)	5
3.5. Sending traffic into the MPLS plane	6
4. Terminology	6
5. BGP families, routes and encoding	7
5.1. New address-families for "MPLS namespace signaling" . . .	8
5.1.1. AFI: MPLS, SAFI: 128	8
5.1.2. AFI: MPLS, SAFI: 1	8
5.2. Routes and Operational procedures	9
5.2.1. "Context-Nexthop" discovery route	9
5.2.2. MPLS namespace "Private Label" routes	10
6. Example of Usecases	13
6.1. Mezanine transport layer in a Seamless-MPLS network . . .	13
6.2. Service Forwarding Helper usecase	14
6.3. Standard BGP API to a MPLS network's forwarding-plane . .	14
6.4. Traffic engineering and Security advantages	14
7. IANA Considerations	15
8. Security Considerations	15
9. Acknowledgements	15
10. Normative References	15
Authors' Addresses	16

1. Introduction

The MPLS-forwarding-layer in a core network is a shared resource. The MPLS FIB at nodes in this layer contains labels that are dynamically allocated and locally significant at that node.

For some usecases like upstream-label-allocation, it is useful to be able to create virtual private MPLS-forwarding-layers over this shared MPLS-forwarding-layer. This allows installing deterministic private label-values in the private-FIBs in this private forwarding-layer, while preserving the "locally significant" nature of the underlying shared 'public' MPLS-forwarding-layer.

It can be noted that, mechanism described in this document is nothing but a [RFC4364] style BGP VPN where the FEC is MPLS-Label, instead of IP-prefix. This document defines new address-families (AFI: MPLS, SAFI: VPN-Unicast, Unicast) and associated signaling mechanisms to create and use MPLS forwarding-contexts in a network. The concepts of MPLS-Context-tables and upstream allocation are described in [RFC5331].

BGP speakers participating in the private MPLS FIB layer create instances of "MPLS forwarding-context" FIBs, which are identified using a "Context-Protocol-Nexthop (CPNH)". A Context-label MAY be advertised in conjunction with the Context Protocol Nexthop (CPNH) using new BGP address-family to other speakers.

2. Motivation

A provider's core network consists of a global-domain (default forwarding-tables in P and PE nodes) that is shared by all tenants in the network and may also contain multiple private user-domains (e.g. VRF route tables).

The global MPLS forwarding-layer can be viewed as the collection of all default MPLS forwarding-tables. This global MPLS Fib layer contains labels locally significant to each node. The "local-significance of labels" gives the nodes freedom to participate in MPLS-forwarding with whatever label-ranges they can support in forwarding hardware.

In emerging usecases some applications using the MPLS-network may benefit from a "static labels" view of the MPLS-network. In some other usecases, a standard mechanism to do Upstream label-allocation is beneficial.

It is desirable to leave the global MPLS FIB layer intact, and build private MPLS FIB-layers on top of it to achieve these requirements. The private-MPLS-FIBs can then be used by the applications as desired. The private MPLS-FIBs need to be created only at the nodes in the network where predictable label-values (external label allocation) is desired. E.g. P-routers that need to act as a "Detour-nodes" or "Service-Forwarding-Helpers" that need to mirror service-labels.

In other words, provisioning of these private MPLS-FIBs can be gradual and can co-exist with nodes not supporting the feature described in this document. These private-MPLS-FIBs can be stitched together using either the Context-labels over the existing shared MPLS-network tunnels, or 'private' context-interfaces - to form the "private MPLS-FIB layer".

An application can then install the routes with desired label-values in the private forwarding-contexts with desired forwarding-semantics.

3. Constructs and building blocks

The building-blocks that construct a private MPLS plane are described in this section.

3.1. Context Protocol Nexthop Address

A private MPLS plane (just "MPLS plane" here-after) is identified by an IP-address called Context Protocol Nexthop (CPNH). This address is unique in the core-network, like any other loopback address.

A loopback-address uniquely identifies a specific node in the network, and we call it Global Protocol Nexthop (GPNH) in this document. The CPNH address uniquely identifies a "MPLS-plane".

Each node that has forwarding-context for a MPLS-plane MUST be configured with the same CPNH but a different RD, such that the RD:CPNH will uniquely identify that node in the MPLS-plane.

3.2. MPLS context FIB

An instance of a MPLS forwarding-table at a node in the private MPLS-plane. This Private MPLS FIB contains the private-label routes.

A node can have context-FIB for multiple MPLS-planes. The same label-value can have a different forwarding-semantic in each MPLS-plane. Thus the applications using that MPLS-plane get a deterministic label-value independent of other applications using other MPLS-planes.

The terms "private MPLS FIB-layer" and "private MPLS-plane" are used interchangeably in this document.

3.3. Context Label

A context-label is a non-reserved dynamically allocated label, that is installed in the global MPLS FIB, and points to a MPLS-Context-FIB. The Context-Label have forwarding semantics as follows in the global MPLS-FIB:

Context-Label -> Pop and Lookup in MPLS-Context-Fib

Advertising the "Context-Label in conjunction with the GPNH" tells the network how to reach a "RD:CPNH".

3.4. Roles of nodes in a MPLS-plane

The node roles in a MPLS-plane can be classified into "edge nodes" (call them PLER) or "transit-nodes" (call them PLSR).

3.4.1. Edge-nodes (PLER)

Private Label Edge-routers (PLER) have MPLS context-FIB that belong to the MPLS-plane. They advertise the presence of this context-FIB using transport layer address families like BGP-CT [BGP-CT] or BGP-LU, and private-label routes from this FIB are advertised using new BGP AFI/SAFI described in this document.

3.4.2. Transit-nodes (PLSR)

These are just Border-nodes that do label-swap forwarding for the Context-Labels they see in the Context-Protocol-Nexthop advertisement routes (BGP-CT or BGP-LU) going thru them. They basically stitch/extend the label switched path to a PLER's CPNH when they re-advertise the CPNH routes with nexthop-self.

PLSRs don't have MPLS context-FIBs. PLSRs don't have Context Protocol-Nexthop. Because they don't have Private label routes to originate.

However a node in the network can play both roles, of PLER and PLSR.

3.5. Sending traffic into the MPLS plane

At a PLER, MPLS-traffic arriving with private-label hits the correct private MPLS-FIB by virtue of either arriving on a "private network-interface" that is attached to the MPLS context-FIB, or arriving with a "Context-label" on a network-interface attached to the global MPLS-FIB.

To send data traffic into this private MPLS plane, the sender MUST use as handle either a "Context-label" advertised by a node or a "Private-interface" owned by the MPLS context-FIB at the node. The MPLS context-FIB is created for an application that needs a private MPLS-plane.

The Context-Label is the only dynamic label-value the application needs to learn from the network (PLER node it is connected to), to be able to use the private MPLS-plane. The application can chose predictable value for the labels to be programmed in the private MPLS-FIBs.

Once the packet enters the private MPLS plane at an edge-node (PLER), the node will forward the packet to the next node (PLSR or PLER), by pushing the Context-label advertised by that next-node, and the transport-label to reach that node's GPNH. This will repeat until the packet reaches the PLER's private MPLS-FIB that originated that private MPLS-label.

At each PLER in the MPLS-plane, the private-label value remains the same, and points towards the same resource attached to the MPLS-plane. This allows the applications using the MPLS-network a static-labels view of the resources attached to the private MPLS-plane.

At each PLSR in the MPLS-plane, the context-label value will change (be swapped in forwarding), but is transparent to the application.

4. Terminology

P-router : A Provider core router, also called a LSR

LSR : Label Switch Router (pure transport node speaking LDP, RSVP etc)

PLSR: a BGP-CT or BGP-LU transit node in a private MPLS-plane, that does label-swap forwarding for Context-Label.

PLER: an edge node in a private MPLS-plane. It has a forwarding-context for private-labels.

Detour-router : A BGP border node that is used as a loose-hop in a traffic-engineered path

PE-router : Provider Edge router, that hosts a service (Internet, L3VPN etc)

SE-router : Service Edge router. Same as PE.

SFH-router : Service Forwarding Helper. A node helping an SE-router with service-traffic forwarding, using Service-routes mirrored by the SE.

MPLS FIB : MPLS Forwarding table

Global MPLS FIB : Global MPLS Forwarding table, to which shared-interfaces are connected

Private MPLS FIB : Private MPLS Forwarding table, to which private-interfaces are connected

Private MPLS FIB Layer (Private MPLS plane): The group of Private MPLS FIBs in the network, connected together via Context-Labels

Context-Label : Locally-significant Non-reserved label pointing to a private MPLS FIB

Context nexthop IP-address (CPNH) : An IP-address that identifies the "Private MPLS FIB Layer". RD:CPNH identifies a Private MPLS FIB at a specific BGP node.

Global nexthop IP-address (GPNH) : Global Protocol Nexthop address. E.g. a loopback address used as transport tunnel end-point.

5. BGP families, routes and encoding

This section describes the new constructs defined by this document.

5.1. New address-families for "MPLS namespace signaling"

This document defines a new AFI: "MPLS" (IANA code TBD). And two new address-families, using SAFIs 128 and 1. These address families are used to signal "MPLS namespaces" in BGP. To send or receive routes of these address families, these AFI, SAFI pair of values MUST be negotiated in Multiprotocol Extensions capability described in RFC4760 [RFC4760]

5.1.1. AFI: MPLS, SAFI: 128

This address-family is used to exchange private label-routes in private MPLS-FIBs at routers that are connected using a common network interface. The private label route has NLRI prefix format "RD:PrivateLabel" and contains Route-Target extended-community identifying the private-FIB-Layer (VPN) the route belongs to. The nexthop of these routes is set to either the GPNH or the CPNH of the BGP-speaker advertising the RFC-8277 label.

Any transport layer protocol is used to advertise the Context-Label that the receiving router uses to send traffic into the private MPLS-FIB. The Context-Label installed in the global MPLS-FIB points to the private MPLS-FIB. The Context-Label is required when the connecting-interface is a shared common interface that terminates into the global MPLS FIB.

Routes of this address-family can be sent with either IPv4 or IPv6 nexthop. The type of nexthop is inferred from the length of the nexthop.

When the length of Next Hop Address field is 24 (or 48) the nexthop address is of type VPN-IPv6 with 8-octet RD set to zero (potentially followed by the link-local VPN-IPv6 address of the next hop with an 8-octet RD).

When the length of Next Hop Address field is 12 the nexthop address is of type VPN-IPv4 with 8-octet RD.

5.1.2. AFI: MPLS, SAFI: 1

This address-family is used to exchange private label-routes in private MPLS-FIBs to routers that are connected using a private network-interface.

Because the interface is private, and terminates directly into the private MPLS-FIB, a Context-Label is not required to access the private MPLS-FIB and NLRI prefix format is just "PrivateLabel/24", without the RD.

Routes of this address-family can be sent with either IPv4 or IPv6 nexthop. The type of nexthop is inferred from the length of the nexthop.

When the length of Next Hop Address field is 16 (or 32) the nexthop address is of type IPv6 (potentially followed by the link-local IPv6 address of the next hop).

When the length of Next Hop Address field is 4 the nexthop address is a 4 octet IPv4 address.

5.2. Routes and Operational procedures

5.2.1. "Context-Nexthop" discovery route

The Context-NH discovery route may be a BGP-LU or [BGP-CT] family route that carries CPNH in the "Prefix" portion of the NLRI. And the Context-Label is carried in the "Label" field in the [RFC8277] format NLRI.

This route is advertised with the following path-attributes:

- * BGP Nexthop attribute (code 14, MP_REACH) carrying GPNH address.
- * Route-Target extended community, identifying the Transport class, if applicable.

The "Context-Nexthop discovery route" is originated by each speaker who acts as a PLER. The "RD:Context-nexthop" uniquely identifies the private-MPLS-FIB at the speaker. The "Context-nexthop address" uniquely identifies the private-MPLS-plane in the network. The Context-Label advertised in this route has a local forwarding semantic of "Pop, Lookup in Private-MPLS-FIB".

A BGP speaker readvertising a BGP-CT Context-Nexthop for RD:CPNH discovery-route MUST follow the mechanisms described in [BGP-CT]. Specifically when re-advertising with "next-hop self" MUST allocate a new Label with a forwarding semantic of "Swap Received-Context-Label, Forward to Received-GPNH". This extends reachability to the CPNH across tunnel domains.

5.2.2. MPLS namespace "Private Label" routes

The Private Label routes are carried in the new address-family "MPLS VpnUnicast" (AFI:MPLS, SAFI:128) aka "MPLS-namespace signaling", defined in this document.

The NLRI format follows the specifications in [RFC8277], with the "Prefix" portion of the NLRI comprising of the RD and "Private MPLS Label" encoded as shown below.

In a MP_REACH_NLRI attribute whose AFI/SAFI is MPLS/128, the "Length" field will be 112 bits or less, comprising of the Label, RD and "Private MPLS Label".

In a MP_REACH_NLRI attribute whose AFI/SAFI is MPLS/1, the "Length" field will be 48 bits or less, comprising of the Label, and "Private MPLS Label".

NLRI Prefix (Private Label route, AFI:MPLS, SAFI:128)

This picture shows NLRI format when the RFC-8277 Multiple Labels Capability is not used:

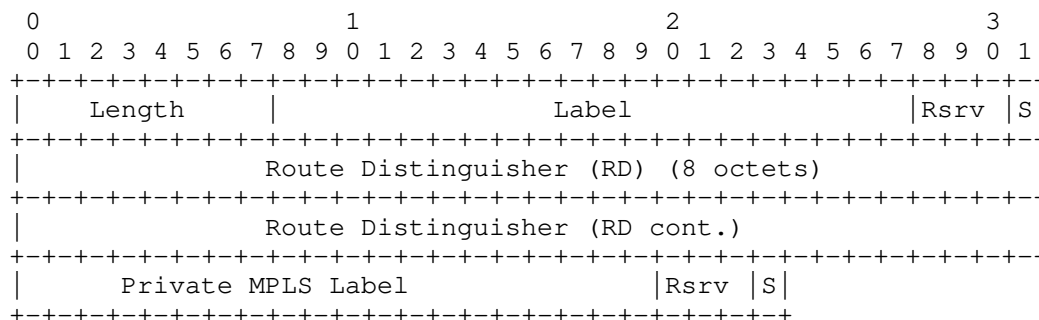


Fig 1: RFC-8277 NLRI with one Label.

- Length:

The Length field consists of a single octet. It specifies the length in bits of the remainder of the NLRI field.

In a MP_REACH_NLRI attribute whose AFI/SAFI is MPLS/128, the "Length" field will be 112 bits or less, comprising of the Label, RD and "Private MPLS Label".

As specified in [RFC4760], the actual length of the NLRI field will be the number of bits specified in the Length field,

rounded up to the nearest integral number of octets.

- Label:
The Label field is a 20-bit field containing an MPLS label value (see [RFC3032]). This label is locally significant, downstream allocated at the speaker identified in the BGP Nexthop field in MP_REACH_NLRI (code 14). This label is pushed in nexthop of the route installed in MPLS context FIB at receiving router.
- Route Distinguisher (RD):
The 8 byte Route Distinguisher as specified in [RFC4760].
- Private MPLS Label:
The "Private MPLS Label" field is a 20-bit field containing an MPLS label value (see [RFC3032]). This is an upstream assigned MPLS label, used as destination of route installed in MPLS context FIB at the receiving router.
- Rsrv:
This 3-bit field SHOULD be set to zero on transmission and MUST be ignored on reception.
- S:
This 1-bit field MUST be set to one on transmission and MUST be ignored on reception.

Attributes on this route:

- * BGP Nexthop attribute (code 14, MP_REACH) carrying a GPNH address.
(OR)
- * The Multi-nexthop attribute [MULTI-NH] with forwarding-semantic:
 - "Forward to RD:CPNH"
- * Route-Target extended-community, identifying the private FIB-layer

MultiNexthop BGP-attribute (Private Label route)

MultiNH.Num-Nexthops = 1
FwdSemanticsTLV.FwdAction = Forward
NHDescrTLV.NhopDescrType = RD:CPNH or GPNH

Fig 2: MultiNexthop attr of Private Label route

A speaker MAY readvertise a private-label-route without changing the Nexthop (RD:CPNH) carried in it, if the speaker is a pure PLSR.

If it does alter the nexthop to SelfRD:CPNH, it SHOULD act as a PLER, and for e.g. originate a "Context-Nexthop discovery route" for prefix "SelfRD:CPNH".

Even if the speaker sets nexthop-address to Self because of regular BGP readvertisement-rules, Label Prefix MUST NOT be altered, and the received NLRI "RD:Private-Label1" MUST be re-advertised as-is. Such that value of label "Private-Label1" doesn't change while the packet traverses multiple nodes in the private-MPLS-FIB-layer.

The Route-target attached to the route is the one identifying the private MPLS FIB layer (VPN). The Private-label routes resolve over the Context-nexthop route that belong to the same VPN.

A node receiving a "Private-Label route" RD:L1 MUST install the label L1 in the private MPLS Forwarding-context identified by the Route-Target attached to the route.

The label route MUST be installed with forwarding-semantic as specified in the received Multi-nexthop attribute. As an example, a Detour node MAY receive the private-label-route with a forwarding-semantic of "Forward to RD:CPNH" operation. And an Egress node MAY receive a private-label-route with a forwarding-semantic pointing to a resource it houses. Note that such a Private-label BGP-route MAY be received from external-application also.

5.2.2.1. Resolving received Private Label-routes

A node receiving a "Context-nexthop discovery route" MUST be capable of using either the CPNH or the RD:CPNH carried in the NLRI, to resolve other routes received with this CPNH address or RD:CPNH in the "Nexthop-attributes".

The receiver of a private-label route MUST recursively resolve the received nexthop (RD:CPNH) over the Context-Nexthop discovery-route for prefix "RD:CPNH" to determine the label stack "Context-Label, Transport-Label" to push, so that the MPLS packet with private-label reaches the private MPLS FIB originating the route.

If a node receives multiple "Context-nexthop discovery route" for a CPNH, it SHOULD run path-selection after stripping the RD, to find the closest ingress to the private-MPLS-plane identified by the CPNH. This best path SHOULD be used to resolve a received private-label-route.

6. Example of Usecases

6.1. Mezanine transport layer in a Seamless-MPLS network

Typically service-routes in a MPLS network bind to the following entities that identify point-of-presence of a service:

- * Protocol Nexthop - PE loopback address (GPNH)
- * Service Label - PE advertised locally significant label that identifies the service

In this model, whenever a PE is taken out of service the GPNH changes, and Service-Label changes - which causes maintenance a heavy convergence event. Because the service-routes with massive-scale need to be readvertised with new service-label or PE-address.

An alternate model could be: to advertise the Service-routes with a protocol-nexthop of CPNH (without RD), with a forwarding-semantic of:

- * "Push <Private-Label>, and Forward to CPNH"

This model fully decouples the service-layer from the transport-layer identifiers, by making the Service-routes refer to the CPNH and Private-Labels. Thus the underlying transport-layer can change (nodes representing a Private-label can be added or removed) without any changes to the service-routes. Which present good scaling properties for the network.

This model also allows anycast traffic forwarding to any resource in the network. Multiple PEs can advertise the same Private-Label to identify a specific service (e.g. peering with an AS) they are offering.

Once the service-route traffic enters the private-FIB-layer, at the closest entry-point determined by path-selection of CPNH auto-discovery routes; then the Private-Labels (with pre-determined values) pushed will determine the loose hop path taken by the traffic and also the destination-resource.

6.2. Service Forwarding Helper usecase

In a virtualized environment a Service-PE node (that comprises of a vCP and multiple vFPs) can mirror MPLS labels (GL1) in its global MPLS-FIB to a private forwarding context at an upstream node (SFH) with information on which vFPs are optimal exit-points for that label. Such that the SFH can optimally forward traffic to GL1 to the right vFPs, thus avoiding intra fabric traffic hops.

To do this, the service-PE advertises a private-label route with RD:GL1 to the SFH node. The route is advertised with a Multi-nexthop attribute with one or more legs that have a "Forward to SEPx" semantics. Where SEPx is one of many exit-points at the Service-PE node.

6.3. Standard BGP API to a MPLS network's forwarding-plane

This mechanism facilitates predictable (external-allocator determined) label-values, using a standard BGP-family as the API. It gives the external applications a separate MPLS-FIB to play with, totally separate from other applications.

This also avoids vendor specific-API dependencies for external-allocators (controller softwares), and vice-versa.

This mechanism also increases the overall MPLS label-space available in the network, because it creates per-app label-forwarding-contexts (namespaces), instead of reserving/splitting the global MPLS FIB among various applications.

6.4. Traffic engineering and Security advantages

- * Ability of ingress to steer mpls-traffic thru specific detour loose-hop nodes using predictable-labels' stack.
- * Provide label-spoofing protection at edge-nodes - by virtue of using separate mpls-forwarding-contexts
- * Allow private-MPLS label usage to spread across multiple-domains/ AS and work seamlessly with existing technologies like Inter-AS VPN option C.

7. IANA Considerations

This document makes following requests of IANA.

New BGP AFI code ("Address Family Numbers" registry):

* 16399 for "MPLS Namespaces"

Note to RFC Editor: this section may be removed on publication as an RFC.

8. Security Considerations

Using separate mpls-forwarding-contexts for separate applications and stitching them into separate MPLS-planes increases the security attributes of the MPLS network.

9. Acknowledgements

The authors thank Jeffrey (Zhaohui) Zhang, Ron Bonica, Jeff Haas and John Scudder for the valuable discussions.

10. Normative References

- [BGP-CT] Vairavakkalai, K., "BGP Classful Transport Planes", 25 August 2021, <<https://tools.ietf.org/html/draft-kaliraj-idr-bgp-classful-transport-planes-12#section-11.3>>.
- [MULTI-NH] Vairavakkalai, K., "BGP MultiNexthop attribute", 28 December 2021, <<https://tools.ietf.org/html/draft-kaliraj-idr-multinexthop-attribute-04>>.
- [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>>.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <<https://www.rfc-editor.org/info/rfc4364>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, DOI 10.17487/RFC5331, August 2008, <<https://www.rfc-editor.org/info/rfc5331>>.

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

Authors' Addresses

Kaliraj Vairavakkalai
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: kaliraj@juniper.net

Minto Jeyananth
Juniper Networks, Inc.
1133 Innovation Way,
Sunnyvale, CA 94089
United States of America

Email: minto@juniper.net

BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2022

A. Sajassi
P. Brissette
M. Mishra
S. Thoria
Cisco Systems
J. Rabadan
Nokia
J. Drake
Juniper Networks
July 11, 2021

AC-Aware Bundling Service Interface in EVPN
draft-sajassi-bess-evpn-ac-aware-bundling-04

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 Broadcast Domain.

EVPN technology defines three different types of service interface which serve different requirements but none of them address the requirement of supporting multiple subnets within single Broadcast Domain. In this draft we define new service interface type to support multiple subnets in single Broadcast Domain. Service interface proposed in this draft will be applicable to multihoming case only.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119] and RFC 8174 [RFC8174].

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute

working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on January 12, 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Problem with Unicast MAC route	6
1.2. Problem with Multicast route synchronization	6
1.3. Potential Security concern caused by misconfiguration	6
2. Terminology	6
3. Requirements	7
4. Solution Description	8
4.1. Control Plane Operation	8
4.1.1. MAC/IP Address Advertisement	8
4.1.1.1. Local Unicast MAC learning	8
4.1.1.2. Remote Unicast MAC learning	8
4.1.2. Multicast route Advertisement	8
4.1.2.1. Local multicast state	9
4.1.2.2. Remote multicast state	9
4.2. Data Plane Operation	9
4.2.1. Unicast Forwarding	9
4.2.2. Multicast Forwarding	10
5. Mis-configuration across multihoming peers	10
6. BGP Encoding	10
6.1. Attachment Circuit ID Extended Community	10
6.2. Ethernet-tag field vs AC ID Extended Community	11

7. Security Considerations	11
8. IANA Considerations	11
9. Acknowledgement	11
10. References	11
10.1. Normative References	11
10.2. Informative References	12
Authors' Addresses	12

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 Broadcast Domain. Each subnet would be differentiated by VLAN. Thus, single IRB interface can still serve multiple subnet.

Motivation behind such deployments are

1. **Manageability:** The support to have multiple subnets using single Broadcast Domain requires only one Broadcast Domain and one IRB for "N" subnets compare to "N" Broadcast Domain and "N" IRB interface to manage.
2. **Simplicity:** It avoids extra configuration by configuring VLAN Range with single BD and IRB as compare to individual VLAN, BD and IRB interface per subnet.

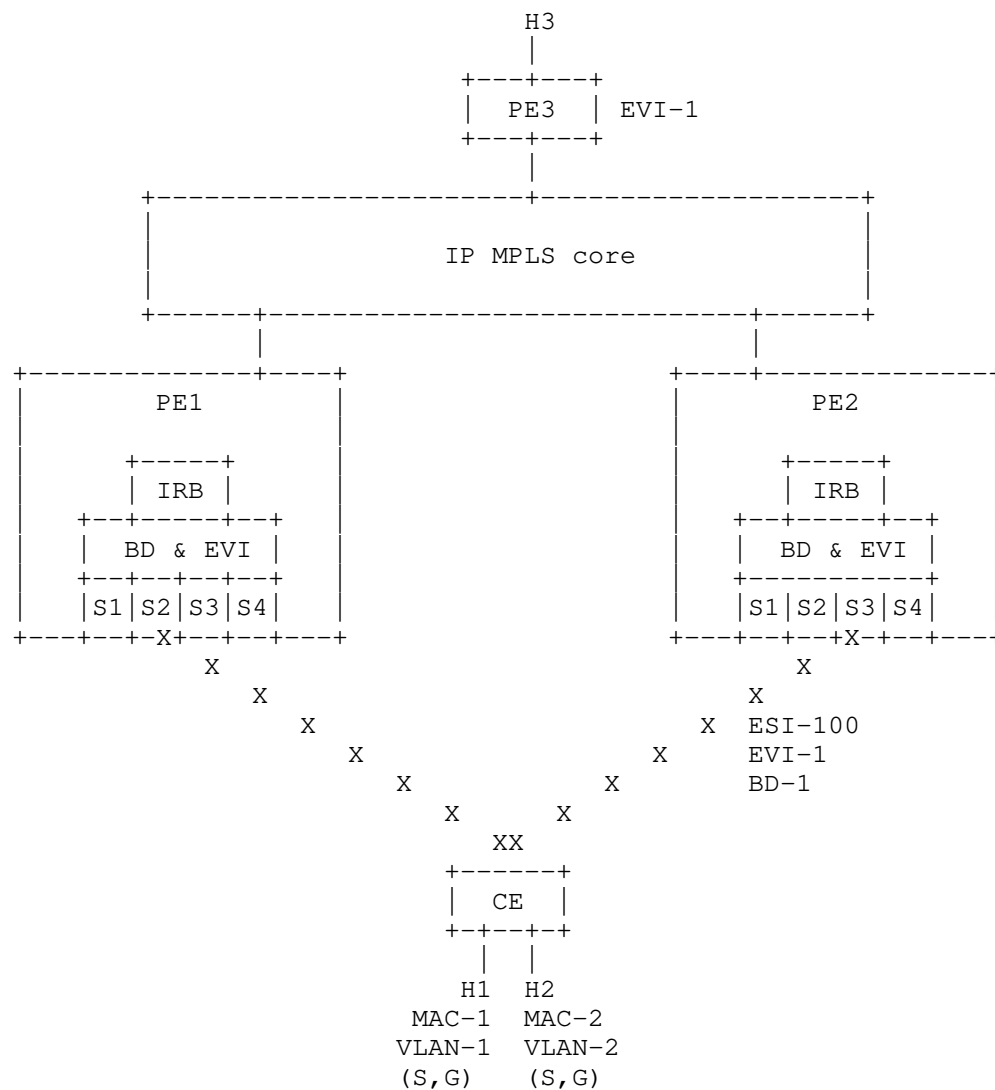
[RFC7432] defines three types of service interface. None of them provide flexibility to achieve multiple subnets within single Broadcast Domain. The different types of service interface from [RFC7432] are:

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.

From definition, it seems like VLAN Bundle Service Interface does provide flexibility to support multiple subnets within single Broadcast Domain. However, the requirement is to have multiple subnets from same ES on multi-homing all active mode; that would not work. For example, let's take the case from Figure 1 where PE1 learns MAC of H1 on VLAN 1 (subnet S1). PE1 originates EVPN MAC route, as per [RFC7432], where the Ethernet Tag would be set to 0. Incoming packets from IRB interface, at PE2, are untagged packet. PE2 does not have any associated AC information from EVPN MAC routes advertised by PE1. 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 Broadcast Domain. This extension is applicable only for multi-homed EVPN peers.



EVPN topology with multi-homing and non multihoming peer.

Figure 1

Figure 1 shows sample EVPN topology where PE1 and PE2 are multihomed peers. PE3 is remote peer participating in the same EVPN instance (EVI-1). It illustrates four subnets S1, S2, S3 and S4 where numerical value provides associated VLAN information.

1.1. Problem with Unicast MAC route

BD-1 has multiple subnets 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 providing 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 does not know the destination AC since multiple bridge ports have the same ESI assignment.

1.2. Problem with Multicast route synchronization

[I-D.ietf-bess-evpn-igmp-mld-proxy] defines mechanism to synchronize multicast routes between multihome peers. In above case, if receiver behind S1 send IGMP membership request, CE could hash it to either of the PEs. When multicast route is originated, it does not contain any AC information. Once it reaches to peering PE, it does not have any information about which subnet this IGMP membership request belong to. Similarly to unicast traffic problem, the incoming multicast traffic from IRB cannot be forwarded to proper AC.

1.3. Potential Security concern caused by misconfiguration

In case of single subnet per Broadcast Domain, there is potential case of security issue. For example, PE1 has BD1 configured with VLAN-1 where as multihome peer PE2 has BD1 configured VLAN-2. Each of the IGMP membership requests 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. Again, similar issue can potentially be seen with unicast traffic.

2. Terminology

- o AC: Attachment Circuit.
- o ARP: Address Resolution Protocol.
- o 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.

- o 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.
- o BT: Bridge Table. The instantiation of a BD in a MAC-VRF, as per [RFC7432].
- o Ethernet A-D route: Ethernet Auto-Discovery (A-D) route, as per [RFC7432].
- o EVI: EVPN Instance spanning the NVE/PE devices that are participating on that EVPN, as per [RFC7432].
- o EVPN: Ethernet Virtual Private Networks, as per [RFC7432].
- o IRB: Integrated Routing and Bridging interface. It connects an IP-VRF to a BD (or subnet).
- o 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.
- o ND: Neighbor Discovery Protocol.
- o RD: BGP Route Distinguisher.
- o RT-2: EVPN route type 2, i.e., MAC/IP advertisement route, as defined in [RFC7432].
- o RT-5: EVPN route type 5, i.e., IP Prefix route. As defined in Section 3 of [EVPN-PREFIX].
- o SN: Subnet.
- o TS: Tenant System.
- o VLAN: The usage of VLAN refers to 802.1Q or 802.1AD tag.
- o (S,G): Multicast membership request
- o This document also assumes familiarity with the terminology of [RFC7432], [RFC8365], [RFC7365].

3. Requirements

1. A service interface represents an attachment-circuit where multiple VLAN are configured on. Each of these VLANs are represented by a different AC under a single Broadcast Domain.

2. Single Broadcast Domain MUST support service interfaces.
3. Service interface MUST be applicable to multihomed peers only.
4. Service interface MUST have an Ethernet-Segment identifier assignment.
5. New service interface handling procedures MUST be backward compatible with implementation procedures defined in [RFC7432]
6. New service interface MUST support EVPN multicast routes defined in [I-D.ietf-bess-evpn-igmp-mld-proxy] too.

4. Solution Description

4.1. Control Plane Operation

4.1.1. MAC/IP Address Advertisement

4.1.1.1. Local Unicast MAC learning

[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.

MAC/IP route construction follows mechanism defined in [RFC7432] section 9.2.1. An attach Attachment Circuit ID Extended Community (Section 6.1) must be attached to EVPN RT-2.

4.1.1.2. Remote Unicast MAC learning

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]

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

From Figure 1, 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 AC in given Broadcast Domain receives IGMP membership request, it MUST synchronize multicast state by originating multicast route defined in [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 1 when H2 sends IGMP membership request for (S,G), CE hashed it to one of the PE. Lets say PE1 received IGMP membership request. 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.

PE1 must originate multicast route updates for any subsequent IGMP membership requests under same or different subnet attaching adequate Attachment Circuit ID Extended Community (Section 6.1).

4.1.2.2. Remote multicast state

If multihomed PE receives remote multicast route on Broadcast Domain for given ES, route MUST be programmed to correct subnet. Subnet information MUST be extracted from Attachment Circuit ID Extended Community. That value maps to the VLAN of a local AC where the multicast route is associated to.

4.2. Data Plane Operation

4.2.1. Unicast Forwarding

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

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

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

If destination MAC lookup is performed on known unicast packet, destination MAC lookup MUST provide VLAN and local AC information. 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 VLAN value.

4.2.2. Multicast Forwarding

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

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 across multihoming peers

If there is mis-configuration of VLAN or VLAN range across multihoming peers, same MAC address would be learnt with different VLAN per Broadcast Domain. In this case Error message MUST be thrown for operator to make configuration changes. Furthermore, the errored MAC route MUST be ignored.

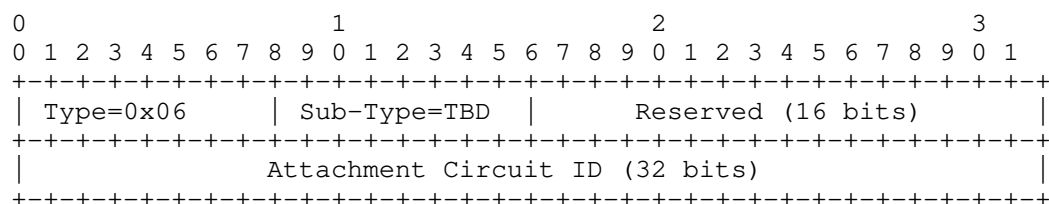
6. BGP Encoding

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. 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. It may also be advertised along with EVPN Multicast Route (Route Type 7 and 8) as per [I-D.ietf-bess-evpn-igmp-mld-proxy]. Generically speaking, the new extended community must be attached to any routes which require specific VLAN identification.

The Attachment Circuit ID Extended Community is encoded as an 8-octet value as follows:



Attachment Circuit ID Extended Community

The attachment circuit ID plays the role of normalized VID. It is defined as per [I-D.ietf-bess-evpn-vpws-fxc].

6.2. Ethernet-tag field vs AC ID Extended Community

The current proposal is entirely backward compatible with [RFC7432] VLAN-aware bundling mode since the Ethernet-tag field remains intact. However, it has its own drawbacks. For instance with multicast, the same (S,G) maybe be used over different subnets. In that case, the same route MUST carry multiple AC ID Extended Community; one per attachment Circuit ID / VLAN. It may happen that the number of VLAN is fairly large. Multiple routes with different RD may be required to carry such amount of Extended Community. This approach is complexifying the overall solution and implementation.

To remedy to that situation, the attachment Circuit ID MAY be set to 0xFFFF_FFFF. That value tells peer PE that the attachment Circuit ID is carried has part of the Ethernet Tag field of the associated route. Since the key of the EVPN route is unique, multiple AC ID Extended Community per route is no longer required. There is drawback. It pose backward interoperability issue with PE expecting a zero Ethernet-TAG ID.

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-vpws-fxc]

Sajassi, A., Brissette, P., Uttaro, J., Drake, J., Boutros, S., and J. Rabadan, "EVPN VPWS Flexible Cross-Connect Service", draft-ietf-bess-evpn-vpws-fxc-03 (work in progress), June 2021.

- [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

- [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

Email: sajassi@cisco.com

Patrice Brissette
Cisco Systems

Email: pbrisset@cisco.com

Mankamana Mishra
Cisco Systems

Email: mankamis@cisco.com

Samir Thoria
Cisco Systems

Email: sthoria@cisco.com

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com

John Drake
Juniper Networks

Email: jdrake@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 10, 2021

A. Sajassi, Ed.
G. Badoni
P. Warade
S. Pasupula
Cisco Systems
J. Drake, Ed.
Juniper
J. Rabadan, Ed.
Nokia
June 8, 2021

EVPN Support for L3 Fast Convergence and Aliasing/Backup Path
draft-sajassi-bess-evpn-ip-aliasing-02

Abstract

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

Status of This Memo

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

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

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

This Internet-Draft will expire on December 10, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Ethernet Segments for Host Routes in Symmetric IRB . . .	3
1.2. Inter-subnet Forwarding for Prefix Routes in the Interface-less IP-VRF-to-IP-VRF Model	4
1.3. Ethernet Segments for Prefix routes in IP-VRF-to-IP-VRF use-cases	4
1.4. Terminology and Conventions	5
2. IP Aliasing and Backup Path	6
2.1. Constructing the IP A-D per EVI Route	7
3. Fast Convergence for Routed Traffic	8
3.1. Constructing IP A-D per Ethernet Segment Route	9
3.1.1. IP A-D per ES Route Targets	9
3.2. Avoiding convergence issues by synchronizing IP prefixes	9
3.3. Handling Silent Host MAC/IP route for IP Aliasing	9
3.4. MAC Aging	10
4. Determining Reachability to Unicast IP Addresses	10
4.1. Local Learning	10
4.2. Remote Learning	11
4.3. Constructing the IP Routes	11
4.3.1. Route Resolution	11
5. Forwarding Unicast Packets	11
6. Load Balancing of Unicast Packets	12
7. Security Considerations	12
8. IANA Considerations	12
9. Contributors	12
10. Acknowledgments	12
11. References	12
11.1. Normative References	12
11.2. Informative References	13
Authors' Addresses	13

1. Introduction

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

- a. Inter-subnet forwarding for host routes in symmetric IRB [I-D.ietf-bess-evpn-inter-subnet-forwarding].
- b. Inter-subnet forwarding for prefix routes in the interface-less IP-VRF-to-IP-VRF model [I-D.ietf-bess-evpn-prefix-advertisement].
- c. Inter-subnet forwarding for prefix routes when the ESI is used exclusively as an L3 construct [I-D.ietf-bess-evpn-prefix-advertisement].

1.1. Ethernet Segments for Host Routes in Symmetric IRB

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

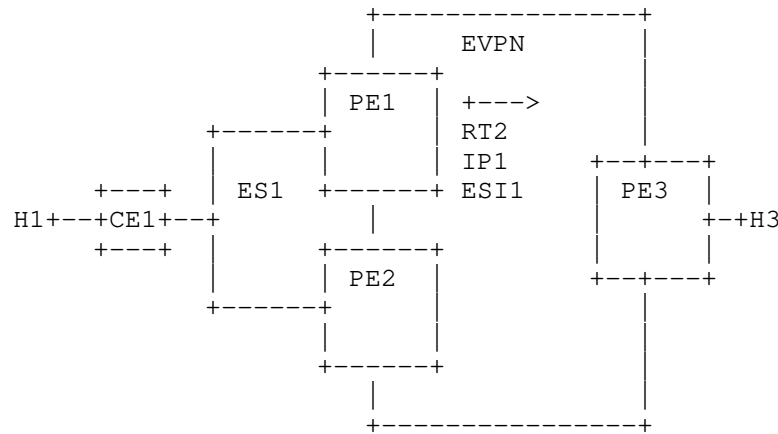


Figure 1: Inter-subnet traffic between Multihoming PEs and Remote PE

With Asymmetric IRB [I-D.ietf-bess-evpn-inter-subnet-forwarding], if H3 sends inter-subnet traffic to H1, routing will happen at PE3. PE3 will be attached to the destination IRB interface and will trigger ARP/ND requests if it does not have an ARP/ND adjacency to H1. A subsequent routing lookup will resolve the destination MAC to H1's MAC address. Furthermore, H1's MAC will point to an ECMP EVPN destination on PE1 and PE2, either due to host route advertisement from both PE1 and PE2, or due to Ethernet Segment MAC Aliasing as detailed in [RFC7432].

With Symmetric IRB [I-D.ietf-bess-evpn-inter-subnet-forwarding], if H3 sends inter-subnet traffic to H1, a routing lookup will happen at

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

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

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

In this model there is no Overlay Index and hence no recursive resolution of the IP Prefix route to either a MAC/IP Advertisement or an Ethernet A-D per ES/EVI route, which means that the fast convergence and aliasing/backup path functions are disabled. In a sense it is already described in section 4.3 of [I-D.ietf-bess-evpn-prefix-advertisement], Bump-in-the-Wire Use-Case, but that section does not describe aliasing. I.e., this document can be considered to be adding the aliasing/backup path function to the Bump-in-the-Wire Use-Case.

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

This document also enables fast convergence and aliasing/backup path to be used even when the ESI is used exclusively as an L3 construct.

As an example, consider the scenario in Figure 2 in which PE1 and PE2 are multi-homed to CE1. However, and contrary to CE1 in Figure 1, in this case the links between CE1 and PE1/PE2 are used exclusively for L3 protocols and L3 forwarding in different BDs, and a BGP session established between CE1's loopback address and PE1's IRB address.

In these use-cases, sometimes the CE supports a single BGP session to one of the PEs (through which it advertises a number of IP Prefixes seating behind itself) and yet, it is desired that remote PEs can build an IP ECMP list or backup IP list including all the PEs multi-homed to the same CE. For example, in Figure 2, CE1 has a single eBGP neighbor, i.e., PE1. Load-balancing for traffic from CE1 to H4 can be accomplished by a default route with next-hops PE1 and PE2, however, load-balancing from H4 to any of the prefixes attached to CE1 would not be possible since only PE1 would advertise EVPN IP Prefix routes for CE1's prefixes. This document provides a solution so that PE3 considers PE2 as a next-hop in the IP ECMP list for CE1's

- CE: Customer Edge device, e.g., a host, router, or switch.s
- EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.
- MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.
- Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.
- Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.
- IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by any routing protocol, E.g., EVPN, IP-VPN and BGP PE-CE IP address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.
- IP route: An IP Prefix route or a MAC/IP Advertisement route that contains a host route.
- LACP: Link Aggregation Control Protocol.
- PE: Provider Edge device.
- Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.
- All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.
- RT5: EVPN IP Prefix route, as specified in [I-D.ietf-bess-evpn-prefix-advertisement].

2. IP Aliasing and Backup Path

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

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

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

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

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

To address the third use-case described in Section 1, where the links between a CE and its multihomed PEs are used exclusively for L3 protocols and L3 forwarding, a PE uses the procedures described in 1) and 2), above. The ESI is of type 4 [RFC7432] and set to the router ID of the CE.

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

2.1. Constructing the IP A-D per EVI Route

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

- The Route-Distinguisher is for the corresponding IP-VRF.
- The Ethernet Tag should be set to 0.
- The route SHOULD carry the RT of the corresponding IP-VRF.

- The route MUST carry the PE's MAC Extended Community if the encapsulation used between the PEs for inter-subnet forwarding is an Ethernet NVO tunnel [I-D.ietf-bess-evpn-prefix-advertisement].
- The route SHOULD carry the Layer 2 Extended Community [RFC8214]. For all-active multihoming, all PEs attached to the specified ES will advertise P=1. For backup path, the Primary PE will advertise P=1 and the Backup PE will advertise P=0, B=1.
 - o The Primary PE SHOULD be a PE with a routing adjacency to the attached CE.
 - o The Primary PE MAY be determined by policy or MAY be elected by a DF Election as in [RFC8584].

3. Fast Convergence for Routed Traffic

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

To accomplish this, EVPN defined the fast convergence function specified in [RFC7432]. This document extends fast convergence to inter-subnet forwarding by having each PE advertise a set of one or more IP A-D per ES routes for each locally attached Ethernet segment (refer to Section 3.1 below for details on how these routes are constructed). A PE may need to advertise more than one IP A-D per ES route for a given ES because the ES may be in a multiplicity of IP-VRFs and the Route-Targets for all of these IP-VRFs may not fit into a single route. Advertising a set of IP A-D per ES routes for the ES allows each route to contain a subset of the complete set of Route Targets. Each IP A-D per ES route is differentiated from the other routes in the set by a different Route Distinguisher (RD).

Upon failure in connectivity to the attached ES, the PE withdraws the corresponding set of IP A-D per ES routes. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all IP addresses associated with the Ethernet Segment in question, across IP-VRFs. If no other PE has advertised an IP A-D per ES route for the same Ethernet Segment, then the PE that received the withdrawal simply invalidates the IP entries for that segment. Otherwise, the PE updates its next-hop adjacencies accordingly.

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

3.1. Constructing IP A-D per Ethernet Segment Route

This section describes the procedures used to construct the IP A-D per ES route, which is used for fast convergence (as discussed in Section 3). The usage/construction of this route remains similar to that described in section 8.2.1. of [RFC7432] with a few notable exceptions as explained in following sections.

3.1.1. IP A-D per ES Route Targets

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

3.2. Avoiding convergence issues by synchronizing IP prefixes

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

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

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

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

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

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

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

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

- a. When PE1 fails, the next hop tracking to PE1 in the underlay routing protocols can help detect the failure.
- b. Upon the failure of its link to CE1, PE1 will withdraw its IP A-D route(s) and PE2 can use this as a trigger to detect failure.

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

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

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

3.4. MAC Aging

In the same example as in Section 3.3, PE1 would do ARP/ND refresh for H1 before it ages out. During this process, H1 can age out genuinely or due to the ARP/ND reply landing on PE2. PE1 must withdraw the local entry from BGP when H1 entry ages out. PE1 deletes the entry from the local forwarding only when there are no remote synced entries.

4. Determining Reachability to Unicast IP Addresses

4.1. Local Learning

The procedures for local learning do not change from [RFC7432] or [I-D.ietf-bess-evpn-prefix-advertisement].

4.2. Remote Learning

The procedures for remote learning do not change from [RFC7432] or [I-D.ietf-bess-evpn-prefix-advertisement].

4.3. Constructing the IP Routes

The procedures for constructing MAC/IP Address or IP Prefix Advertisements do not change from [RFC7432] or [I-D.ietf-bess-evpn-prefix-advertisement].

4.3.1. Route Resolution

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

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

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

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

If after (1) PE1 withdraws the IP route, then PE3 will do delayed deletion of H1, as described in Section 3.3.

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

5. Forwarding Unicast Packets

Refer to Section 5 in [I-D.ietf-bess-evpn-inter-subnet-forwarding] and [I-D.ietf-bess-evpn-prefix-advertisement].

6. Load Balancing of Unicast Packets

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

7. Security Considerations

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

8. IANA Considerations

No IANA considerations.

9. Contributors

10. Acknowledgments

11. References

11.1. 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>>.
- [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>>.
- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

11.2. Informative References

- [I-D.ietf-bess-evpn-inter-subnet-forwarding]
Sajassi, A., Salam, S., Thoria, S., Drake, J. E., and J. Rabadan, "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-13 (work in progress), February 2021.
- [I-D.ietf-bess-evpn-prefix-advertisement]
Rabadan, J., Henderickx, W., Drake, J. E., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-11 (work in progress), May 2018.

Authors' Addresses

A. Sajassi (editor)
Cisco Systems

Email: sajassi@cisco.com

G. Badoni
Cisco Systems

Email: gbadoni@cisco.com

P. Warade
Cisco Systems

Email: pwarade@cisco.com

S. Pasupula
Cisco Systems

Email: surpasup@cisco.com

J. Drake (editor)
Juniper

Email: jdrake@juniper.net

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

Email: jorge.rabadan@nokia.com

BESS WG
Internet-Draft
Intended status: Standards Track
Expires: 12 January 2022

Y. Wang
ZTE Corporation
11 July 2021

AC-Influenced DF Election per EVI
draft-wang-bess-evpn-ac-df-per-evi-01

Abstract

The AC-influenced DF Election (AC-DF) per [RFC8584] is too dependent on EAD/EVI routes. For example, when no failures occurred on an ESI, that AC-DF procedures will give incorrect results if no EAD/EVI routes are advertised. But in some light-weighted EVPNs (i.e. PBB EVPNs), no EAD/EVI routes will be advertised at all.

This draft proposes some new extensions to support AC-influenced DF Election without any EAD/EVI routes advertised in advance of any <ESI, EVI> failures.

Status of This Memo

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

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

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

This Internet-Draft will expire on 12 January 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components

extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Terminology	3
2. AC-DF per EVI	3
2.1. Use Case	3
2.2. AC-DF per EVI Capability	4
2.2.1. Capability Negotiation Procedures	4
2.2.2. AC-DF Capability vs AC-DF per EVI Capability	4
2.3. DF Election Procedures	4
2.3.1. Initiation of AC-DF per EVI Mode	4
2.3.2. Reverse EAD/EVI Route	5
2.3.3. DF Election Rules	5
2.3.4. Route Filtering and RT Constraints Mechanism	6
3. Other considerations	6
3.1. Reverse EAD/EVI Route in PBB EVPNs	6
3.2. Why no EAD/EVI Routes Advertised	6
4. Comparison with other solutions	7
5. Security Considerations	8
6. IANA Considerations	8
6.1. New DF Election Capability	8
6.2. New EVPN Layer 2 Attributes Control Flags	8
7. Acknowledgements	8
8. Normative References	8
9. Informative References	9
Author's Address	9

1. Introduction

When the EAD/EVI route is not advertised before the corresponding ESI sub-interface fails, The AC-influenced DF Election procedures should elect the right DF before and after that failure.

Note that according to [RFC8584], the AC-influenced DF Election will be incorrect when no EAD/EVI route is advertised, even if no ESI sub-interface has failed at all.

This draft proposes some extension to DF-Capability negotiation and DF-Election procedures to support AC-influenced DF-election when no EAD/EVI routes are advertised.

1.1. Terminology

Most of the terminology used in this documents comes from [RFC8584] and [RFC7623] except for the following:

- * Light-weighted EVPN: The EVPN solution without EAD/EVI Route advertisement.
- * EAD/ES: Ethernet A-D route per EVI, or RT-1 per ES route.
- * EAD/EVI: Ethernet A-D route per EVI, or RT-1 per EVI route.

2. AC-DF per EVI

2.1. Use Case

The ethernet segment ES1's ESI is ESI1, AC1/AC2 is two sub-interfaces on ES1, and AC3/AC4 is two sub-interfaces on ES2. AC1 and AC3 are attached to EVPN Instance EVI1, while AC2 and AC4 are attached to EVI2. The redundancy mode of ES1 is all-active.

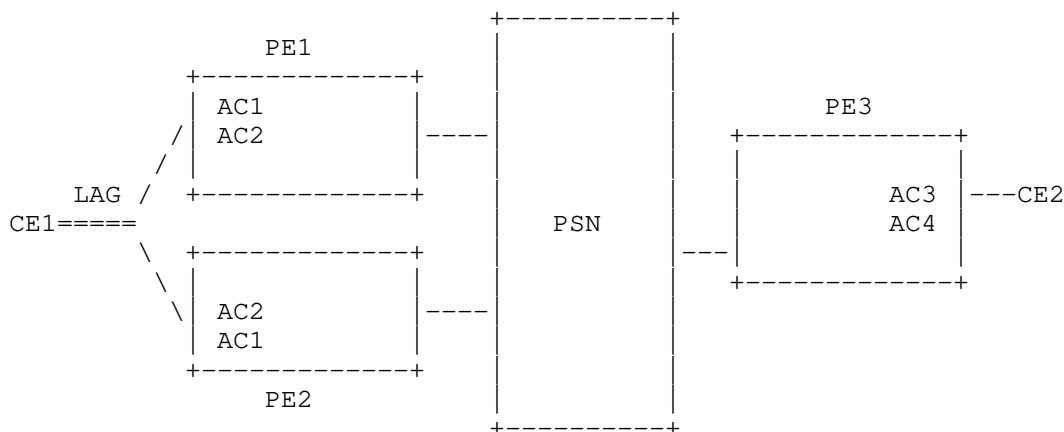


Figure 1: AC-DF per EVI Usecase

EVI1 and EVI2 are two EVPN Instances, and there are no EAD/EVI routes advertised for them. Such EVPN Instances are called Light-weighted EVPNs in this draft. For example, EVI1 and EVI2 can be the I-Components of PBB EVPNs, because no EAD/EVI routes are advertised for these I-Components.

Initially PE1 is the DF for <ESI1, EVI1>, and PE2 is the DF for <ESI1, EVI2>. When the AC1 of PE1 fails (but ESI1 and AC2 of PE1 still works well), the DF for <ESI1, EVI1> should switch to PE2.

The DF-election should be done without any EAD/EVI routes advertised before any ESI sub-interface fails. Otherwise those EAD/EVI routes are advertised just for the adaptation of AC-influenced DF-election procedures per [RFC8584], but will do nothing good for the unicast packet forwarding in data plane (because data plane MAC learning don't use EAD/EVI routes).

2.2. AC-DF per EVI Capability

We introduce a new bit named "AC-DF per EVI" in the "bitmap" field of the DF Election extended community. The "AC-DF per EVI" bit means that the DF-Election for an <ESI, EVI> will not take EAD/EVI routes into considerations untill a Reverse EAD/EVI route for that <ESI, EVI> is received from any of the PEs.

2.2.1. Capability Negotiation Procedures

Only when all RT-4 routes of the same ESI indicate the "AC-DF per EVI" Capability, The DF Election will be executed in "AC-DF per EVI" mode. We can say that the DF Election mode for that ESI is negotiated as "AC-DF per EVI" mode.

Note that when any of the PEs on that ES is an old PE that don't support "AC-DF per EVI" mode, the RT-4 route from that PE will not indicate the "AC-DF per EVI" Capability, So the DF election will not be executed in "AD-DF per EVI" mode. This is for compatibility purpose.

2.2.2. AC-DF Capability vs AC-DF per EVI Capability

When all RT-4 routes of the same ESI indicate both the "AC-DF per EVI" Capability and the old AC-DF Capability, The DF Election will be executed in "AD-DF per EVI" mode.

Note that when any of the PEs on that ES is an old PE that don't support "AC-DF per EVI" mode, the RT-4 route from that PE may indicate the "old AC-DF" Capability only, So the DF election will still be executed in old AD-DF mode per [RFC8584]. This is for compatibility purpose.

2.3. DF Election Procedures

2.3.1. Initiation of AC-DF per EVI Mode

According to [RFC8584], the AC-influenced DF Election will be incorrect when no EAD/EVI route is advertised, even if no ESI sub-interface has failed at all.

But when the DF Election mode for an ESI is negotiated as AC-DF per EVI mode, no normal EAD/EVI routes can impact the DF Election procedures. The DF election will be done following that ESI's RT-4 routes only until at least one reverse EAD/EVI route is received.

2.3.2. Reverse EAD/EVI Route

In order to do AC-influenced DF election after a sub-interface of that ES fails, we introduce the "Reverse EAD/EVI Route". The reverse EAD/EVI Route is a special type of EAD/EVI Route that is advertised on the failure of corresponding <ESI, EVI>, not on the activation of that <ESI, EVI>.

Reverse EAD/EVI routes can use the same format as EAD/EVI Routes except for the following differences:

- * A Reverse EAD/EVI Route carries an EVPN Layer 2 Attributes Extended Community whose "Control Flags" field includes a new flag named "AC Down". The "AC Down" flag means that the corresponding AC (for which the Reverse EAD/EVI route is advertised) is down.
- * It is recommended to carry an ES-Import RT ([RFC7432]) and an EVI-RT ([I-D.ietf-bess-evpn-igmp-mld-proxy]) along with a reverse EAD/EVI route, no traditional Route-Targets have to be carried for DF-election purpose.
- * A Reverse EAD/EVI Route should make its MPLS label field be set to zero.

Note that when the corresponding sub-interface fails, the MP_REACH_NLRI of the reverse EAD/EVI route is advertised, and when the corresponding sub-interface recovers, the MP_UNREACH_NLRI of the reverse EAD/EVI route is sent. This is the opposite of normal EAD/EVI route. So it is called as reverse EAD/EVI route.

2.3.3. DF Election Rules

When a Reverse EAD/EVI Route for an <ESI, EVI> is received from a remote PE X, the RT-4 Route of that PE x are expelled from that <ESI, EVI>'s DF election. Then the DF election for that <ESI, EVI> will be updated according to the corresponding DF Alg.

Note that the DF election for other <ESI, EVI>s will not be affected by that Reverse EAD/EVI Route.

2.3.4. Route Filtering and RT Constraints Mechanism

When PE Y receives a reverse EAD/EVI route from PE X, and the ES-Import RT of that route can't match any local ES of PE Y, PE Y will not import that route into the EVI that is identified by that route's EVI-RT.

When RT Constraints Mechanism is enabled, each reverse EAD/EVI route will be distributed to the adjacent PEs of its ES only. Because that only the ES-Import RT are visible to the RT Constraints Mechanism, The EVI-RT is not visible to the RT Constraints Mechanism.

3. Other considerations

3.1. Reverse EAD/EVI Route in PBB EVPNs

The AC-DF per EVI mode is not confined to PBB EVPN which is just an example of light-weighted EVPNs. But in PBB EVPN, the construction of reverse EAD/EVI route need some special considerations.

- * It's EVI-RT should be the export route-target of the B-Component, not the C-Component.

- * It's Ethernet Tag ID (ETI) should be the I-SID of the I-Component.

Note that when PE Z receives a reverse EAD/EVI route whose EVI-RT matches a local B-Component but whose ETI matches none of the I-Components of that B-Component, PE Z may not import that reverse EAD/EVI route.

Note that the reverse EAD/EVI route don't have to carry any B-MAC along with it. Because that the B-MAC can do nothing helpful for the DF election.

3.2. Why no EAD/EVI Routes Advertised

When no RT-2 Routes advertised, no EAD/EVI routes need to be advertised either. PBB EVPN is an example of that. In PBB EVPN, the C-MACs are learnt in the data plane.

Other light-weighted EVPNs also do data plane C-MAC learning, so they don't have to advertise EAD/EVI routes either. In such EVPNs, AC-DF per EVI will help.

4. Comparison with other solutions

PBB EVPN can assign a dedicated vES to each sub-interface, in such case, the RT-4 routes are advertised per each sub-interface (or vESI). But this will bring out some other disadvantages:

- * The uniformity of service carving can't be achieved without careful configuring.

With service carving, it is possible to elect multiple DFs per ES (one per EVI) in order to perform load balancing of traffic destined to a given ES. The objective is that the load-balancing procedures should carve up the BD space among the redundant PE nodes evenly, in such a way that every PE is the DF for a distinct set of EVIs.

When each EVI use a dedicated vESI to advertise the corresponding Ethernet Segment Routes for that <ES, EVI>, The service carving mechanisms can not work without manual configuration.

- * The amount of B-MACs will be greatly increased.

The brief comparisons are listed as the following table:

Items	AC-DF per EVI	vESI
ESIs	one per port	one per sub-interface
RT-4 Routes	one per port	one per sub-interface
B-MACs	one per port	one per sub-interface
Service Carving	auto	manual-configured
EAD per EVI routes	none	none
Reverse EAD per EVI routes	one per failed sub-interfaces	none

Table 1: Comparisons with vESI for PBB-EVPN

Using AC-DF per EVI mode, the service-carving is automatically achieved, and no extra B-MACs should be configured and advertised.

5. Security Considerations

Security considerations will be added in future versions.

6. IANA Considerations

6.1. New DF Election Capability

IANA will be requested to allocate a new DF Election Capability in the "DF Election Capabilities" Registry. This capability is called "AC-DF per EVI Capability".

Bit ----	Name -----	Reference -----
4	AC-DF per EVI Capability	This draft

Figure 2: AC-DF per EVI Capability

6.2. New EVPN Layer 2 Attributes Control Flags

IANA will be requested to allocate a new Control Flag in the "EVPN Layer 2 Attributes Control Flags" Registry. This Control Flag is called "D" Flag, where "D" means AC-Down.

Bit ----	Name -----	Reference -----
D	AC-Down on Advertising PE	This Draft

Figure 3: AC Failure Flag

7. Acknowledgements

The authors would like to thank the following for their comments and review of this document:

Chunning Dai.

8. Normative References

[I-D.ietf-bess-evpn-igmp-mld-proxy]
 Sajassi, A., Thoria, S., Mishra, M. P., Patel, K., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-igmp-mld-proxy-09, 19 April 2021,
<https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-igmp-mld-proxy-09>.

[I-D.ietf-bess-srv6-services]

Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay Services", Work in Progress, Internet-Draft, draft-ietf-bess-srv6-services-07, 11 April 2021, <<https://datatracker.ietf.org/doc/html/draft-ietf-bess-srv6-services-07>>.

[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>>.

[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>>.

[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>>.

9. Informative References

[RFC7041] Balus, F., Ed., Sajassi, A., Ed., and N. Bitar, Ed., "Extensions to the Virtual Private LAN Service (VPLS) Provider Edge (PE) Model for Provider Backbone Bridging", RFC 7041, DOI 10.17487/RFC7041, November 2013, <<https://www.rfc-editor.org/info/rfc7041>>.

[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>>.

[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>>.

Author's Address

Yubao Wang
ZTE Corporation
No.68 of Zijinghua Road, Yuhuatai District
Nanjing
China

Email: wang.yubao2@zte.com.cn

BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 13, 2021

W. Wang
A. Wang
China Telecom
H. Wang
Huawei Technologies
March 12, 2021

Layer-3 Accessible EVPN Services
draft-wang-bess-l3-accessible-evpn-04

Abstract

This draft describes layer-3 accessible EVPN service interfaces according to [RFC7432], and proposes a new solution which can simplify the deployment of layer-3 accessible EVPN service. This solution allows each PE in EVPN network to maintain only one IP-VRF.

Status of This Memo

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

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

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

This Internet-Draft will expire on September 13, 2021.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

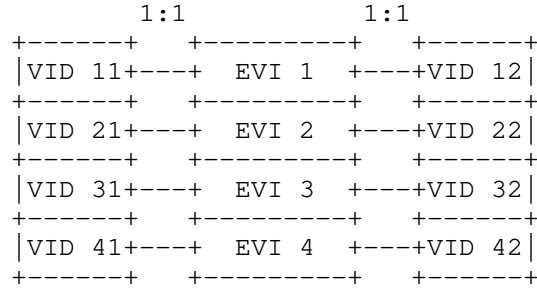
the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

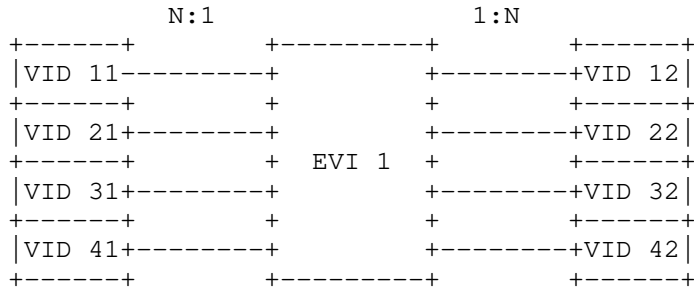
1. Introduction	2
2. Conventions used in this document	4
3. Terminology	4
4. Service Interfaces in layer-3 accessible EVPN	5
5. Solutions of LSI-aware bundle service interface	6
6. Protocol Extensions	8
6.1. Forwarding Plane	8
6.1.1. Extensions to VxLAN	8
6.2. Control Plane	8
7. Security Considerations	9
8. IANA Considerations	9
9. Normative References	9
Authors' Addresses	10

1. Introduction

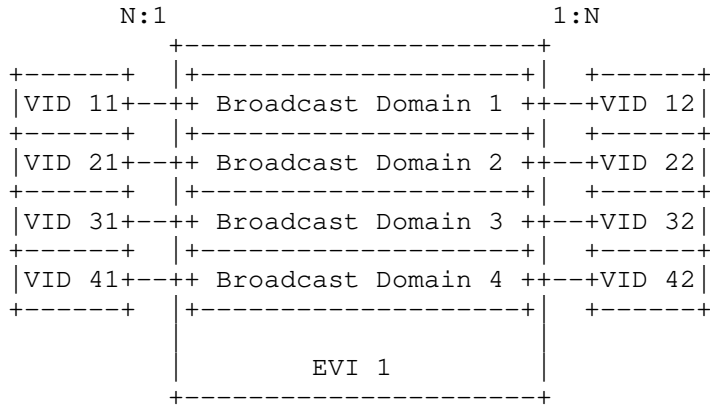
[RFC7432] defines three service interfaces for layer-2 accessible EVPN: VLAN-Based Service Interface, VLAN-Bundle Service Interface and VLAN-Aware Bundle Service Interface. These three types of service interfaces can realize the isolation of layer-2 traffic of customers in different ways, as shown in Figure 1.



VLAN-based Service Interface



VLAN-bundle Service Interface



VLAN-Aware Bundle Service Interface

Figure 1: EVPN Service Interfaces Overview

For VLAN-based service interface, there is a one to one mapping between VID and EVI. Each EVI has a single broadcast domain so that traffic from different customers can be isolated.

For VLAN-bundle service interface, there is a N to one mapping between VID and EVI. Each EVI has a single broadcast domain, but the MAC address MUST be unique that can be used for customer traffic isolation.

For VLAN-aware bundle service interface, there is a N to one mapping between VID and EVI. Each EVI has multiple broadcast domains while the MAC address can overlap. One broadcast domain corresponds to one VID, which can be used to customer traffic isolation.

In the scenarios corresponding to these service interfaces, CE-PE should be placed in the same Layer-2 network. In most of provider network, CE-PE need to cross a Layer-3 network, then the above service interfaces should be extended to adapt to the layer-3 network.

In this draft, we describe three layer-3 accessible interfaces for EVPN, summarize the existing layer-3 accessible EVPN solutions, and propose a new solution which can simplify the depolyment of layer-3 accessible EVPN service.

2. Conventions used in this document

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] .

3. Terminology

The following terms are defined in this draft:

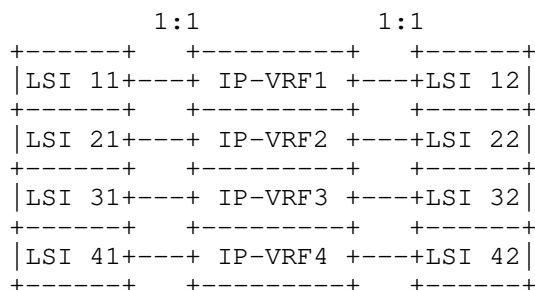
- o CE: Client Edge
- o PE: Provider Edge
- o EVPN: BGP/MPLS Ethernet VPN, defined in [RFC7432]
- o VxLAN: Virtual eXtensible Local Area Network, defined in [RFC7348]
- o IPSec: Internet Protocol Security, defined in [RFC4301]

4. Service Interfaces in layer-3 accessible EVPN

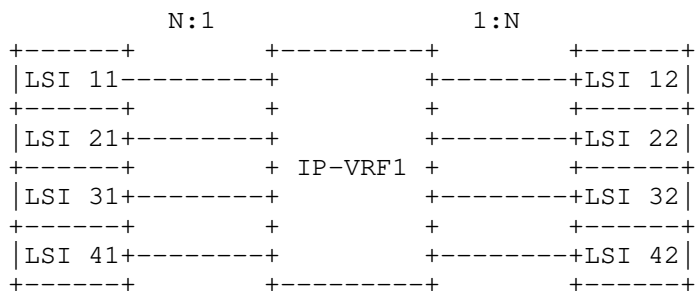
In most of provider network, CE-PE need to cross a Layer-3 network. With this scenario, service interfaces defined in [RFC7432] should be extended to adapt to the layer-3 network. To achieve the traffic isolation, tunnel encapsulation technologies can be used.

We define Logical Session Identifier(LSI) to distinguish the packets from different tunnels, which is related to VNI/SPI. The length of LSI is 16 bits.

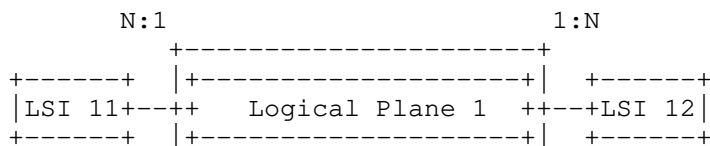
The layer-3 accessible interfaces for EVPN are shown in Figure 2, refer to [RFC7432]

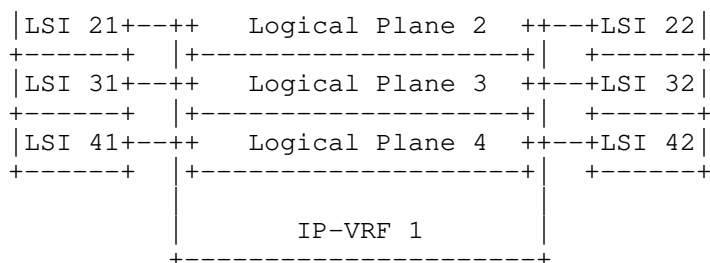


LSI-based Service Interface



LSI-bundle Service Interface





LSI-Aware Bundle Service Interface

Figure 2: Layer-3 accessible EVPN Service Interfaces Overview

For LSI-based service interface, there is a one to one mapping between LSI and IP-VRF. Each IP-VRF has a single logical plane so that traffic from different customers can be isolated.

For LSI-bundle service interface, there is a N to one mapping between LSI and IP-VRF. Each IP-VRF has a single logical plane, but the IP address MUST be unique that can be used for customer traffic isolation.

For LSI-aware bundle service interface, there is a N to one mapping between LSI and IP-VRF. Each IP-VRF has multiple logical planes while the IP address can overlap. One logical plane corresponds to one LSI, which can be used to customer traffic isolation.

5. Solutions of LSI-aware bundle service interface

Let's assume a scenario as shown in Figure 3. PE1, PE2 and PE3 are EVPN peers, the customer data transmission between PEs relies on VxLAN. CE1, CE2 and CE3 are connected to the sites of customer for its department A and B.

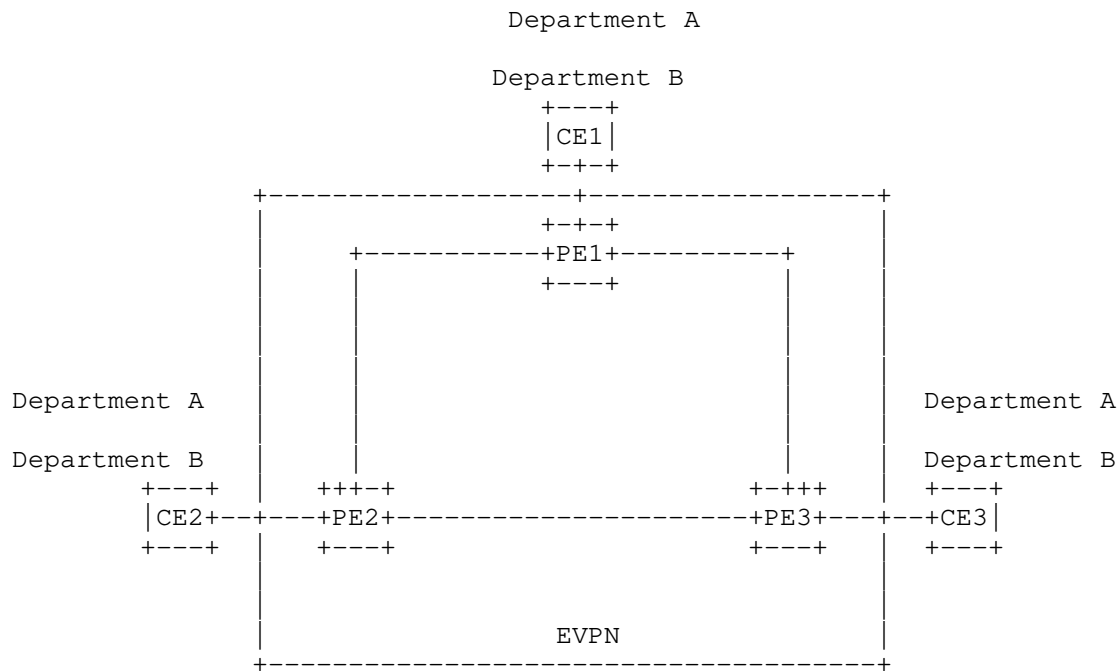


Figure 3: LSI-aware bundle service interface scenario

If each VNI has its own IP-VRF, each PE and CE maintain an IP-VRF for each deployment. In this situation, customer traffic can be isolated by different VNIs, and there is no need for extending control plane/forwarding plane protocols.

For deployment, we expect a simpler way, such as assign an IP-VRF to each customer, not to each department. That is to say, all VNIs share one IP-VRF on PEs. In this situation, each CE still maintain an IP-VRF for each deployment, but each PE maintains only one VRF for all deployments. In this situation, customer traffic cannot be isolated by VNIs. We propose a solution for this scenario:

- o Using LSI information to identify different customer routes / traffic. As described above, LSI can be generated by VNI/SPI, and there is a one to one mapping between LSI and VNI/SPI. PEs should maintain the mapping table of LSI and VNI/SPI, so that they can distinguish different customer routes / traffic. LSI information can be transmitted by using Ethernet Tag ID or a newly defined ESI type.

- o TBD (more solutions are welcome).

6. Protocol Extensions

6.1. Forwarding Plane

6.1.1. Extensions to VxLAN

When the forwarding plane uses VxLAN tunnel technologies, we should extend the VxLAN GPE header to carry the LSI information, the extensions to the VxLAN GPE header is shown in Figure 4:

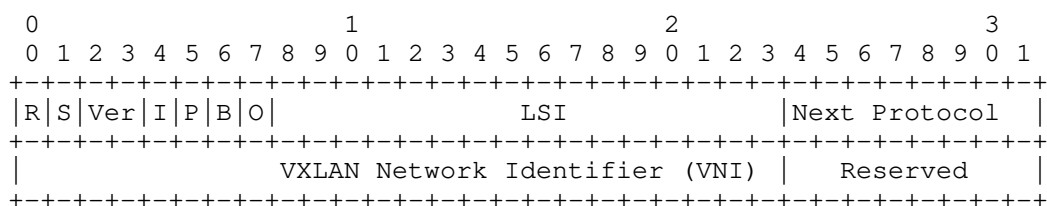


Figure 4: The extensions to VxLAN GPE header

We define a S bit. If S is set to 1, it means the field after O bit contains LSI information.

6.2. Control Plane

We proposed two methods to identify the routes that related to different LSI information:

- o Reusing the Ethernet Tag ID. This method requires the update of [I-D.ietf-bess-evpn-prefix-advertisement] (Ethernet Tag ID is set to 0 for route type 5), and may arises some confuse with the original defination of Ethernet Tag ID.
- o Using the newly defined ESI type as shown in Figure 5. This method can preserve the original purpose of ESI defination (multi-homing).

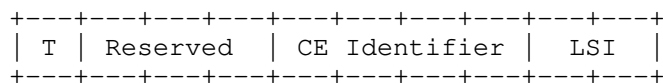


Figure 5: The format of new ESI type

Where:

- o T (1 octet): specifies the ESI Type. The recommended value is 0x06.
- o CE Identifier (3 octets): the route ID/IPv4 address of CE.
- o LSI (2 octets): the LSI information.

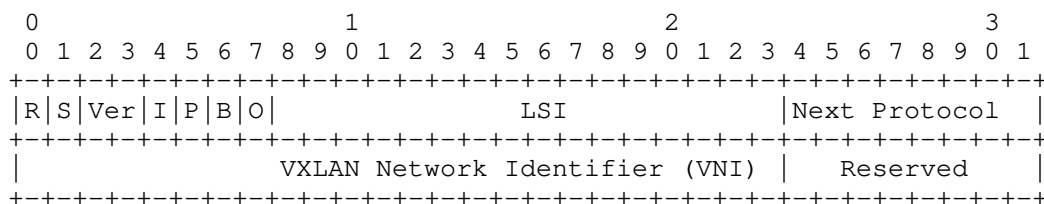
Since the length of LSI is 16 bits, while the length of Ethernet Tag ID and ESI are 80 bits and 32 bits, respectively. We can only use the lower 16 bits of Ethernet Tag ID / ESI field to carry LSI information, the other locations MUST set to 0.

7. Security Considerations

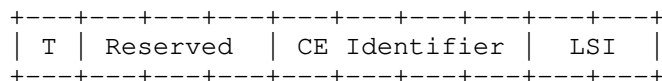
TBD

8. IANA Considerations

This draft extends the VxLAN GPE header, S bit of Flag and LSI field are added:



This draft also define a new ESI type:



9. Normative References

```
[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-bess-mvpn-evpn-aggregation-label]
```

Zhang, Z., Rosen, E., Lin, W., Li, Z., and I. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", draft-ietf-bess-mvpn-evpn-aggregation-label-05 (work in progress), January 2021.

- [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>>.
- [RFC2890] Dommetty, G., "Key and Sequence Number Extensions to GRE", RFC 2890, DOI 10.17487/RFC2890, September 2000, <<https://www.rfc-editor.org/info/rfc2890>>.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [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, <<https://www.rfc-editor.org/info/rfc7432>>.

Authors' Addresses

Wei Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: weiwang94@foxmail.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, Beijing 102209
China

Email: wangaj3@chinatelecom.cn

Haibo Wang
Huawei Technologies
Huawei Building, No.156 Beiqing Rd.
Beijing, Beijing 100095
China

Email: rainsword.wang@huawei.com