Network Working Group
INTERNET-DRAFT
Category: Standards Track

A. Sajassi
Cisco

R. Aggarwal
Arktan

N. Bitar
Verizon

W. Henderickx
F. Balus
Alcatel-Lucent

S. Boutros
K. Patel
S. Salam
Cisco

Aldrin Isaac
Bloomberg

J. Drake
R. Shekhar
Juniper Networks

J. Uttaro
AT&T

Expires: August 25, 2013

February 25, 2013

BGP MPLS Based Ethernet VPN
draft-ietf-l2vpn-evpn-03

Status of this Memo

Copyright and License Notice

Abstract

This document describes procedures for BGP MPLS based Ethernet VPNs (E-VPN).

Table of Contents

1. Specification of requirements 5
2. Contributors 5
3. Introduction 5
4. Terminology 5
5. BGP MPLS Based E-VPN Overview 6
6. Ethernet Segment 7
7. Ethernet Tag 9
7.1 VLAN Based Service Interface 9
7.2 VLAN Bundle Service Interface 9
7.2.1 Port Based Service Interface 10
7.3 VLAN Aware Bundle Service Interface 10
7.3.1 Port Based VLAN Aware Service Interface 10
8. BGP E-VPN NLRI 10
8.1. Ethernet Auto-Discovery Route 11
8.2. MAC Advertisement Route 12
8.3. Inclusive Multicast Ethernet Tag Route 12
8.4 Ethernet Segment Route 13
8.5 ESI Label Extended Community 13
8.6 ES-Import Route Target 14
8.7 MAC Mobility Extended Community 14
8.8 Default Gateway Extended Community 15
9. Multi-homing Functions 15
9.1 Multi-homed Ethernet Segment Auto-Discovery 15
9.1.1 Constructing the Ethernet Segment Route 15
9.2 Fast Convergence 16
9.2.1 Constructing the Ethernet A-D Route per Ethernet Segment 16
9.2.1.1. Ethernet A-D Route Targets 17
9.3 Split Horizon 17
9.3.1 ESI Label Assignment 18
9.3.1.1 Ingress Replication 18

9.3.1.2. P2MP MPLS LSPs 19
9.3.1.3. MP2MP LSPs 20
9.4 Aliasing and Backup-Path 20
9.4.1 Constructing the Ethernet A-D Route per EVI 21
9.4.1.1 Ethernet A-D Route Targets 22
9.5 Designated Forwarder Election 22
10. Determining Reachability to Unicast MAC Addresses 24
10.1. Local Learning 25
10.2. Remote learning 25
10.2.1. Constructing the BGP E-VPN MAC Address Advertisement . 25
10.2.2 Route Resolution 27
11. ARP and ND 28
11.1 Default Gateway 29
12. Handling of Multi-Destination Traffic 29
12.1. Construction of the Inclusive Multicast Ethernet Tag Route 30
12.2. P-Tunnel Identification 30
13. Processing of Unknown Unicast Packets 31
13.1. Ingress Replication 32
13.2. P2MP MPLS LSPs 32
14. Forwarding Unicast Packets 32
14.1. Forwarding packets received from a CE 32
14.2. Forwarding packets received from a remote PE 34
14.2.1. Unknown Unicast Forwarding 34
14.2.2. Known Unicast Forwarding 34
15. Load Balancing of Unicast Frames 34
15.1. Load balancing of traffic from an PE to remote CEs 34
15.1.1 Single-Active Redundancy Mode 34
15.1.2 All-Active Redundancy Mode 35
15.2. Load balancing of traffic between an PE and a local CE . . 37
15.2.1. Data plane learning 37
15.2.2. Control plane learning 37
16. MAC Mobility 37
16.1. MAC Duplication Issue 39
17. Multicast 39
17.1. Ingress Replication 40
17.2. P2MP LSPs 40
17.3. MP2MP LSPs 40
17.3.1. Inclusive Trees 40
17.3.2. Selective Trees 41
17.4. Explicit Tracking 42
18. Convergence 42
18.1. Transit Link and Node Failures between PEs 42
18.2. PE Failures 42
18.2.1. Local Repair 42
18.3. PE to CE Network Failures 42
19. LACP State Synchronization 43
20. Acknowledgements 44

21. Security Considerations 44
22. IANA Considerations 44
23. References 44
23.1 Normative References 44
23.2 Informative References 45
24. Author's Address 45

1. Specification of requirements

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

2. Contributors

In addition to the authors listed above, the following individuals also contributed to this document:

Quaizar Vohra
Kireeti Kompella
Apurva Mehta
Nadeem Mohammad
Juniper Networks

Clarence Filsfils
Dennis Cai
Cisco

3. Introduction

This document describes procedures for BGP MPLS based Ethernet VPNs (E-VPN). The procedures described here are intended to meet the requirements specified in [EVPN-REQ]. Please refer to [EVPN-REQ] for the detailed requirements and motivation. E-VPN requires extensions to existing IP/MPLS protocols as described in this document. In addition to these extensions E-VPN uses several building blocks from existing MPLS technologies.

4. Terminology

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

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

E-VPN Instance (EVI): An E-VPN routing and forwarding instance on a PE.

Ethernet segment identifier (ESI): If a CE is multi-homed to two or more PEs, the set of Ethernet links that attaches the CE to the PEs is an ’Ethernet segment’. Ethernet segments MUST have a unique non-zero identifier, the ’Ethernet Segment Identifier’.

Ethernet Tag: An Ethernet Tag identifies a particular broadcast domain, e.g., a VLAN. An E-VPN instance consists of one or more broadcast domains. Ethernet tag(s) are assigned to the broadcast domains of a given E-VPN instance by the provider of that E-VPN, and each PE in that E-VPN instance performs a mapping between broadcast domain identifier(s) understood by each of its attached CEs and the corresponding Ethernet tag.

Link Aggregation Control Protocol (LACP):

Multipoint to Multipoint (MP2MP):

Point to Multipoint (P2MP):

Point to Point (P2P):

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

5. BGP MPLS Based E-VPN Overview

This section provides an overview of E-VPN.

An E-VPN comprises CEs that are connected to PEs that form the edge of the MPLS infrastructure. A CE may be a host, a router or a switch. The PEs provide virtual Layer 2 bridged connectivity between the CEs. There may be multiple E-VPNs in the provider’s network.

The PEs may be connected by an MPLS LSP infrastructure which provides the benefits of MPLS technology such as fast-reroute, resiliency, etc. The PEs may also be connected by an IP infrastructure in which case IP/GRE tunneling or other IP tunneling can be used between the PEs. The detailed procedures in this version of this document are specified only for MPLS LSPs as the tunneling technology. However these procedures are designed to be extensible to IP tunneling as the PSN tunneling technology.

In an E-VPN, MAC learning between PEs occurs not in the data plane (as happens with traditional bridging) but in the control plane. Control plane learning offers greater control over the MAC learning process, such as restricting who learns what, and the ability to apply policies. Furthermore, the control plane chosen for advertising MAC reachability information is multi-protocol (MP) BGP (similar to IP VPNs (RFC 4364)). This provides greater scalability

and the ability to preserve the "virtualization" or isolation of groups of interacting agents (hosts, servers, virtual machines) from each other. In E-VPN, PEs advertise the MAC addresses learned from the CEs that are connected to them, along with an MPLS label, to other PEs in the control plane using MP-BGP. Control plane learning enables load balancing of traffic to and from CEs that are multi-homed to multiple PEs. This is in addition to load balancing across the MPLS core via multiple LSPs between the same pair of PEs. In other words it allows CEs to connect to multiple active points of attachment. It also improves convergence times in the event of certain network failures.

However, learning between PEs and CEs is done by the method best suited to the CE: data plane learning, IEEE 802.1x, LLDP, 802.1aq, ARP, management plane or other protocols.

It is a local decision as to whether the Layer 2 forwarding table on an PE is populated with all the MAC destination addresses known to the control plane, or whether the PE implements a cache based scheme. For instance the MAC forwarding table may be populated only with the MAC destinations of the active flows transiting a specific PE.

The policy attributes of E-VPN are very similar to those of IP-VPN. An EVI requires a Route-Distinguisher (RD) and one or more Route-Targets (RTs). A CE attaches to an E-VPN instance (EVI) on an PE, on an Ethernet interface which may be configured for one or more Ethernet Tags, e.g., VLANs. Some deployment scenarios guarantee uniqueness of VLANs across E-VPNs: all points of attachment of a given EVI use the same VLAN, and no other EVI uses this VLAN. This document refers to this case as a "Unique VLAN E-VPN" and describes simplified procedures to optimize for it.

6. Ethernet Segment

If a CE is multi-homed to two or more PEs, the set of Ethernet links constitutes an "Ethernet Segment". An Ethernet segment may appear to the CE as a Link Aggregation Group (LAG). Ethernet segments have an identifier, called the "Ethernet Segment Identifier" (ESI) which is encoded as a ten octets integer. The following two ESI values are reserved:

- ESI 0 denotes a single-homed CE.

- ESI {0xFF} (repeated 10 times) is known as MAX-ESI and is reserved.

In general, an Ethernet segment MUST have a non-reserved ESI that is unique network wide (e.g., across all EVPNs on all the PEs). If the

CE(s) constituting an Ethernet Segment is (are) managed by the network operator, then ESI uniqueness should be guranteed; however, if the CE(s) is (are) not managed, then the operator MUST configure a network-wide unique ESI for that Ethernet Segment. This is required to enable auto-discovery of Ethernet Segments and DF election. The ESI can be assigned using various mechanisms:

1. If IEEE 802.1AX LACP is used between the PEs and CEs, then the ESI is determined from LACP by concatenating the following parameters:

   + CE LACP System Identifier comprised of two octets of System Priority and six octets of System MAC address, where the System Priority is encoded in the most significant two octets. The CE LACP identifier MUST be encoded in the high order eight octets of the ESI.

   + CE LACP two octets Port Key. The CE LACP port key MUST be encoded in the low order two octets of the ESI.

   As far as the CE is concerned, it would treat the multiple PEs that it is connected to as the same switch. This allows the CE to aggregate links that are attached to different PEs in the same bundle.

   This mechanism could be used only if it produces ESIs that satisfy the uniqueness requirement specified above.

   2. If LLDP is used between the PEs and CEs that are hosts, then the ESI is determined by LLDP. The ESI will be specified in a following version.

   This mechanism could be used only if it produces ESIs that satisfy the uniqueness requirement specified above.

   3. In the case of indirectly connected hosts via a bridged LAN between the CEs and the PEs, the ESI is determined based on the Layer 2 bridge protocol as follows: If MST is used in the bridged LAN then the value of the ESI is derived by listening to BPDUs on the Ethernet segment. To achieve this the PE is not required to run MST. However the PE must learn the Root Bridge MAC address and Bridge Priority of the root of the Internal Spanning Tree (IST) by listening to the BPDUs. The ESI is constructed as follows:

   {Bridge Priority (16 bits) , Root Bridge MAC Address (48 bits)}

   This mechanism could be used only if it produces ESIs that satisfy

the uniqueness requirement specified above.

4. The ESI may be configured.

7. Ethernet Tag

An Ethernet Tag identifies a particular broadcast domain, e.g. a VLAN, in an EVI. An EVI consists of one or more broadcast domains. Ethernet Tags are assigned to the broadcast domains of a given EVI by the provider of the E-VPN service. Each PE, in a given EVI, performs a mapping between the Ethernet Tag and the corresponding broadcast domain identifier(s) understood by each of its attached CEs (e.g. CE VLAN Identifiers or CE-VIDs).

If the broadcast domain identifiers(s) are understood consistently by all of the CEs in an EVI, the broadcast domain identifier(s) MAY be used as the corresponding Ethernet Tag(s). In other words, the Ethernet Tag ID assigned by the provider is numerically equal to the broadcast domain identifier (e.g., CE-VID = Ethernet Tag).

Further, some deployment scenarios guarantee uniqueness of broadcast domain identifiers across all EVIs; all points of attachment of a given EVI use the same broadcast domain identifier(s) and no other EVI uses these broadcast domain identifier(s). This allows the RT(s) for each EVI to be derived automatically, as described in section 9.4.1.1.1 "Auto-Derivation from the Ethernet Tag ID".

The following subsections discuss the relationship between Ethernet Tags, EVIs and broadcast domain identifiers as well as the setting of the Ethernet Tag Identifier, in the various E-VPN BGP routes (defined in section 8), for the different types of service interfaces described in [EVPN-REQ].

7.1 VLAN Based Service Interface

With this service interface, there is a one-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there is a single bridge domain per PE for the EVI. Different CEs connected to different PE ports MAY use different broadcast domain identifiers (e.g. CE-VIDs) for the same EVI. If said identifiers are different, the frames SHOULD remain tagged with the originating CE’s broadcast domain identifier (e.g. CE-VID). When the CE broadcast domain identifiers are not consistent, a tag translation function MUST be supported in the data path and MUST be performed on the disposition PE. The Ethernet Tag Identifier in all E-VPN routes MUST be set to 0.

7.2 VLAN Bundle Service Interface

With this service interface, there is a many-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there is a single bridge domain per PE for the EVI. Different CEs connected to different PE ports MUST use the same broadcast domain identifiers (e.g. CE-VIDs) for the same EVI. The MPLS encapsulated frames MUST remain tagged with the originating CE's broadcast domain identifier (e.g. CE-VID). Tag translation is NOT permitted. The Ethernet Tag Identifier in all E-VPN routes MUST be set to 0.

7.2.1 Port Based Service Interface

This service interface is a special case of the VLAN Bundle service interface, where all of the VLANs on the port are part of the same service and map to the same bundle. The procedures are identical to those described in section 7.2.

7.3 VLAN Aware Bundle Service Interface

With this service interface, there is a many-to-one mapping between the broadcast domain identifier understood by a CE on a port (e.g. CE-VID) and an EVI. Furthermore, there are multiple bridge domains per PE for the EVI: one broadcast domain per CE broadcast domain identifier. In the case where the CE broadcast domain identifiers are not consistent for different CEs, a normalized Ethernet Tag MUST be carried in the MPLS encapsulated frames and a tag translation function MUST be supported in the data path. This translation MUST be performed on both the imposition as well as the disposition PEs. The Ethernet Tag Identifier in all E-VPN routes MUST be set to the normalized Ethernet Tag assigned by the E-VPN provider.

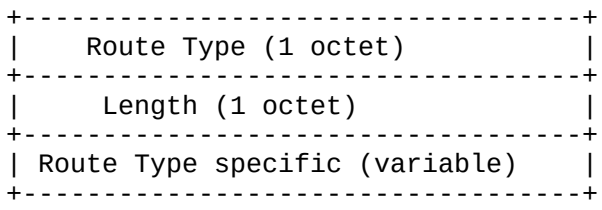
7.3.1 Port Based VLAN Aware Service Interface

This service interface is a special case of the VLAN Aware Bundle service interface, where all of the VLANs on the port are part of the same service and map to the same bundle. The procedures are identical to those described in section 7.3.

8. BGP E-VPN NLRI

This document defines a new BGP NLRI, called the E-VPN NLRI.

Following is the format of the E-VPN NLRI:



The Route Type field defines encoding of the rest of the E-VPN NLRI (Route Type specific E-VPN NLRI).

The Length field indicates the length in octets of the Route Type specific field of E-VPN NLRI.

This document defines the following Route Types:

+ 1 - Ethernet Auto-Discovery (A-D) route
+ 2 - MAC advertisement route
+ 3 - Inclusive Multicast Route
+ 4 - Ethernet Segment Route

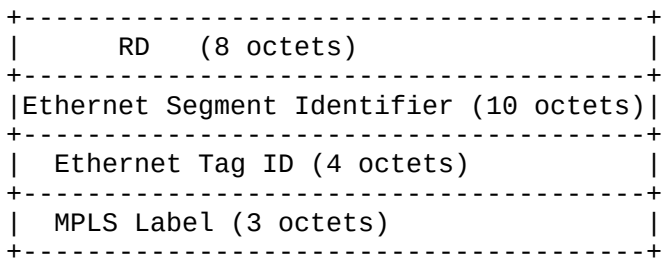
The detailed encoding and procedures for these route types are described in subsequent sections.

The E-VPN NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an AFI of 25 (L2VPN) and a SAFI of 70 (E-VPN). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the E-VPN NLRI (encoded as specified above).

In order for two BGP speakers to exchange labeled E-VPN NLRI, they must use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of 25 (L2VPN) and a SAFI of 70 (E-VPN).

8.1. Ethernet Auto-Discovery Route

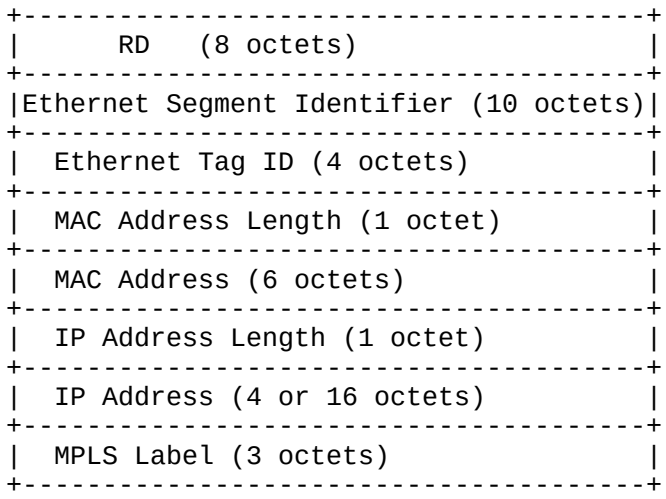
A Ethernet A-D route type specific E-VPN NLRI consists of the following:



For procedures and usage of this route please see section 9.2 "Fast Convergence" and section 9.4 "Aliasing".

8.2. MAC Advertisement Route

A MAC advertisement route type specific E-VPN NLRI consists of the following:

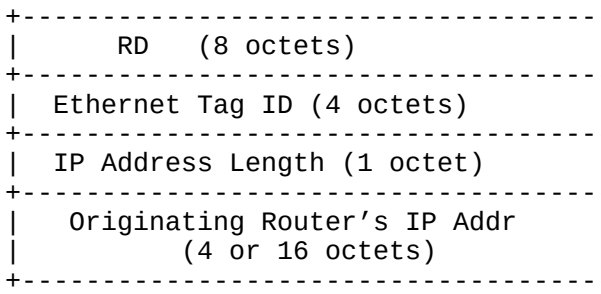


For the purpose of BGP route key processing, only the Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP Address Address fields are considered to be part of the prefix in the NLRI. The Ethernet Segment Identifier and MPLS Label fields are to be treated as route attributes as opposed to being part of the "route".

For procedures and usage of this route please see section 10 "Determining Reachability to Unicast MAC Addresses" and section 15 "Load Balancing of Unicast Packets".

8.3. Inclusive Multicast Ethernet Tag Route

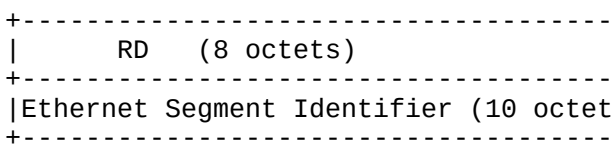
An Inclusive Multicast Ethernet Tag route type specific E-VPN NLRI consists of the following:



For procedures and usage of this route please see section 12 "Handling of Multi-Destination Traffic", section 13 "Processing of Unknown Unicast Traffic" and section 17 "Multicast".

8.4 Ethernet Segment Route

The Ethernet Segment Route is encoded in the E-VPN NLRI using the Route Type value of 4. The Route Type Specific field of the NLRI is formatted as follows:

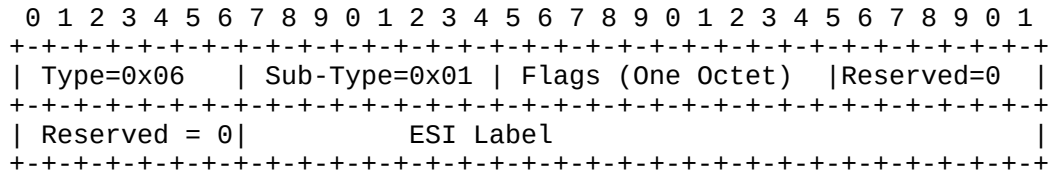


For procedures and usage of this route please see section 9.5 "Designated Forwarder Election".

8.5 ESI Label Extended Community

This extended community is a new transitive extended community with the Type field is 0x06, and the Sub-Type of 0x01. It may be advertised along with Ethernet Auto-Discovery routes and it enables split-horizon procedures for multi-homed sites as described in section 9.3 "Split Horizon".

Each ESI Label Extended Community is encoded as a 8-octet value as follows:

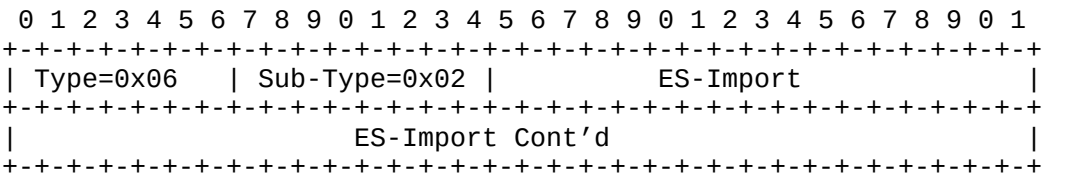


The low order bit of the flags octet is defined as the "Active-Standby" bit and may be set to 1. A value of 0 means that the multi-homed site is operating in All-Active mode; whereas, a value of 1 means that the multi-homed site is operating in Single-Active mode.

The second low order bit of the flags octet is defined as the "Root-Leaf". A value of 0 means that this label is associated with a Root site; whereas, a value of 1 means that this label is associate with a Leaf site. The other bits must be set to 0.

8.6 ES-Import Route Target

This is a new transitive Route Target extended community carried with the Ethernet Segment route. When used, it enables all the PEs connected to the same multi-homed site to import the Ethernet Segment routes. The value is derived automatically from the ESI by encoding the 6-byte MAC address portion of the ESI in the ES-Import Route Target. The format of this extended community is as follows:



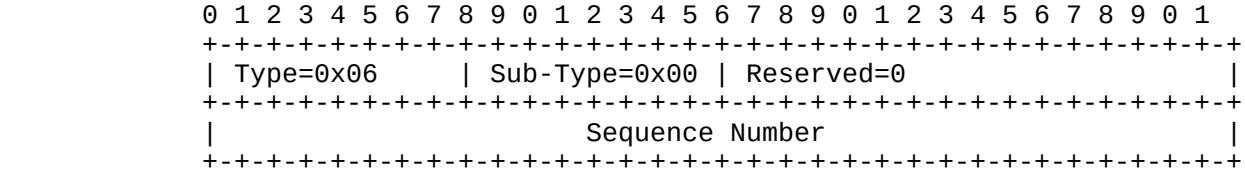
This document expands the definition of the Route Target extended community to allow the value of high order octet (Type field) to be 0x06 (in addition to the values specified in rfc4360). The value of low order octet (Sub-Type field) of 0x02 indicates that this extended community is of type "Route Target". The new value for Type field of 0x06 indicates that the structure of this RT is a six bytes value (e.g., a MAC address). A BGP speaker that implements RT-Constrain (RFC4684) MUST apply the RT-Constrain procedures to the ES-import RT as-well.

For procedures and usage of this attribute, please see section 9.1 "Redundancy Group Discovery".

8.7 MAC Mobility Extended Community

This extended community is a new transitive extended community with the Type field of 0x06 and the Sub-Type of 0x00. It may be advertised along with MAC Advertisement routes. The procedures for using this Extended Community are described in section 16 "MAC Mobility".

The MAC Mobility Extended Community is encoded as a 8-octet value as follows:



8.8 Default Gateway Extended Community

The Default Gateway community is an Extended Community of an Opaque Type (see 3.3 of rfc4360). It is a transitive community, which means that the first octet is 0x03. The value of the second octet (Sub-Type) is 0x030d (Default Gateway) as defined by IANA. The Value field of this community is reserved (set to 0 by the senders, ignored by the receivers).

9. Multi-homing Functions

This section discusses the functions, procedures and associated BGP routes used to support multi-homing in E-VPN. This covers both multi-homed device (MHD) as well as multi-homed network (MHN) scenarios.

9.1 Multi-homed Ethernet Segment Auto-Discovery

PEs connected to the same Ethernet segment can automatically discover each other with minimal to no configuration through the exchange of the Ethernet Segment route.

9.1.1 Constructing the Ethernet Segment Route

The Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value field comprises an IP address of the MES (typically, the loopback address) followed by 0’s.

The Ethernet Segment Identifier MUST be set to the ten octet ESI identifier described in section 6.

The BGP advertisement that advertises the Ethernet Segment route MUST also carry an ES-Import extended community attribute, as defined in

section 8.6.

The Ethernet Segment Route filtering MUST be done such that the Ethernet Segment Route is imported only by the PEs that are multi-homed to the same Ethernet Segment. To that end, each PE that is connected to a particular Ethernet segment constructs an import filtering rule to import a route that carries the ES-Import extended community, constructed from the ESI.

Note that the new ES-Import extended community is not the same as the Route Target Extended Community. The Ethernet Segment route carries this new ES-Import extended community. The PEs apply filtering on this new extended community. As a result the Ethernet Segment route is imported only by the PEs that are connected to the same Ethernet segment.

9.2 Fast Convergence

In E-VPN, MAC address reachability is learnt via the BGP control-plane over the MPLS network. As such, in the absence of any fast protection mechanism, the network convergence time is a function of the number of MAC Advertisement routes that must be withdrawn by the PE encountering a failure. For highly scaled environments, this scheme yields slow convergence.

To alleviate this, E-VPN defines a mechanism to efficiently and quickly signal, to remote PE nodes, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment. This is done by having each PE advertise an Ethernet A-D Route per Ethernet segment for each locally attached segment (refer to section 9.2.1 below for details on how this route is constructed). Upon a failure in connectivity to the attached segment, the PE withdraws the corresponding Ethernet A-D route. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all MAC addresses associated with the Ethernet segment in question. If no other PE had advertised an Ethernet A-D route for the same segment, then the PE that received the withdrawal simply invalidates the MAC entries for that segment. Otherwise, the PE updates the next-hop adjacencies to point to the backup PE(s).

9.2.1 Constructing the Ethernet A-D Route per Ethernet Segment

This section describes procedures to construct the Ethernet A-D route when a single such route is advertised by an PE for a given Ethernet Segment. This flavor of the Ethernet A-D route is used for fast convergence (as discussed above) as well as for advertising the ESI label used for split-horizon filtering (as discussed in section 9.3). Support of this route flavor is MANDATORY.

Route-Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by 0.

The Ethernet Segment Identifier MUST be a ten octet entity as described in section "Ethernet Segment". This document does not specify the use of the Ethernet A-D route when the Segment Identifier is set to 0.

The Ethernet Tag ID MUST be set to 0.

The MPLS label in the NLRI MUST be set to 0.

The "ESI Label Extended Community" MUST be included in the route. If all-Active multi-homing is desired, then the "Active-Standby" bit in the flags of the ESI Label Extended Community MUST be set to 0 and the MPLS label in that extended community MUST be set to a valid MPLS label value. The MPLS label in this Extended Community is referred to as an "ESI label". This label MUST be a downstream assigned MPLS label if the advertising PE is using ingress replication for receiving multicast, broadcast or unknown unicast traffic from other PEs. If the advertising PE is using P2MP MPLS LSPs for sending multicast, broadcast or unknown unicast traffic, then this label MUST be an upstream assigned MPLS label. The usage of this label is described in section 9.3.

If the Ethernet Segment is connected to more than one PE and Single-Active multi-homing is desired, then the "Active-Standby" bit in the flags of the ESI Label Extended Community MUST be set to 1 and ESI label MUST be set to zero.

9.2.1.1. Ethernet A-D Route Targets

The Ethernet A-D route MUST carry one or more Route Target (RT) attributes. These RTs MUST be the set of RTs associated with all the EVIs to which the Ethernet Segment, corresponding to the Ethernet A-D route, belongs.

9.3 Split Horizon

Consider a CE that is multi-homed to two or more PEs on an Ethernet segment ES1 operating in All-Active mode. If the CE sends a broadcast, unknown unicast, or multicast (BUM) packet to one of the non-DF (Designated Forwarder) PEs, say PE1, then PE1 will forward that packet to all or subset of the other PEs in the EVI including the DF PE for that Ethernet segment. In this case the DF PE that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This filtering is referred to as "split horizon" filtering in

this document.

In order to achieve this split horizon function, every BUM packet originating from a non-DF PE is encapsulated with an MPLS label that identifies the Ethernet segment of origin (i.e. the segment from which the frame entered the E-VPN network). This label is referred to as the ESI label, and MUST be distributed by all PEs when operating in All-Active multi-homing mode using the "Ethernet A-D route per Ethernet Segment" as per the procedures in section 9.2.1 above. This route is imported by the PEs connected to the Ethernet Segment and also by the PEs that have at least one EVI in common with the Ethernet Segment in the route. As described in section 9.1.1, the route MUST carry an ESI Label Extended Community with a valid ESI label. The disposition DF PE rely on the value of the ESI label to determine whether or not a BUM frame is allowed to egress a specific Ethernet segment. It should be noted that if the BUM frame is originated from the DF PE operating in All-Active multi-homing mode, then the DF PE MAY not encapsulate the frame with the ESI label. Furthermore, if the multi-homed PEs operate in active/standby mode, then the packet MUST not be encapsulated with the ESI label and the label value MUST be set to zero in ESI Label Extended Community per section 9.2.1 above.

9.3.1 ESI Label Assignment

The following subsections describe the assignment procedures for the ESI label, which differ depending on the type of tunnels being used to deliver multi-destination packets in the E-VPN network.

9.3.1.1 Ingress Replication

All PEs operating in an All-Active multi-homing mode that rely on ingress replication for the reception of BUM traffic, distribute to other PEs, that belong to the Ethernet segment, a downstream assigned "ESI label" in the Ethernet A-D route per ESI. This label MUST be programmed in the platform label space by the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for.

Consider PE1 and PE2 that are multi-homed to CE1 on ES1 and operating in All-Active multi-homing mode. Further consider that PE1 is using P2P or MP2P LSPs to send packets to PE2. Consider that PE1 is the non-DF for VLAN1 and PE2 is the DF for VLAN1, and PE1 receives a BUM packet from CE1 on VLAN1 on ES1. In this scenario, PE2 distributes an Inclusive Multicast Ethernet Tag route for VLAN1 in the associated EVI. So, when PE1 sends a BUM packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that PE2 has

distributed for ES1. It MUST then push on the MPLS label distributed by PE2 in the Inclusive Multicast Ethernet Tag route for VLAN1. The resulting packet is further encapsulated in the P2P or MP2P LSP label stack required to transmit the packet to PE2. When PE2 receives this packet, it determines the set of ESIs to replicate the packet to from the top MPLS label, after any P2P or MP2P LSP labels have been removed. If the next label is the ESI label assigned by PE2 for ES1, then PE2 MUST NOT forward the packet onto ES1. If the next label is an ESI label which has not been assigned by PE2, then PE2 MUST drop the packet. It should be noted that in this scenario, if PE2 receives a BUM traffic for VLAN1 from CE1, then it doesn't need to encapsulate the packet with an ESI label when sending it to the PE1 since PE1 can use its DF logic to filter the BUM packets and thus doesn't need to use split-horizon filtering for ES1.

9.3.1.2. P2MP MPLS LSPs

The non-DF PEs operating in an All-Active multi-homing mode that is using P2MP LSPs for sending BUM traffic, distribute to other PEs, that belong to the Ethernet segment or have an E-VPN in common with the Ethernet Segment, an upstream assigned "ESI label" in the Ethernet A-D route. This label is upstream assigned by the PE that advertises the route. This label MUST be programmed by the other PEs, that are connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must result in NOT forwarding packets received with this label onto the Ethernet segment that the label was distributed for. This label MUST also be programmed by the other PEs, that import the route but are not connected to the ESI advertised in the route, in the context label space for the advertising PE. Further the forwarding entry for this label must be a POP with no other associated action.

Consider PE1 and PE2 that are multi-homed to CE1 on ES1 and operating in All-Active multi-homing mode. Also consider PE3 that is in the same EVI as one of the EVIs to which ES1 belongs. Further, assume that PE1 which is the non-DF, using P2MP MPLS LSPs to send BUM packets. When PE1 sends a BUM packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI that the packet was received on. The resulting packet is further encapsulated in the P2MP MPLS label stack necessary to transmit the packet to the other PEs. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for E-VPN. When PE2 receives this packet, it de-capsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1, then PE2 MUST NOT forward the packet onto ES1. When PE3 receives this packet, it de-capsulates the

top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by PE1 to ES1 and PE3 is not connected to ES1, then PE3 MUST pop the label and flood the packet over all local ESIs in the EVI. It should be noted that when PE2 sends a BUM frame over a P2MP LSP, it does not need to encapsulate the frame with an ESI label because it is the DF for that VLAN.

9.3.1.3. MP2MP LSPs

The procedures for ESI Label assignment and usage for MP2MP LSPs will be described in a future version.

9.4 Aliasing and Backup-Path

In the case where a CE is multi-homed to multiple PE nodes, using a LAG with All-Active redundancy, it is possible that only a single PE learns a set of the MAC addresses associated with traffic transmitted by the CE. This leads to a situation where remote PE nodes receive MAC advertisement routes, for these addresses, from a single PE even though multiple PEs are connected to the multi-homed segment. As a result, the remote PEs are not able to effectively load-balance traffic among the PE nodes connected to the multi-homed Ethernet segment. This could be the case, for e.g. when the PEs perform data-path learning on the access, and the load-balancing function on the CE hashes traffic from a given source MAC address to a single PE. Another scenario where this occurs is when the PEs rely on control plane learning on the access (e.g. using ARP), since ARP traffic will be hashed to a single link in the LAG.

To alleviate this issue, E-VPN introduces the concept of 'Aliasing'. Aliasing refers to the ability of a PE to signal that it has reachability to a given locally attached Ethernet segment, even when it has learnt no MAC addresses from that segment. The Ethernet A-D route per EVI is used to that end. Remote PEs which receive MAC advertisement routes with non-reserved ESI SHOULD consider the advertised MAC address as reachable via all PEs which have advertised reachability to the relevant Segment using: (1) Ethernet A-D routes per EVI with the same ESI (and Ethernet Tag if applicable) AND (2)Ethernet A-D routes per ESI with the same ESI and with the Active/Standby bit set to 0 in the ESI Label Extended Community.

This flavor of Ethernet A-D route per EVI, associated with aliasing, can arrive at target PEs asynchronously relative to the flavor of Ethernet A-D route associated with split-horizon and mass-withdraw (i.e. per ESI). Therefore, if the Ethernet A-D route per EVI arrives ahead of the Ethernet A-D route per ESI, then the former must NOT be used for traffic forwarding till the latter arrives. This will take

care of corner cases and race conditions where the Ethernet A-D route associated with mass-withdraw is withdrawn but a PE still receives the route associated with aliasing.

Backup-Path is a closely related function, albeit it applies to the case where the redundancy mode is Active/Standby. In this case, the PE advertises that it has reachability to a given locally attached Ethernet Segment using the Ethernet A-D route as well. Remote PEs which receive the MAC advertisement routes, with non-reserved ESI, MUST consider the MAC address as reachable via the advertising PE. Furthermore, the remote PEs SHOULD install a Backup-Path, for said MAC, to the PE which had advertised reachability to the relevant Segment using (1) an Ethernet A-D routes per EVI with the same ESI (and Ethernet Tag if applicable) AND (2)Ethernet A-D routes per ESI with the same ESI and with the Active/Standby bit set to 1 in the ESI Label Extended Community.

9.4.1 Constructing the Ethernet A-D Route per EVI

This section describes procedures to construct the Ethernet A-D route when one or more such routes are advertised by an PE for a given EVI. This flavor of the Ethernet A-D route is used for aliasing, and support of this route flavor is OPTIONAL.

Route-Distinguisher (RD) MUST be set to the RD of the EVI that is advertising the NLRI. An RD MUST be assigned for a given EVI on an PE. This RD MUST be unique across all EVIs on an PE. It is RECOMMENDED to use the Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE. This number may be generated by the PE. Or in the Unique VLAN E-VPN case, the low order 12 bits may be the 12 bit VLAN ID, with the remaining high order 4 bits set to 0.

The Ethernet Segment Identifier MUST be a ten octet entity as described in section "Ethernet Segment Identifier". This document does not specify the use of the Ethernet A-D route when the Segment Identifier is set to 0.

The Ethernet Tag ID is the identifier of an Ethernet Tag on the Ethernet segment. This value may be a 12 bit VLAN ID, in which case the low order 12 bits are set to the VLAN ID and the high order 20 bits are set to 0. Or it may be another Ethernet Tag used by the E-VPN. It MAY be set to the default Ethernet Tag on the Ethernet segment or to the value 0.

Note that the above allows the Ethernet A-D route to be advertised with one of the following granularities:

+ One Ethernet A-D route for a given <ESI, Ethernet Tag ID> tuple per EVI. This is applicable when the PE uses MPLS-based disposition.

+ One Ethernet A-D route per <ESI, EVI> (where the Ethernet Tag ID is set to 0). This is applicable when the PE uses MAC-based disposition, or when the PE uses MPLS-based disposition when no VLAN translation is required.

The usage of the MPLS label is described in the section on "Load Balancing of Unicast Packets".

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

9.4.1.1 Ethernet A-D Route Targets

The Ethernet A-D route MUST carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically.

If an PE uses Route Target Constrain [RT-CONSTRAIN], the PE SHOULD advertise all such RTs using Route Target Constrains. The use of RT Constrains allows each Ethernet A-D route to reach only those PEs that are configured to import at least one RT from the set of RTs carried in the Ethernet A-D route.

9.4.1.1.1 Auto-Derivation from the Ethernet Tag ID

The following is the procedure for deriving the RT attribute automatically from the Ethernet Tag ID associated with the advertisement:

+ The Global Administrator field of the RT MUST be set to the Autonomous System (AS) number that the PE belongs to.

+ The Local Administrator field of the RT contains a 4 octets long number that encodes the Ethernet Tag-ID. If the Ethernet Tag-ID is a two octet VLAN ID then it MUST be encoded in the lower two octets of the Local Administrator field and the higher two octets MUST be set to zero.

For the "Unique VLAN E-VPN" this results in auto-deriving the RT from the Ethernet Tag, e.g., VLAN ID for that E-VPN.

9.5 Designated Forwarder Election

Consider a CE that is a host or a router that is multi-homed directly to more than one PE in an E-VPN on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

- Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the CE.

- Flooding unknown unicast traffic (i.e. traffic for which an PE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the CE, if the environment requires flooding of unknown unicast traffic.

Note that this behavior, which allows selecting a DF at the granularity of <ESI, EVI> for multicast, broadcast and unknown unicast traffic, is the default behavior in this specification. Optional mechanisms, which will be specified in the future, will allow selecting a DF at the granularity of <ESI, EVI, S, G>.

Note that a CE always sends packets belonging to a specific flow using a single link towards an PE. For instance, if the CE is a host then, as mentioned earlier, the host treats the multiple links that it uses to reach the PEs as a Link Aggregation Group (LAG). The CE employs a local hashing function to map traffic flows onto links in the LAG.

If a bridged network is multi-homed to more than one PE in an E-VPN via switches, then the support of All-Active points of attachments, as described in this specification, requires the bridge network to be connected to two or more PEs using a LAG. In this case the reasons for doing DF election are the same as those described above when a CE is a host or a router.

If a bridged network does not connect to the PEs using LAG, then only one of the links between the switched bridged network and the PEs must be the active link for a given Ethernet Tag. In this case, the Ethernet A-D route per Ethernet segment MUST be advertised with the "Active-Standby" flag set to one. Procedures for supporting All-Active points of attachments, when a bridge network connects to the PEs using LAG, are for further study.

The default procedure for DF election at the granularity of <ESI, EVI> is referred to as "service carving". With service carving, it is possible to elect multiple DFs per Ethernet Segment (one per EVI) in order to perform load-balancing of multi-destination traffic destined to a given Segment. The load-balancing procedures carve up the EVI space among the PE nodes evenly, in such a way that every PE is the

DF for a disjoint set of EVIs. The procedure for service carving is as follows:

1. When a PE discovers the ESI of the attached Ethernet Segment, it advertises an Ethernet Segment route with the associated ES-Import extended community attribute.

2. The PE then starts a timer (default value = 3 seconds) to allow the reception of Ethernet Segment routes from other PE nodes connected to the same Ethernet Segment. This timer value MUST be same across all PEs connected to the same Ethernet Segment.

3. When the timer expires, each PE builds an ordered list of the IP addresses of all the PE nodes connected to the Ethernet Segment (including itself), in increasing numeric value. Every PE is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the numerically lowest IP address. The ordinals are used to determine which PE node will be the DF for a given EVI on the Ethernet Segment using the following rule: Assuming a redundancy group of N PE nodes, the PE with ordinal i is the DF for an EVI with an associated Ethernet Tag value V when (V mod N) = i. In the case where multiple Ethernet Tags are associated with a single EVI, then the numerically lowest Ethernet Tag value in that EVI MUST be used in the modulo function.

4. The PE that is elected as a DF for a given EVI will unblock traffic for the Ethernet Tags associated with that EVI. Note that the DF PE unblocks multi-destination traffic in the egress direction towards the Segment. All non-DF PEs continue to drop multi-destination traffic (for the associated EVIs) in the egress direction towards the Segment.

In the case of link or port failure, the affected PE withdraws its Ethernet Segment route. This will re-trigger the service carving procedures on all the PEs in the RG. For PE node failure, or upon PE commissioning or decommissioning, the PEs re-trigger the service carving. In case of a Single-Active multi-homing, when a service moves from one PE in the RG to another PE as a result of re-carving, the PE, which ends up being the elected DF for the service, must trigger a MAC address flush notification towards the associated Ethernet Segment. This can be done, for e.g. using IEEE 802.1ak MVRP 'new' declaration.

10. Determining Reachability to Unicast MAC Addresses

PEs forward packets that they receive based on the destination MAC address. This implies that PEs must be able to learn how to reach a given destination unicast MAC address.

There are two components to MAC address learning, "local learning" and "remote learning":

10.1. Local Learning

A particular PE must be able to learn the MAC addresses from the CEs that are connected to it. This is referred to as local learning.

The PEs in a particular E-VPN MUST support local data plane learning using standard IEEE Ethernet learning procedures. An PE must be capable of learning MAC addresses in the data plane when it receives packets such as the following from the CE network:

- DHCP requests

- ARP request for its own MAC.

- ARP request for a peer.

Alternatively PEs MAY learn the MAC addresses of the CEs in the control plane or via management plane integration between the PEs and the CEs.

There are applications where a MAC address that is reachable via a given PE on a locally attached Segment (e.g. with ESI X) may move such that it becomes reachable via another PE on another Segment (e.g. with ESI Y). This is referred to as a "MAC Mobility". Procedures to support this are described in section "MAC Mobility".

10.2. Remote learning

A particular PE must be able to determine how to send traffic to MAC addresses that belong to or are behind CEs connected to other PEs i.e. to remote CEs or hosts behind remote CEs. We call such MAC addresses as "remote" MAC addresses.

This document requires an PE to learn remote MAC addresses in the control plane. In order to achieve this, each PE advertises the MAC addresses it learns from its locally attached CEs in the control plane, to all the other PEs in the EVI, using MP-BGP and specifically the MAC Advertisement route.

10.2.1. Constructing the BGP E-VPN MAC Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC Advertisement route type in the E-VPN NLRI.

The RD MUST be the RD of the EVI that is advertising the NLRI. The

procedures for setting the RD for a given EVI are described in section 9.4.1.

The Ethernet Segment Identifier is set to the ten octet ESI described in section "Ethernet Segment".

The Ethernet Tag ID may be zero or may represent a valid Ethernet Tag ID. This field may be non-zero when there are multiple bridge domains in the EVI (e.g., the PE needs to perform qualified learning for the VLANs in that EVI).

When the the Ethernet Tag ID in the NLRI is set to a non-zero value, for a particular bridge domain, then this Ethernet Tag may either be the Ethernet tag value associated with the CE, e.g., VLAN ID, or it may be the Ethernet Tag Identifier, e.g., VLAN ID assigned by the E-VPN provider and mapped to the CE's Ethernet tag. The latter would be the case if the CE Ethernet tags, e.g., VLAN ID, for a particular bridge domain are different on different CEs.

The MAC address length field is typically set to 48. However this specification enables specifying the MAC address as a prefix; in which case, the MAC address length field is set to the length of the prefix. This provides the ability to aggregate MAC addresses if the deployment environment supports that. The encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1D-ORIG] [802.1D-REV]. If the MAC address is advertised as a prefix then the trailing bits of the prefix MUST be set to 0 to ensure that the entire prefix is encoded as 6 octets.

The IP Address Length field value is set to the number of octets in the IP Address field.

The IP Address field is optional. By default, the IP Address Length field is set to 0 and the IP address field is omitted from the route. When a valid IP address needs to be advertised (e.g., for ARP suppression purposes or for inter-subnet switching), it is then encoded in this route.

The MPLS label field carries one or more labels (that corresponds to the stack of labels [MPLS-ENCAPS]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [MPLS-ENCAPS]). The MPLS label stack MUST be the downstream assigned E-VPN MPLS label stack that is used by the PE to forward MPLS-encapsulated Ethernet frames received from remote PEs, where the destination MAC address in the Ethernet frame is the MAC address advertised in the above NLRI. The forwarding procedures are specified in section "Forwarding Unicast Packets" and "Load Balancing of Unicast Packets".

An PE may advertise the same single E-VPN label for all MAC addresses in a given EVI. This label assignment methodology is referred to as a per EVI label assignment. Alternatively, an PE may advertise a unique E-VPN label per <ESI, Ethernet Tag> combination. This label assignment methodology is referred to as a per <ESI, Ethernet Tag> label assignment. As a third option, an PE may advertise a unique E-VPN label per MAC address. All of these methodologies have their tradeoffs.

Per EVI label assignment requires the least number of E-VPN labels, but requires a MAC lookup in addition to an MPLS lookup on an egress PE for forwarding. On the other hand, a unique label per <ESI, Ethernet Tag> or a unique label per MAC allows an egress PE to forward a packet that it receives from another PE, to the connected CE, after looking up only the MPLS labels without having to perform a MAC lookup. This includes the capability to perform appropriate VLAN ID translation on egress to the CE.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the advertising PE.

The BGP advertisement for the MAC advertisement route MUST also carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs), or may be derived automatically from the Ethernet Tag ID, in the Unique VLAN case, as described in section "Ethernet A-D Route per E-VPN".

It is to be noted that this document does not require PEs to create forwarding state for remote MACs when they are learnt in the control plane. When this forwarding state is actually created is a local implementation matter.

10.2.2 Route Resolution

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to the reserved ESI value of 0 or MAX-ESI, then the receiving PE MUST install forwarding state for the associated MAC Address based on the MAC Advertisement route alone.

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, and the receiving PE is locally attached to the same ESI, then the PE does not alter its forwarding state based on the received route. This ensures that local routes are preferred to remote routes.

If the Ethernet Segment Identifier field in a received MAC Advertisement route is set to a non-reserved ESI, then the receiving PE MUST install forwarding state for a given MAC address only when

both the MAC Advertisement route AND the associated Ethernet A-D route per ESI have been received.

To illustrate this with an example, consider two PEs (PE1 and PE2) connected to a multi-homed Ethernet Segment ES1. All-Active redundancy mode is assumed. A given MAC address M1 is learnt by PE1 but not PE2. On PE3, the following states may arise:

T1- When the MAC Advertisement Route from PE1 and the Ethernet A-D routes per ESI from PE1 and PE2 are received, PE3 can forward traffic destined to M1 to both PE1 and PE2.

T2- If after T1, PE1 withdraws its Ethernet A-D route per ESI, then PE3 forwards traffic destined to M1 to PE2 only.

T3- If after T1, PE2 withdraws its Ethernet A-D route per ESI, then PE3 forwards traffic destined to M1 to PE1 only.

T4- If after T1, PE1 withdraws its MAC Advertisement route, then PE3 treats traffic to M1 as unknown unicast. Note, here, that had PE2 also advertised a MAC route for M1 before PE1 withdraws its MAC route, then PE3 would have continued forwarding traffic destined to M1 to PE2.

11. ARP and ND

The IP address field in the MAC advertisement route may optionally carry one of the IP addresses associated with the MAC address. This provides an option which can be used to minimize the flooding of ARP or Neighbor Discovery (ND) messages over the MPLS network and to remote CEs. This option also minimizes ARP (or ND) message processing on end-stations/hosts connected to the E-VPN network. An PE may learn the IP address associated with a MAC address in the control or management plane between the CE and the PE. Or, it may learn this binding by snooping certain messages to or from a CE. When an PE learns the IP address associated with a MAC address, of a locally connected CE, it may advertise this address to other PEs by including it in the MAC Advertisement route. The IP Address may be an IPv4 address encoded using four octets, or an IPv6 address encoded using sixteen octets. The IP Address length field MUST be set to 32 for an IPv4 address or to 128 for an IPv6 address.

If there are multiple IP addresses associated with a MAC address, then multiple MAC advertisement routes MUST be generated, one for each IP address. For instance, this may be the case when there are both an IPv4 and an IPv6 address associated with the MAC address. When the IP address is dissociated with the MAC address, then the MAC advertisement route with that particular IP address MUST be

withdrawn.

When an PE receives an ARP request for an IP address from a CE, and if the PE has the MAC address binding for that IP address, the PE SHOULD perform ARP proxy and respond to the ARP request.

11.1 Default Gateway

A PE MAY choose to terminate ARP messages instead of performing ARP proxy for them. Such scenarios arises when the PE needs to perform inter-subnet forwarding where each subnet is represented by a different bridge domain/EVI. In such scenarios the inter-subnet forwarding is performed at layer 3 and the PE that performs such function is called the default gateway.

Each PE that acts as a default gateway for a given E-VPN advertises in the E-VPN control plane its default gateway IP and MAC address using the MAC advertisement route, and indicates that such route is associated with the default gateway. This is accomplished by requiring the route to carry the Default Gateway extended community defined in [Section 8.8 Default Gateway Extended Community].

Each PE that receives this route and imports it as per procedures specified in this document follows the procedures in this section when replying to ARP Requests that it receives if such Requests are for the IP address in the received E-VPN route.

Each PE that acts as a default gateway for a given E-VPN that receives this route and imports it as per procedures specified in this document MUST create MAC forwarding state that enables it to apply IP forwarding to the packets destined to the MAC address carried in the route.

12. Handling of Multi-Destination Traffic

Procedures are required for a given PE to send broadcast or multicast traffic, received from a CE encapsulated in a given Ethernet Tag in an EVI, to all the other PEs that span that Ethernet Tag in the EVI. In certain scenarios, described in section "Processing of Unknown Unicast Packets", a given PE may also need to flood unknown unicast traffic to other PEs.

The PEs in a particular E-VPN may use ingress replication, P2MP LSPs or MP2MP LSPs to send unknown unicast, broadcast or multicast traffic to other PEs.

Each PE MUST advertise an "Inclusive Multicast Ethernet Tag Route" to

enable the above. The following subsection provides the procedures to construct the Inclusive Multicast Ethernet Tag route. Subsequent subsections describe in further detail its usage.

12.1. Construction of the Inclusive Multicast Ethernet Tag Route

The RD MUST be the RD of the EVI that is advertising the NLRI. The procedures for setting the RD for a given E-VPN are described in section 9.4.1.

The Ethernet Tag ID is the identifier of the Ethernet Tag. It MAY be set to 0 or to a valid Ethernet Tag value.

The Originating Router's IP address MUST be set to an IP address of the PE. This address SHOULD be common for all the EVIs on the PE (e.,g., this address may be PE's loopback address).

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the same IP address as the one carried in the Originating Router's IP Address field.

The BGP advertisement for the Inclusive Multicast Ethernet Tag route MUST also carry one or more Route Target (RT) attributes. The assignment of RTs described in the section on "Constructing the BGP E-VPN MAC Address Advertisement" MUST be followed.

12.2. P-Tunnel Identification

In order to identify the P-Tunnel used for sending broadcast, unknown unicast or multicast traffic, the Inclusive Multicast Ethernet Tag route MUST carry a "PMSI Tunnel Attribute" as specified in [BGP MVPN].

Depending on the technology used for the P-tunnel for the E-VPN on the PE, the PMSI Tunnel attribute of the Inclusive Multicast Ethernet Tag route is constructed as follows.

+ If the PE that originates the advertisement uses a P-Multicast tree for the P-tunnel for E-VPN, the PMSI Tunnel attribute MUST contain the identity of the tree (note that the PE could create the identity of the tree prior to the actual instantiation of the tree).

+ An PE that uses a P-Multicast tree for the P-tunnel MAY aggregate two or more Ethernet Tags in the same or different EVIs present on the PE onto the same tree. In this case, in addition to carrying the identity of the tree, the PMSI Tunnel attribute MUST carry an MPLS upstream assigned label which

the PE has bound uniquely to the Ethernet Tag for the EVI associated with this update (as determined by its RTs).

If the PE has already advertised Inclusive Multicast Ethernet Tag routes for two or more Ethernet Tags that it now desires to aggregate, then the PE MUST re-advertise those routes. The re-advertised routes MUST be the same as the original ones, except for the PMSI Tunnel attribute and the label carried in that attribute.

+ If the PE that originates the advertisement uses ingress replication for the P-tunnel for E-VPN, the route MUST include the PMSI Tunnel attribute with the Tunnel Type set to Ingress Replication and Tunnel Identifier set to a routable address of the PE. The PMSI Tunnel attribute MUST carry a downstream assigned MPLS label. This label is used to demultiplex the broadcast, multicast or unknown unicast E-VPN traffic received over a MP2P tunnel by the PE.

+ The Leaf Information Required flag of the PMSI Tunnel attribute MUST be set to zero, and MUST be ignored on receipt.

13. Processing of Unknown Unicast Packets

The procedures in this document do not require the PEs to flood unknown unicast traffic to other PEs. If PEs learn CE MAC addresses via a control plane protocol, the PEs can then distribute MAC addresses via BGP, and all unicast MAC addresses will be learnt prior to traffic to those destinations.

However, if a destination MAC address of a received packet is not known by the PE, the PE may have to flood the packet. Flooding must take into account "split horizon forwarding" as follows: The principles behind the following procedures are borrowed from the split horizon forwarding rules in VPLS solutions [RFC4761] and [RFC4762]. When an PE capable of flooding (say PEx) receives a broadcast or multicast Ethernet frame, or one with an unknown destination MAC address, it must flood the frame. If the frame arrived from an attached CE, PEx must send a copy of the frame to every other attached CE participating in the EVI, on a different ESI than the one it received the frame on, as long as the PE is the DF for the egress ESI. In addition, the PE must flood the frame to all other PEs participating in the EVI. If, on the other hand, the frame arrived from another PE (say PEy), PEx must send a copy of the packet only to attached CEs as long as it is the DF for the egress ESI. PEx MUST NOT send the frame to other PEs, since PEy would have already done so. Split horizon forwarding rules apply to broadcast and multicast packets, as well as packets to an unknown MAC address.

Whether or not to flood packets to unknown destination MAC addresses should be an administrative choice, depending on how learning happens between CEs and PEs.

The PEs in a particular E-VPN may use ingress replication using RSVP-TE P2P LSPs or LDP MP2P LSPs for sending broadcast, multicast and unknown unicast traffic to other PEs. Or they may use RSVP-TE P2MP or LDP P2MP or LDP MP2MP LSPs for sending such traffic to other PEs.

13.1. Ingress Replication

If ingress replication is in use, the P-Tunnel attribute, carried in the Inclusive Multicast Ethernet Tag routes for the EVI, specifies the downstream label that the other PEs can use to send unknown unicast, multicast or broadcast traffic for the EVI to this particular PE.

The PE that receives a packet with this particular MPLS label MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the PE MUST treat the packet as an unknown unicast packet.

13.2. P2MP MPLS LSPs

The procedures for using P2MP LSPs are very similar to VPLS procedures [VPLS-MCAST]. The P-Tunnel attribute used by an PE for sending unknown unicast, broadcast or multicast traffic for a particular EVI is advertised in the Inclusive Ethernet Tag Multicast route as described in section "Handling of Multi-Destination Traffic".

The P-Tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree in [VPLS-MCAST]. Note that multiple Ethernet Tags, which may be in different EVIs, may use the same P2MP LSP, using upstream labels [VPLS-MCAST]. This is the equivalent of an Aggregate Inclusive tree in [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

The PE that receives a packet on the P2MP LSP specified in the PMSI Tunnel Attribute MUST treat the packet as a broadcast, multicast or unknown unicast packet. Further if the MAC address is a unicast MAC address, the PE MUST treat the packet as an unknown unicast packet.

14. Forwarding Unicast Packets

14.1. Forwarding packets received from a CE

When an PE receives a packet from a CE, on a given Ethernet Tag, it

must first look up the source MAC address of the packet. In certain environments the source MAC address MAY be used to authenticate the CE and determine that traffic from the host can be allowed into the network. Source MAC lookup MAY also be used for local MAC address learning.

If the PE decides to forward the packet, the destination MAC address of the packet must be looked up. If the PE has received MAC address advertisements for this destination MAC address from one or more other PEs or learned it from locally connected CEs, it is considered as a known MAC address. Otherwise, the MAC address is considered as an unknown MAC address.

For known MAC addresses the PE forwards this packet to one of the remote PEs or to a locally attached CE. When forwarding to a remote PE, the packet is encapsulated in the E-VPN MPLS label advertised by the remote PE, for that MAC address, and in the MPLS LSP label stack to reach the remote PE.

If the MAC address is unknown and if the administrative policy on the PE requires flooding of unknown unicast traffic then:

- The PE MUST flood the packet to other PEs. The PE MUST first encapsulate the packet in the ESI MPLS label as described in section 9.3.
  If ingress replication is used, the packet MUST be replicated one or more times to each remote PE with the outermost label being an MPLS label determined as follows: This is the MPLS label advertised by the remote PE in a PMSI Tunnel Attribute in the Inclusive Multicast Ethernet Tag route for an <EVI, Ethernet Tag> combination. The Ethernet Tag in the route must be the same as the Ethernet Tag associated with the interface on which the ingress PE receives the packet. If P2MP LSPs are being used the packet MUST be sent on the P2MP LSP that the PE is the root of for the Ethernet Tag in the EVI. If the same P2MP LSP is used for all Ethernet Tags, then all the PEs in the EVI MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP is used for a given Ethernet Tag in the EVI, then only the PEs in the Ethernet Tag MUST be the leaves of the P2MP LSP. The packet MUST be encapsulated in the P2MP LSP label stack.

If the MAC address is unknown then, if the administrative policy on the PE does not allow flooding of unknown unicast traffic:

- The PE MUST drop the packet.

14.2. Forwarding packets received from a remote PE

14.2.1. Unknown Unicast Forwarding

When an PE receives an MPLS packet from a remote PE then, after processing the MPLS label stack, if the top MPLS label ends up being a P2MP LSP label associated with an EVI or the downstream label advertised in the P-Tunnel attribute, and after performing the split horizon procedures described in section "Split Horizon":

- If the PE is the designated forwarder of unknown unicast, broadcast or multicast traffic, on a particular set of ESIs for the Ethernet Tag, the default behavior is for the PE to flood the packet on these ESIs. In other words, the default behavior is for the PE to assume that the destination MAC address is unknown unicast, broadcast or multicast and it is not required to perform a destination MAC address lookup. As an option, the PE may perform a destination MAC lookup to flood the packet to only a subset of the CE interfaces in the Ethernet Tag. For instance the PE may decide to not flood an unknown unicast packet on certain Ethernet segments even if it is the DF on the Ethernet segment, based on administrative policy.

- If the PE is not the designated forwarder on any of the ESIs for the Ethernet Tag, the default behavior is for it to drop the packet.

14.2.2. Known Unicast Forwarding

If the top MPLS label ends up being an E-VPN label that was advertised in the unicast MAC advertisements, then the PE either forwards the packet based on CE next-hop forwarding information associated with the label or does a destination MAC address lookup to forward the packet to a CE.

15. Load Balancing of Unicast Frames

This section specifies the load balancing procedures for sending known unicast frames to a multi-homed CE.

15.1. Load balancing of traffic from an PE to remote CEs

Whenever a remote PE imports a MAC advertisement for a given <ESI, Ethernet Tag> in an EVI, it MUST examine all imported Ethernet A-D routes for that ESI in order to determine the load-balancing characteristics of the Ethernet segment.

15.1.1 Single-Active Redundancy Mode

For a given ESI, if the remote PE has imported an Ethernet A-D route per Ethernet Segment from at least one PE, where the "Active-Standby"

flag in the ESI Label Extended Community is set, then the remote PE MUST deduce that the Ethernet segment is operating in Single-Active redundancy mode. As such, the MAC address will be reachable only via the PE announcing the associated MAC Advertisement route - this is referred to as the primary PE. The set of other PE nodes advertising Ethernet A-D routes per Ethernet Segment for the same ESI serve as backup paths, in case the active PE encounters a failure. These are referred to as the backup PEs. It should be noted that the primary PE for a given <ESI, EVI> is the DF for that <ESI, EVI>.

If the primary PE encounters a failure, it MAY withdraw its Ethernet A-D route for the affected segment prior to withdrawing the entire set of MAC Advertisement routes.

In the case where only a single other backup PE in the network had advertised an Ethernet A-D route for the same ESI, the remote PE can then use the Ethernet A-D route withdrawal as a trigger to update its forwarding entries, for the associated MAC addresses, to point towards the backup PE. As the backup PE starts learning the MAC addresses over its attached Ethernet segment, it will start sending MAC Advertisement routes while the failed PE withdraws its own. This mechanism minimizes the flooding of traffic during fail-over events.

In the case where multiple other backup PE in the network had advertised an Ethernet A-D route for the same ESI, the remote PE MUST then use the Ethernet A-D route withdrawal as a trigger to start flooding traffic destined to the associated MAC addresses (as long as flooding of unknown unicast is administratively allowed). It is not possible to select a single backup path in this case.

15.1.2 All-Active Redundancy Mode

If for the given ESI, none of the Ethernet A-D routes per Ethernet Segment imported by the remote PE have the "Active-Standby" flag set in the ESI Label Extended Community, then the remote PE MUST treat the Ethernet segment as operating in All-Active redundancy mode. The remote PE would then treat the MAC address as reachable via all of the PE nodes from which it has received both an Ethernet A-D route per Ethernet Segment as well as an Ethernet A-D route per EVI for the ESI in question. The remote PE MUST use the MAC advertisement and eligible Ethernet A-D routes to construct the set of next-hops that it can use to send the packet to the destination MAC. Each next-hop comprises an MPLS label stack that is to be used by the egress PE to forward the packet. This label stack is determined as follows:

-If the next-hop is constructed as a result of a MAC route then this label stack MUST be used. However, if the MAC route doesn’t exist, then the next-hop and MPLS label stack is constructed as a result of

the Ethernet A-D routes. Note that the following description applies to determining the label stack for a particular next-hop to reach a given PE, from which the remote PE has received and imported Ethernet A-D routes that have the matching ESI and Ethernet Tag as the one present in the MAC advertisement. The Ethernet A-D routes mentioned in the following description refer to the ones imported from this given PE.

-If an Ethernet A-D route per Ethernet Segment for that ESI exists, together with an Ethernet A-D route per EVI, then the label from that latter route must be used.

The following example explains the above.

Consider a CE (CE1) that is dual-homed to two PEs (PE1 and PE2) on a LAG interface (ES1), and is sending packets with MAC address MAC1 on VLAN1. A remote PE, say PE3, is able to learn that MAC1 is reachable via PE1 and PE2. Both PE1 and PE2 may advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If this is not the case, and if MAC1 is advertised only by PE1, PE3 still considers MAC1 as reachable via both PE1 and PE2 as both PE1 and PE2 advertise a Ethernet A-D route per ESI for ES1 as well as an Ethernet A-D route per EVI for <ES1, VLAN1>.

The MPLS label stack to send the packets to PE1 is the MPLS LSP stack to get to PE1 and the E-VPN label advertised by PE1 for CE1’s MAC.

The MPLS label stack to send packets to PE2 is the MPLS LSP stack to get to PE2 and the MPLS label in the Ethernet A-D route advertised by PE2 for <ES1, VLAN1>, if PE2 has not advertised MAC1 in BGP.

We will refer to these label stacks as MPLS next-hops.

The remote PE (PE3) can now load balance the traffic it receives from its CEs, destined for CE1, between PE1 and PE2. PE3 may use N-Tuple flow information to hash traffic into one of the MPLS next-hops for load balancing of IP traffic. Alternatively PE3 may rely on the source MAC addresses for load balancing.

Note that once PE3 decides to send a particular packet to PE1 or PE2 it can pick one out of multiple possible paths to reach the particular remote PE using regular MPLS procedures. For instance, if the tunneling technology is based on RSVP-TE LSPs, and PE3 decides to send a particular packet to PE1, then PE3 can choose from multiple RSVP-TE LSPs that have PE1 as their destination.

When PE1 or PE2 receive the packet destined for CE1 from PE3, if the packet is a unicast MAC packet it is forwarded to CE1. If it is a

multicast or broadcast MAC packet then only one of PE1 or PE2 must forward the packet to the CE. Which of PE1 or PE2 forward this packet to the CE is determined based on which of the two is the DF.

If the connectivity between the multi-homed CE and one of the PEs that it is attached to fails, the PE MUST withdraw the Ethernet Tag A-D routes, that had been previously advertised, for the Ethernet Segment to the CE. When the MAC entry on the PE ages out, the PE MUST withdraw the MAC address from BGP. Note that to aid convergence, the Ethernet Tag A-D routes MAY be withdrawn before the MAC routes. This enables the remote PEs to remove the MPLS next-hop to this particular PE from the set of MPLS next-hops that can be used to forward traffic to the CE. For further details and procedures on withdrawal of E-VPN route types in the event of PE to CE failures please section "PE to CE Network Failures".

15.2. Load balancing of traffic between an PE and a local CE

A CE may be configured with more than one interface connected to different PEs or the same PE for load balancing, using a technology such as LAG. The PE(s) and the CE can load balance traffic onto these interfaces using one of the following mechanisms.

15.2.1. Data plane learning

Consider that the PEs perform data plane learning for local MAC addresses learned from local CEs. This enables the PE(s) to learn a particular MAC address and associate it with one or more interfaces, if the technology between the PE and the CE supports multi-pathing. The PEs can now load balance traffic destined to that MAC address on the multiple interfaces.

Whether the CE can load balance traffic that it generates on the multiple interfaces is dependent on the CE implementation.

15.2.2. Control plane learning

The CE can be a host that advertises the same MAC address using a control protocol on both interfaces. This enables the PE(s) to learn the host’s MAC address and associate it with one or more interfaces. The PEs can now load balance traffic destined to the host on the multiple interfaces. The host can also load balance the traffic it generates onto these interfaces and the PE that receives the traffic employs E-VPN forwarding procedures to forward the traffic.

16. MAC Mobility

It is possible for a given host or end-station (as defined by its MAC

address) to move from one Ethernet segment to another; this is referred to as 'MAC Mobility' or 'MAC move' and it is different from the multi-homing situation in which a given MAC address is reachable via multiple PEs for the same Ethernet segment. In a MAC move, there would be two sets of MAC Advertisement routes, one set with the new Ethernet segment and one set with the previous Ethernet segment, and the MAC address would appear to be reachable via each of these segments.

In order to allow all of the PEs in the E-VPN to correctly determine the current location of the MAC address, all advertisements of it being reachable via the previous Ethernet segment MUST be withdrawn by the PEs, for the previous Ethernet segment, that had advertised it.

If local learning is performed using the data plane, these PEs will not be able to detect that the MAC address has moved to another Ethernet segment and the receipt of MAC Advertisement routes, with the MAC Mobility extended community attribute, from other PEs serves as the trigger for these PEs to withdraw their advertisements. If local learning is performed using the control or management planes, these interactions serve as the trigger for these PEs to withdraw their advertisements.

In a situation where there are multiple moves of a given MAC, possibly between the same two Ethernet segments, there may be multiple withdrawals and re-advertisements. In order to ensure that all PEs in the E-VPN receive all of these correctly through the intervening BGP infrastructure, it is necessary to introduce a sequence number into the MAC Mobility extended community attribute.

Since the sequence number is an unsigned 32 bit integer, all sequence number comparisons must be performed modulo 2**32. This unsigned arithmetic preserves the relationship of sequence numbers as they cycle from 2**32 - 1 to 0.

Every MAC mobility event for a given MAC address will contain a sequence number that is set using the following rules:

- A PE advertising a MAC address for the first time advertises it with no MAC Mobility extended community attribute.

- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with a different Ethernet segment identifier advertises the MAC address in a MAC Advertisement route tagged with a MAC Mobility extended community attribute with a sequence number one greater than the sequence number in the MAC mobility attribute of the received MAC Advertisement

route. In the case of the first mobility event for a given MAC address, where the received MAC Advertisement route does not carry a MAC Mobility attribute, the value of the sequence number in the received route is assumed to be 0 for purpose of this processing.

- A PE detecting a locally attached MAC address for which it had previously received a MAC Advertisement route with the same Ethernet segment identifier advertises it with:
  i. no MAC Mobility extended community attribute, if the received route did not carry said attribute.

  ii. a MAC Mobility extended community attribute with the sequence number equal to the highest of the sequence number(s) in the received MAC Advertisement route(s), if the received route(s) is (are) tagged with a MAC Mobility extended community attribute.

A PE receiving a MAC Advertisement route for a MAC address with a different Ethernet segment identifier and a higher sequence number than that which it had previously advertised, withdraws its MAC Advertisement route. If two (or more) PEs advertise the same MAC address with same sequence number but different Ethernet segment identifiers, a PE that receives these routes selects the route advertised by the PE with lowest IP address as the best route.

16.1. MAC Duplication Issue

A situation may arise where the same MAC address is learned by different PEs in the same VLAN because of two (or more hosts) being mis-configured with the same (duplicate) MAC address. In such situation, the traffic originating from these hosts would trigger continuous MAC moves among the PEs attached to these hosts. It is important to recognize such situation and avoid incrementing the sequence number (in the MAC Mobility attribute) to infinity. In order to remedy such situation, a PE that detects a MAC mobility event by way of local learning starts an M-second timer (default value of M = 5) and if it detects N MAC moves before the timer expires (default value for N = 3), it concludes that a duplicate MAC situation has occurred. The PE MUST alert the operator and stop sending and processing any BGP MAC Advertisement routes for that MAC address till a corrective action is taken by the operator. The values of M and N MUST be configurable to allow for flexibility in operator control.

17. Multicast

The PEs in a particular E-VPN may use ingress replication or P2MP LSPs to send multicast traffic to other PEs.

17.1. Ingress Replication

The PEs may use ingress replication for flooding unknown unicast, multicast or broadcast traffic as described in section "Handling of Multi-Destination Traffic". A given unknown unicast or broadcast packet must be sent to all the remote PEs. However a given multicast packet for a multicast flow may be sent to only a subset of the PEs. Specifically a given multicast flow may be sent to only those PEs that have receivers that are interested in the multicast flow. Determining which of the PEs have receivers for a given multicast flow is done using explicit tracking described below.

17.2. P2MP LSPs

An PE may use an "Inclusive" tree for sending an unknown unicast, broadcast or multicast packet or a "Selective" tree. This terminology is borrowed from [VPLS-MCAST].

A variety of transport technologies may be used in the SP network. For inclusive P-Multicast trees, these transport technologies include point-to-multipoint LSPs created by RSVP-TE or mLDP. For selective P-Multicast trees, only unicast PE-PE tunnels (using MPLS or IP/GRE encapsulation) and P2MP LSPs are supported, and the supported P2MP LSP signaling protocols are RSVP-TE, and mLDP.

17.3. MP2MP LSPs

The root of the MP2MP LDP LSP advertises the Inclusive Multicast Tag route with the PMSI Tunnel attribute set to the MP2MP Tunnel identifier. This advertisement is then sent to all PEs in the E-VPN. Upon receiving the Inclusive Multicast Tag routes with a PMSI Tunnel attribute that contains the MP2MP Tunnel identifier, the receiving PEs initiate the setup of the MP2MP tunnel towards the root using the procedures in [MLDP].

17.3.1. Inclusive Trees

An Inclusive Tree allows the use of a single multicast distribution tree, referred to as an Inclusive P-Multicast tree, in the SP network to carry all the multicast traffic from a specified set of EVIs on a given PE. A particular P-Multicast tree can be set up to carry the traffic originated by sites belonging to a single E-VPN, or to carry the traffic originated by sites belonging to different E-VPNs. The ability to carry the traffic of more than one E-VPN on the same tree is termed ’Aggregation’. The tree needs to include every PE that is a member of any of the E-VPNs that are using the tree. This implies that an PE may receive multicast traffic for a multicast stream even if it doesn’t have any receivers that are interested in receiving

traffic for that stream.

An Inclusive P-Multicast tree as defined in this document is a P2MP tree. A P2MP tree is used to carry traffic only for E-VPN CEs that are connected to the PE that is the root of the tree.

The procedures for signaling an Inclusive Tree are the same as those in [VPLS-MCAST] with the VPLS-AD route replaced with the Inclusive Multicast Ethernet Tag route. The P-Tunnel attribute [VPLS-MCAST] for an Inclusive tree is advertised in the Inclusive Multicast route as described in section "Handling of Multi-Destination Traffic". Note that an PE can "aggregate" multiple inclusive trees for different EVIs on the same P2MP LSP using upstream labels. The procedures for aggregation are the same as those described in [VPLS-MCAST], with VPLS A-D routes replaced by E-VPN Inclusive Multicast routes.

17.3.2. Selective Trees

A Selective P-Multicast tree is used by an PE to send IP multicast traffic for one or more specific IP multicast streams, originated by CEs connected to the PE, that belong to the same or different E-VPNs, to a subset of the PEs that belong to those E-VPNs. Each of the PEs in the subset should be on the path to a receiver of one or more multicast streams that are mapped onto the tree. The ability to use the same tree for multicast streams that belong to different E-VPNs is termed an PE the ability to create separate SP multicast trees for specific multicast streams, e.g. high bandwidth multicast streams. This allows traffic for these multicast streams to reach only those PE routers that have receivers in these streams. This avoids flooding other PE routers in the E-VPN.

An SP can use both Inclusive P-Multicast trees and Selective P-Multicast trees or either of them for a given E-VPN on an PE, based on local configuration.

The granularity of a selective tree is <RD, PE, S, G> where S is an IP multicast source address and G is an IP multicast group address or G is a multicast MAC address. Wildcard sources and wildcard groups are supported. Selective trees require explicit tracking as described below.

A E-VPN PE advertises a selective tree using a E-VPN selective A-D route. The procedures are the same as those in [VPLS-MCAST] with S-PMSI A-D routes in [VPLS-MCAST] replaced by E-VPN Selective A-D routes. The information elements of the E-VPN selective A-D route are similar to those of the VPLS S-PMSI A-D route with the following differences. A E-VPN Selective A-D route includes an optional Ethernet Tag field. Also an E-VPN selective A-D route may encode a

MAC address in the Group field. The encoding details of the E-VPN selective A-D route will be described in the next revision.

Selective trees can also be aggregated on the same P2MP LSP using aggregation as described in [VPLS-MCAST].

17.4. Explicit Tracking

[VPLS-MCAST] describes procedures for explicit tracking that rely on Leaf A-D routes. The same procedures are used for explicit tracking in this specification with VPLS Leaf A-D routes replaced with E-VPN Leaf A-D routes. These procedures allow a root PE to request multicast membership information for a given (S, G), from leaf PEs. Leaf PEs rely on IGMP snooping or PIM snooping between the PE and the CE to determine the multicast membership information. Note that the procedures in [VPLS-MCAST] do not describe how explicit tracking is performed if the CEs are enabled with join suppression. The procedures for this case will be described in a future version.

18. Convergence

This section describes failure recovery from different types of network failures.

18.1. Transit Link and Node Failures between PEs

The use of existing MPLS Fast-Reroute mechanisms can provide failure recovery in the order of 50ms, in the event of transit link and node failures in the infrastructure that connects the PEs.

18.2. PE Failures

Consider a host host1 that is dual homed to PE1 and PE2. If PE1 fails, a remote PE, PE3, can discover this based on the failure of the BGP session. This failure detection can be in the sub-second range if BFD is used to detect BGP session failure. PE3 can update its forwarding state to start sending all traffic for host1 to only PE2. It is to be noted that this failure recovery is potentially faster than what would be possible if data plane learning were to be used. As in that case PE3 would have to rely on re-learning of MAC addresses via PE2.

18.2.1. Local Repair

It is possible to perform local repair in the case of PE failures. Details will be specified in the future.

18.3. PE to CE Network Failures

When an Ethernet segment connected to an PE fails or when a Ethernet Tag is decommissioned on an Ethernet segment, then the PE MUST withdraw the Ethernet A-D route(s) announced for the <ESI, Ethernet Tags> that are impacted by the failure or decommissioning. In addition, the PE MUST also withdraw the MAC advertisement routes that are impacted by the failure or decommissioning.

The Ethernet A-D routes should be used by an implementation to optimize the withdrawal of MAC advertisement routes. When an PE receives a withdrawal of a particular Ethernet A-D route from an PE it SHOULD consider all the MAC advertisement routes, that are learned from the same <ESI, Ethernet Tag> as in the Ethernet A-D route, from the advertising PE, as having been withdrawn. This optimizes the network convergence times in the event of PE to CE failures.

19. LACP State Synchronization

This section requires review and discussion amongst the authors and will be revised in the next version.

To support CE multi-homing with multi-chassis Ethernet bundles, the PEs connected to a given CE should synchronize [802.1AX] LACP state amongst each other. This ensures that the PEs can present a single LACP bundle to the CE. This is required for initial system bring-up and upon any configuration change.

This includes at least the following LACP specific configuration parameters:

- System Identifier (MAC Address): uniquely identifies a LACP speaker.
- System Priority: determines which LACP speaker’s port priorities are used in the Selection logic.
- Aggregator Identifier: uniquely identifies a bundle within a LACP speaker.
- Aggregator MAC Address: identifies the MAC address of the bundle.
- Aggregator Key: used to determine which ports can join an Aggregator.
- Port Number: uniquely identifies an interface within a LACP speaker.
- Port Key: determines the set of ports that can be bundled.
- Port Priority: determines a port’s precedence level to join a bundle in case the number of eligible ports exceeds the maximum number of links allowed in a bundle.

Furthermore, the PEs should also synchronize operational (run-time) data, in order for the LACP Selection logic state-machines to

execute. This operational data includes the following LACP operational parameters, on a per port basis:

- Partner System Identifier: this is the CE System MAC address.
- Partner System Priority: the CE LACP System Priority
- Partner Port Number: CE's AC port number.
- Partner Port Priority: CE's AC Port Priority.
- Partner Key: CE's key for this AC.
- Partner State: CE's LACP State for the AC.
- Actor State: PE's LACP State for the AC.
- Port State: PE's AC port status.

The above state needs to be communicated between PEs forming a multi-chassis bundle during LACP initial bringup, upon any configuration change and upon the occurrence of a failure.

It should be noted that the above configuration and operational state is localized in scope and is only relevant to PEs which connect to the same multi-homed CE over a given Ethernet bundle.

Furthermore, the communication of state changes, upon failures, must occur with minimal latency, in order to minimize the switchover time and consequent service disruption. The protocol details for synchronizing the LACP state will be described in the following version.

20. Acknowledgements

We would like to thank Yakov Rekhter, Pedro Marques, Kaushik Ghosh, Nischal Sheth, Robert Raszuk, Amit Shukla and Nadeem Mohammed for discussions that helped shape this document. We would also like to thank Han Nguyen for his comments and support of this work. We would also like to thank Steve Kensil and Reshad Rahman for their reviews. Last but not least, many thanks to Jakob Heitz for his help to improve several sections of this draft.

21. Security Considerations

22. IANA Considerations

23. References

23.1 Normative References

[RFC4364] "BGP/MPLS IP VPNs", Rosen, Rekhter, et. al., February 2006

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC4271] Y. Rekhter et. al., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006

[RFC4760] T. Bates et. al., "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007

23.2 Informative References

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

[EVPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-ietf-l2vpn-evpn-req-01.txt

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-11.txt

[RT-CONSTRAIN] P. Marques et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006

[BGP-VPLS-MH] "BGP based Multi-homing in Virtual Private LAN Service", K. Kompella et. al., draft-ietf-l2vpn-vpls-multihoming-04.txt

24. Author’s Address

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Rahul Aggarwal
Email: raggarwa_1@yahoo.com

Wim Henderickx
Alcatel-Lucent
e-mail: wim.henderickx@alcatel-lucent.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

James Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748
USA
Email: uttaro@att.com

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com

Ravi Shekhar
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US
Email: rshekhar@juniper.net

Florin Balus
Alcatel-Lucent
e-mail: Florin.Balus@alcatel-lucent.com

Keyur Patel
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: keyupate@cisco.com

Sami Boutros
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: sboutros@cisco.com

Samer Salam
Cisco
Email: ssalam@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net