Network Working Group
INTERNET-DRAFT
Category: Standards Track
Expires: August 25, 2012

R. Aggarwal
Arktan

A. Sajassi
Cisco

J. Uttaro
AT&T

W. Henderickx
Alcatel-Lucent

A. Isaac
Bloomberg

N. Bitar
Verizon

F. Balus
Alcatel-Lucent

R. Shekhar
J. Drake
Juniper Networks

S. Boutros
K. Patel
Cisco

February 24, 2012

BGP MPLS Based Ethernet VPN
draft-ietf-l2vpn-evpn-00

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

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
Juniper Networks

Samer Salam
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 [E-VPN-REQ]. Please refer to [E-VPN-REQ] for the detailed requirements and motivation.

This document proposes an MPLS based technology, referred to as MPLS-based E-VPN (E-VPN). 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

CE: Customer Edge device e.g., host or router or switch
MES: MPLS Edge Switch
EVI: E-VPN Instance
ESI: Ethernet segment identifier
LACP: Link Aggregation Control Protocol
MP2MP: Multipoint to Multipoint
P2MP: Point to Multipoint
P2P: Point to Point

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, or MPLS Edge Switches (MES), that form the edge of the MPLS infrastructure. A CE

may be a host, a router or a switch. The MPLS Edge Switches provide layer 2 virtual bridge connectivity between the CEs. There may be multiple E-VPNs in the provider's network. An E-VPN routing and forwarding instance on an MES is referred to as an E-VPN Instance (EVI).

The MESes maybe connected by an MPLS LSP infrastructure which provides the benefits of MPLS LSP technology such as fast-reroute, resiliency, etc. The MESes may also be connected by an IP infrastructure in which case IP/GRE tunneling is used between the MESes. 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/GRE as the tunneling technology.

In an E-VPN, MAC learning between MESes 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 (very similar to IP VPNs (RFC 4364)), providing greater scale, and the ability to preserve the "virtualization" or isolation of groups of interacting agents (hosts, servers, Virtual Machines) from each other. In E-VPNs MESes advertise the MAC addresses learned from the CEs that are connected to them, along with an MPLS label, to other MESes 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 MESes. This is in addition to load balancing across the MPLS core via multiple LSPs betwen the same pair of MESes. 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 MESes 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 a MES is populated with all the MAC destinations known to the control plane or whether the MES 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 MES.

The policy attributes of an E-VPN are very similar to those of an IP VPN. An E-VPN instance requires a Route-Distinguisher (RD) and an E-VPN requires one or more Route-Targets (RTs). A CE attaches to an E-VPN instance (EVI) on an MES, 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 E-VPN use the same VLAN, and no other E-VPN uses this VLAN. This document refers to this case as a "Unique Single VLAN E-VPN" and describes simplified procedures to optimize for it.

6. Ethernet Segment Identifier

If a CE is multi-homed to two or more MESes, 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. A single-homed CE is considered to be attached to an Ethernet segment with ESI 0. Otherwise, an Ethernet segment MUST have a unique non-zero ESI. The ESI can be assigned using various mechanisms:

1. The ESI may be configured. For instance when E-VPNs are used to provide a VPLS service the ESI is fairly analogous to the Multi-homing site ID in [BGP-VPLS-MH].

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

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

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

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

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

4. In the case of indirectly connected hosts via a bridged LAN between the CEs and the MESes, the ESI is determined based on the Layer 2 bridge protocol as follows: If STP is used in the bridged LAN then the value of the ESI is derived by listening to BPDUs on

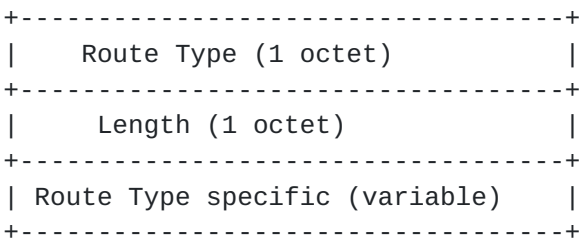
the Ethernet segment. To achieve this the MES is not required to run STP. However the MES must learn the Switch ID, MSTP ID and Root Bridge ID by listening to STP BPDUs. The ESI is constructed as follows:

{Switch ID (6 bits), MSTP ID (6 bits), Root Bridge ID (48 bits)}

7. 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 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
+ 5 - Selective Multicast Auto-Discovery (A-D) Route
+ 6 - Leaf Auto-Discovery (A-D) Route

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

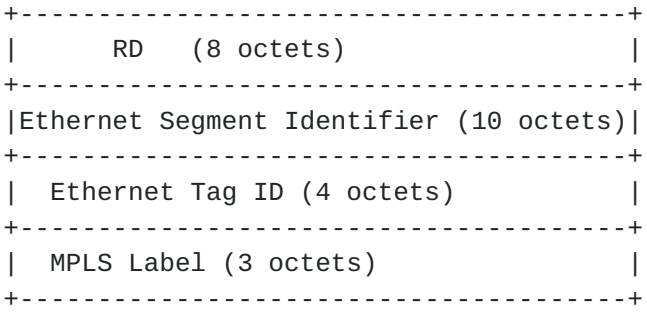
The E-VPN NLRI is carried in BGP [RFC4271] using BGP Multiprotocol Extensions [RFC4760] with an AFI of TBD and an SAFI of E-VPN (To be assigned by IANA). The NLRI field in the MP_REACH_NLRI/MP_UNREACH_NLRI attribute contains the E-VPN NLRI (encoded as specified above).

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 TBD and an SAFI of E-VPN.

7.1. Ethernet Auto-Discovery Route

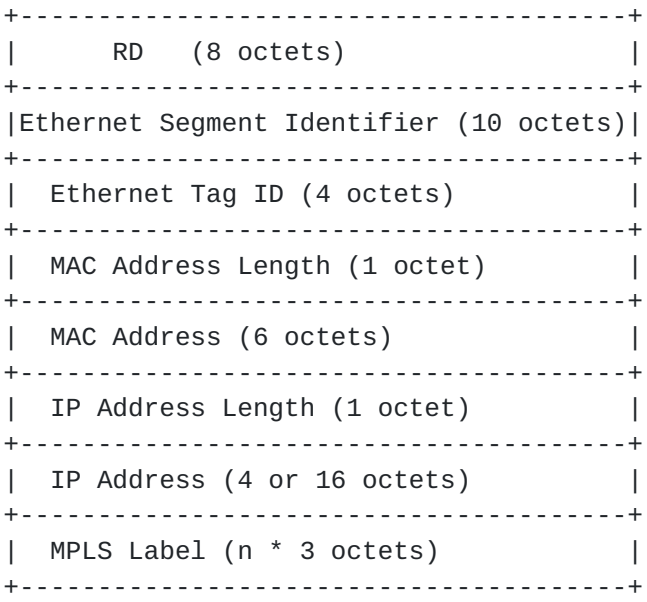
A Ethernet A-D route type specific E-VPN NLRI consists of the following:



For procedures and usage of this route please see the sections on "Auto-Discovery of Ethernet Tags on Ethernet Segments", "Designated Forwarder Election" and "Load Balancing".

7.2. MAC Advertisement Route

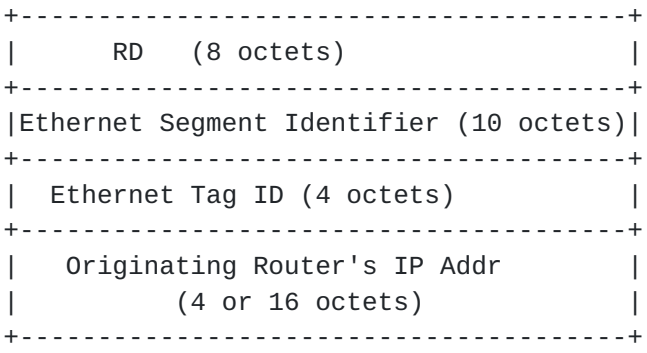
A MAC advertisement route type specific E-VPN NLRI consists of the following:



For procedures and usage of this route please see the sections on "Determining Reachability to Unicast MAC Addresses" and "Load Balancing of Unicast Packets".

7.3. Inclusive Multicast Ethernet Tag Route

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

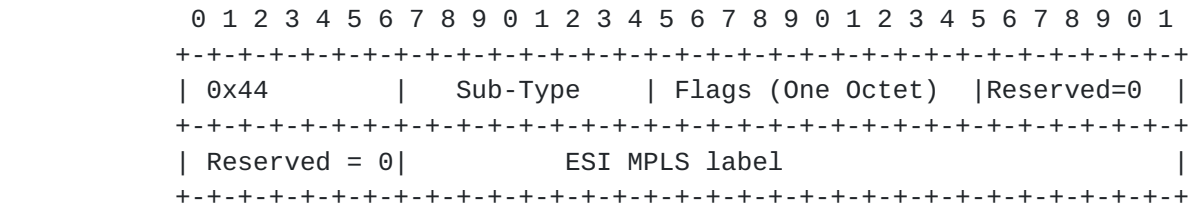


For procedures and usage of this route please see the sections on "Handling of Multi-Destination Traffic", "Unknown Unicast Traffic" and "Multicast".

8. ESI MPLS Label Extended Community

This extended community is a new transitive extended community. It may be advertised along with Ethernet Auto-Discovery routes. When used it carries properties associated with the ESI. Specifically it enables split horizon procedures for multi-homed sites. The procedures for using this Extended Community are described in following sections.

Each ESI MPLS 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. The other bits must be set to 0.

9. Auto-Discovery

E-VPN requires the following types of auto-discovery procedures:

+ E-VPN Auto-Discovery, which allows an MES to discover the other MESes in the E-VPN. Each MES advertises one or more "Inclusive Multicast Tag Routes". The procedures for advertising these

routes are described in the section on "Handling of Multi-Destination Traffic".

+ Auto-Discovery of Ethernet Tags on Ethernet Segments, in a particular E-VPN. The procedures are described in section "Auto-Discovery of Ethernet Tags on Ethernet Segments".

+ Ethernet Segment Auto-Discovery used for auto-discovery of MESes that are multi-homed to the same Ethernet segment. The procedures are described in section "Auto-Discovery of Ethernet Tags on Ethernet Segments".

10. Auto-Discovery of Ethernet Tags on Ethernet Segments

If a CE is multi-homed to two or more MESes on a particular Ethernet segment, each MES MUST advertise, to other MESes in the E-VPN, the information about the Ethernet Tags that are associated with that Ethernet segment. An Ethernet Tag identifies a particular broadcast domain. An example of an Ethernet Tag is a VLAN ID. The MES MAY advertise each Ethernet Tag associated with the Ethernet Segment, or it may advertise a wildcard to cover all the Ethernet Tags enabled on the segment. If a CE is single-homed, then the MES that it is attached to MAY advertise the information about Ethernet Tags (e.g.,VLANs) on the Ethernet segment connected to the CE.

The information about an Ethernet Tag on a particular Ethernet segment is advertised using an "Ethernet Auto-Discovery route (Ethernet A-D route)". This route is advertised using the E-VPN NLRI.

The Ethernet Tag Auto-discovery information SHOULD be used to enable active-active load-balancing among MESes as described in section "Load Balancing of Unicast Packets". In the case of a multi-homed CE this route MUST also carry the "ESI Label Extended Community" to enable split horizon as described in section "Split Horizon". Also, the route can be used for Designated Forwarder (DF) election as described in section "Designated Forwarder Election". Further,it MAY be used to optimize the withdrawal of MAC addresses upon failure as described in section "Convergence".

This section describes procedures for advertising one or more Ethernet A-D routes per Ethernet tag per E-VPN. We will call this as "Ethernet A-D route per E-VPN". This section also describes procedures to advertise and withdraw a single Ethernet A-D route per Ethernet Segment. We will call this as "Ethernet A-D route per Segment".

10.1. Constructing the Ethernet A-D Route

The format of the Ethernet A-D NLRI is specified in section "BGP E-VPN NLRI".

10.1.1. Ethernet A-D Route per E-VPN

This section describes procedures to construct the Ethernet A-D route when one or more such routes are advertised by an MES for a given E-VPN instance.

Route-Distinguisher (RD) MUST be set to the RD of the E-VPN instance that is advertising the NLRI. A RD MUST be assigned for a given E-VPN instance on an MES. This RD MUST be unique across all E-VPN instances on an MES. It is RECOMMENDED to use the Type 1 RD [RFC4364]. The value field comprises an IP address of the MES (typically, the loopback address) followed by a number unique to the MES. This number may be generated by the MES. Or in the Unique Single 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.

Ethernet Segment Identifier MAY be set to 0. When it is not zero the Ethernet Segment Identifier MUST be a ten octet entity as described in section "Ethernet Segment Identifier".

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 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 E-VPN

+ One Ethernet A-D route for a given <Ethernet Tag ID> in a given E-VPN, for all associated Ethernet segments, where the ESI is set to 0.

+ One Ethernet A-D route for the E-VPN where both ESI and Ethernet Tag ID are set to 0.

E-VPN supports both the non-qualified and qualified learning models. When non-qualified learning is used, the Ethernet Tag Identifier specified in this section and in other places in this document MUST be set to the default Ethernet Tag, e.g., VLAN ID. When qualified learning is used, and the Ethernet Tags between MESes and CEs in the

E-VPN are consistently assigned for a given broadcast domain, the Ethernet Tag Identifier MUST be set to the Ethernet Tag, e.g., VLAN ID for the concerned broadcast domain between the advertising MES and the CE. When qualified learning is used, and the Ethernet Tags, e.g., VLAN IDs between MESes and CEs in the E-VPN are not consistently assigned for a given broadcast domain, the Ethernet Tag Identifier, e.g., VLAN ID MUST be set to a common E-VPN provider assigned tag that maps locally on the advertising MES to an Ethernet broadcast domain identifier such as a VLAN ID. The usage of the MPLS label is described in 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 MES.

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

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

10.1.2. 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 MES for a given Ethernet Segment.

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. The reason for such encoding is that the RD cannot be that of a given E-VPN since the ESI can span across one or more E-VPNs.

Ethernet Segment Identifier MUST be a non-zero ten octet entity as described in section "Ethernet Segment Identifier".

The Ethernet Tag ID MUST be set to 0.

If the Ethernet Segment is connected to more than one MES then the "ESI MPLS Label Extended Community" MUST be included in the route. If the Ethernet Segment is connected to more than one MES and active-active multi-homing is desired then the MPLS label in the ESI MPLS Label 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 MES is using ingress replication for receiving multicast, broadcast or unknown unicast traffic, from other MESes. If the advertising MES 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 "Split Horizon".

If the Ethernet Segment is connected to more than one MES and active-standby multi-homing is desired then the "Active-Standby" bit in the flags of the ESI MPLS Label Extended Community MUST be set to 1.

If the per Ethernet Segment Ethernet A-D route is used in conjunction with the per {ESI, VLAN} Ethernet A-D route, for reasons described below, then the MPLS label in the NLRI MUST be set to 0.

10.1.2.1. 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 E-VPN instances to which the Ethernet Segment, corresponding to the Ethernet A-D route, belongs.

10.2. Motivations for Ethernet A-D Route per Ethernet Segment

This section describes various scenarios in which the Ethernet A-D route should be advertised per Ethernet Segment.

10.2.1. Multi-Homing

The per Ethernet Segment Ethernet A-D route MUST be advertised when the Ethernet Segment is multi-homed. This allows Multi-Homed Ethernet Segment Auto-Discovery. It allows the set of MESes connected to the same customer site i.e., CE, to discover each other automatically with minimal to no configuration. It also allows other MESes that have at least one E-VPN in common with the multi-homed Ethernet Segment to discover the properties of the multi-homed Ethernet Segment.

For active-active multi-homing this route is required for split horizon procedures as described in section "Split Horizon" and MUST carry the ESI MPLS Label Extended Community with a valid ESI MPLS label. For active-standby multi-homing this route is required to indicate that active-standby multi-homing and not active-active multi-homing is desired.

This route will be enhanced to carry LAG specific information such as LACP parameters, which will be encoded as new BGP attributes or communities, in the future. Note that this information will be propagated to all MESes that have one or more sites in the VLANs connected to the Ethernet Segment. All the MESes other than the ones that are connected to the MESes will discard this information.

10.2.2. Optimizing Control Plane Convergence

Ethernet A-D route per Ethernet Segment should be advertised when it is desired to optimize the control plane convergence of the withdrawal of the Ethernet A-D routes. If this is done then when an Ethernet segment fails, the single Ethernet A-D route corresponding to the segment can be withdrawn first. This allows all MESes that receive this withdrawal to invalidate the MAC routes learned from the Ethernet segment.

Note that the Ethernet A-D route per Ethernet Segment, when used to optimize control plane convergence, MAY be advertised in addition to the Ethernet Tag A-D routes per E-VPN or MAY be advertised on its own.

10.2.3. Reducing Number of Ethernet A-D Routes

In certain scenarios advertising Ethernet A-D routes per Ethernet segment, instead of per E-VPN, may reduce the number of Ethernet A-D routes in the network. In these scenarios Ethernet A-D routes may be advertised per Ethernet segment instead of per E-VPN.

11. Determining Reachability to Unicast MAC Addresses

MESes forward packets that they receive based on the destination MAC address. This implies that MESes 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":

11.1. Local Learning

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

The MESes in a particular E-VPN MUST support local data plane learning using standard IEEE Ethernet learning procedures. An MES 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

- gratuitous ARP request for its own MAC.

- ARP request for a peer.

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

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

11.2. Remote learning

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

This document requires an MES to learn remote MAC addresses in the control plane. In order to achieve this each MES advertises the MAC addresses it learns from its locally attached CEs in the control plane, to all the other MESes in the E-VPN, using MP-BGP and the MAC address advertisement route.

11.2.1. 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 E-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given E-VPN are described in section "Ethernet A-D Route per E-VPN".

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

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 E-VPN instance (e.g., the MES needs to perform qualified learning for the VLANs in that EVPN instance).

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 IEEE 802 documents [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 MPLS Label Length field value is set to the number of octets in the MPLS Label field. 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 MES to forward MPLS encapsulated Ethernet packets received from remote MESes, where the destination MAC address in the Ethernet packet 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 MES may advertise the same single E-VPN label for all MAC

addresses in a given E-VPN instance. This label assignment methodology is referred to as a per EVI label assigment. Alternatively an MES 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. Or an MES 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 MES for forwarding. On the other hand a unique label per <ESI, Ethernet Tag> or a unique label per MAC allows an egress MES to forward a packet that it receives from another MES, to the connected CE, after looking up only the MPLS labels and not having to do a MAC lookup.

As well as to insert the appropriate VLAN ID on egress to the CE A MES may also advertise more than one label for a given MAC address. For instance an MES may advertise two labels, one of which is for the ESI corresponding to the MAC address and the second is for the Ethernet Tag on the ESI that the MAC address is learnt on.

The IP Address field is optional. By default the IP Address length is set to 0 and the IP address is excluded. When a valid IP address is included it is encoded as specified in the section "Optimizing ARP".

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 MES.

The BGP advertisement that advertises 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 Single VLAN case as described in section "Ethernet A-D Route per E-VPN".

It is to be noted that this document does not require MESes 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.

12. Optimizing ARP

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 messages to MAC VPN CEs and to MESes. This option also minimizes ARP message processing on MAC VPN CEs. A MES may learn the IP address

associated with a MAC address in the control or management plane between the CE and the MES. Or it may learn this binding by snooping certain messages to or from a CE. When a MES learns the IP address associated with a MAC address, of a locally connected CE, it may advertise it to other MESes by including it in the MAC route advertisement. The IP Address may be an IPv4, 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 and 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 is both an IPv4 and an IPv6 address associated with the MAC address. When the IP address is dis-associated with the MAC address then the MAC advertisement route with that particular IP address MUST be withdrawn.

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

Further detailed procedures will be specified in a later version.

13. Designated Forwarder Election

Consider a CE that is a host or a router that is multi-homed directly to more than one MES 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 MESes, 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. Note that this behavior, which allows selecting a DF at the granularity of <ESI, Ethernet Tag> for multicast and broadcast 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, Ethernet Tag, S, G>.

- Flooding unknown unicast traffic (i.e. traffic for which an MES 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 a CE always sends packets belonging to a specific flow

using a single link towards an MES. For instance, if the CE is a host then, as mentioned earlier, the host treats the multiple links that it uses to reach the MESes as a Link Aggregation Group (LAG). The CE employs a local hashing function to map traffic flows onto links in the LAG.

If a bridge network is multi-homed to more than one MES in an E-VPN via switches, then the support of active-active points of attachments as described in this specification requires the bridge network to be connected to two or more MESes 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 bridge network does not connect to the MESes using LAG, then only one of the links between the switched bridged network and the MESes must be the active link. In this case the per Ethernet Segment Ethernet Tag routes MUST be advertised with the "Active-Standby" flag set to one. Procedures for supporting active-active points of attachments, when a bridge network does not connect to the MESes using LAG, are for further study.

The granularity of the DF election MUST be at least the Ethernet segment via which the CE is multi-homed to the MESes. If the DF election is done at the Ethernet segment granularity then a single MES MUST be elected as the DF on the Ethernet segment.

If there are one or more Ethernet Tags (e.g., VLANs) on the Ethernet segment then the granularity of the DF election SHOULD be the combination of the Ethernet segment and Ethernet Tag on that Ethernet segment. In this case a single MES MUST be elected as the DF for a particular Ethernet Tag on that Ethernet segment.

There are two specified mechanisms for performing DF election.

13.1. DF Election Performed by All MESes

The MESes perform a designated forwarder (DF) election, for an Ethernet segment, or <ESI, Ethernet Tag> combination using the Ethernet Tag A-D BGP route described in section "Auto-Discovery of Ethernet Tags on Ethernet Segments".

The DF election for a particular ESI or a particular <ESI, Ethernet Tag> combination proceeds as follows. First an MES constructs a candidate list of MESes. This comprises all the Ethernet A-D routes with that particular ESI or <ESI, Ethernet Tag> tuple that an MES imports in an E-VPN instance, including the Ethernet A-D route(s) generated by the MES itself, if any. The DF MES is chosen from this candidate list. Note that DF election is carried out by all the MESes

that import the DF route.

The default procedure for choosing the DF is the MES with the highest IP address, of all the MESes in the candidate list. This procedure MUST be implemented. It ensures that, except during routing transients each MES chooses the same DF MES for a given ESI and Ethernet Tag combination.

Other alternative procedures for performing DF election are possible and will be described in the future.

13.2. DF Election Performed Only on Multi-Homed MESes

As an MES discovers other MESs that are members of the same multi-homed segment, using per Ethernet Segment Ethernet A-D Routes, it starts building an ordered list based on the originating MES IP addresses. This list is used to select a DF and a backup DF (BDF) on a per group of Ethernet Tag basis. For example, the MES with the numerically highest IP address is considered the DF for a given group of VLANs for that Ethernet segment and the next MES in the list is considered the BDF. To that end, the range of Ethernet Tags associated with the CE must be partitioned into disjoint sets. The size of each set is a function of the total number of CE Ethernet Tags and the total number of MESs that the Ethernet segment is multi-homed to. The DF can employ any distribution function that achieves an even distribution of Ethernet Tags across the MESes that are multi-homed to the Ethernet segment. The DF takes over the Ethernet Tag set of any MES encountering either a node failure or a link/Ethernet segment failure causing that MES to be isolated from the multi-homed segment. In case of a failure that is affecting the DF, then the BDF takes over the DF VLAN set.

It should be noted that once all the MESs participating in an Ethernet segment have the same ordered list for that site, then Ethernet Tag groups can be assigned to each member of that list deterministically without any need to explicitly distribute Ethernet Tags among the member MESs of that list. In other words, the DF election for a group of Ethernet Tags is a local matter and can be done deterministically. As an example, consider, that the ordered list consists of m MESes: (MES1, MES2,., MESm), and there are n Ethernet Tags for that site (V0, V1, V2, ., Vn-1). Then MES1 and MES2 can be the DF and the BDF respectively for all the Ethernet Tags corresponding to (i mod m) for i:0 to n-1. MES2 and MES3 can be the DF and the BDF respectively for all the Ethernet Tags corresponding to (i mod m) + 1 and so on till the last MES in the order list is reached. As a result MESm and MES1 is the DF and the BDF respectively

for the all the VLANs corresponding to (i mod m) + m-1.

14. Handling of Multi-Destination Traffic

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

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

Each MES MUST advertise an "Inclusive Multicast Ethernet Tag Route" to enable the above. Next section provides procedures to construct the Inclusive Multicast Ethernet Tag route. Subsequent sections describe in further detail its usage.

14.1. Construction of the Inclusive Multicast Ethernet Tag Route

The RD MUST be the RD of the E-VPN instance that is advertising the NLRI. The procedures for setting the RD for a given E-VPN are described in section "Ethernet A-D Route per E-VPN".

The Ethernet Segment Identifier MAY be set to the ten octet ESI identifier described in section "Ethernet Segment Identifier". Or it MAY be set to 0. It MUST be set to 0 if the Ethernet Tag is set to 0.

The Ethernet Tag ID is the identifier of the Ethernet Tag. It MAY be set to 0 in which case an egress MES MUST perform a MAC lookup to forward the packet.

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 that advertises 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.

14.2. 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" 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 the 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).

+ A PE that uses a P-Multicast tree for the P-tunnel MAY aggregate two or more Ethernet Tags in the same or different E-VPNs 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 <ESI, Ethernet Tag> for E-VPN 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 the 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 unicast 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.

14.3. Ethernet Segment Identifier and Ethernet Tag

As described above the encoding rules allow setting the Ethernet

Segment Identifier and Ethernet Tag to either non-zero valid values or to 0. If the Ethernet Tag is set to a non-zero valid value, then an egress MES can forward the packet to the set of egress ESIs in the Ethernet Tag, in the E-VPN, by performing an MPLS lookup only. Further if the ESI is also set to non zero then the egress MES does not need to replicate the packet as it is destined for a given Ethernet segment. If both Ethernet Tag and ESI are set to 0 then an egress MES MUST perform a MAC lookup in the EVI determined by the MPLS label, after the MPLS lookup, to forward the packet.

If an MES advertises multiple Inclusive Ethernet Tag routes for a given E-VPN then the PMSI Tunnel Attributes for these routes MUST be distinct.

15. Processing of Unknown Unicast Packets

The procedures in this document do not require MESes to flood unknown unicast traffic to other MESes. If MESes learn CE MAC addresses via a control plane, the MESes 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 MES, the MES 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 [RFC 4761, RFC 4762]. When an MES capable of flooding (say MESx) receives a broadcast Ethernet frame, or one with an unknown destination MAC address, it must flood the frame. If the frame arrived from an attached CE, MESx must send a copy of the frame to every other attached CE, on a different ESI than the one it received the frame on, as well as to all other MESs participating in the E-VPN. If, on the other hand, the frame arrived from another MES (say MESy), MESx must send a copy of the packet only to attached CEs. MESx MUST NOT send the frame to other MESs, since MESy 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 MESes.

The MESes 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 MESes. Or they may use RSVP-TE P2MP or LDP P2MP or LDP MP2MP LSPs for sending such traffic to other MESes.

15.1. Ingress Replication

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

The MES 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 MES MUST treat the packet as an unknown unicast packet.

15.2. P2MP MPLS LSPs

The procedures for using P2MP LSPs are very similar to VPLS procedures [VPLS-MCAST]. The P-Tunnel attribute used by an MES for sending unknown unicast, broadcast or multicast traffic for a particular Ethernet segment, 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 E-VPNs, may use the same P2MP LSP, using upstream labels [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

The MES 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 MES MUST treat the packet as an unknown unicast packet.

16. Forwarding Unicast Packets

16.1. Forwarding packets received from a CE

When an MES 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 used for local MAC address learning.

If the MES decides to forward the packet the destination MAC address of the packet must be looked up. If the MES has received MAC address advertisements for this destination MAC address from one or more other MESes or learned it from locally connected CEs, it is

considered as a known MAC address. Else the MAC address is considered as an unknown MAC address.

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

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

- The MES MUST flood the packet to other MESes. If the ESI over which the MES receives the packet is multi-homed, then the MES MUST first encapsulate the packet in the ESI MPLS label as described in section "Split Horizon".
  If ingress replication is used the packet MUST be replicated one or more times to each remote MES with the bottom label of the stack being an MPLS label determined as follows. This is the MPLS label advertised by the remote MES in a PMSI Tunnel Attribute in the Inclusive Multicast Ethernet Tag route for an <ESI, Ethernet Tag> combination. The Ethernet Tag in the route must be the same as the Ethernet Tag advertised by the ingress MES in its Ethernet Tag A-D route associated with the interface on which the ingress MES receives the packet. If P2MP LSPs are being used the packet MUST be sent on the P2MP LSP that the MES is the root of for the Ethernet Tag in the E-VPN. If the same P2MP LSP is used for all Ethernet Tags then all the MESes in the E-VPN MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP is used for a given Ethernet Tag in the E-VPN then only the MESes 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 admnistrative policy on the MES does not allow flooding of unknown unicast traffic:

- The MES MUST drop the packet.

16.2. Forwarding packets received from a remote MES

16.2.1. Unknown Unicast Forwarding

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

- If the MES 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 MES to flood the packet on the ESIs. In other words the default behavior is for the MES to assume that the destination MAC address is unknown unicast, broadcast or multicast and it is not required to do a destination MAC address lookup, as long as the granularity of the MPLS label included the Ethernet Tag. As an option the MES may do a destination MAC lookup to flood the packet to only a subset of the CE interfaces in the Ethernet Tag. For instance the MES 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 MES is not the designated forwarder on any of the ESIs for the Ethernet Tag, the default behavior is for it to drop the packet.

16.2.2. 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 MES 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.

17. Split Horizon

Consider a CE that is multi-homed to two or more MESes on an Ethernet segment ES1. If the CE sends a multicast, broadcast or unknown unicast packet to a particular MES, say MES1, then MES1 will forward that packet to all or subset of the other MESes in the E-VPN. In this case the MESes, other than MES1, that the CE is multi-homed to MUST drop the packet and not forward back to the CE. This is referred to as "split horizon" in this document.

In order to accomplish this each MES distributes to other MESes the "per Ethernet Segment Ethernet A-D route" as per the procedures in the section "Ethernet A-D Route per Ethernet Segment". This route is imported by the MESes connected to the Ethernet Segment and also by the MESes that have at least one E-VPN in common with the Ethernet Segment in the route. As described in the section "Ethernet A-D Route per Ethernet Segment", the route MUST carry an ESI MPLS Label Extended Community with a valid ESI MPLS label.

17.1. ESI MPLS Label: Ingress Replication

An MES that is using ingress replication for sending broadcast, multicast or unknown unicast traffic, distributes to other MESes, that belong to the Ethernet segment, a downstream assigned "ESI MPLS

label" in the Ethernet A-D route. This label MUST be programmed in the platform label space by the advertising MES. 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 MES1 and MES2 that are multi-homed to CE1 on ES1. Further consider that MES1 is using P2P or MP2P LSPs to send packets to MES2. Consider that MES1 receives a multicast, broadcast or unknown unicast packet from CE1 on VLAN1 on ESI1.

First consider the case where MES2 distributes an unique Inclusive Multicast Ethernet Tag route for VLAN1, for each Ethernet segment on MES2. In this case MES1 MUST NOT replicate the packet to MES2 for <ESI1, VLAN1>.

Next consider the case where MES2 distributes a single Inclusive Multicast Ethernet Tag route for VLAN1 for all Ethernet segments on MES2. In this case when MES1 sends a multicast, broadcast or unknown unicast packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that MES2 has distributed for ESI1. It MUST then push on the MPLS label distributed by MES2 in the Inclusive Ethernet Tag Multicast route for Ethernet Tag1. The resulting packet is further encapsulated in the P2P or MP2P LSP label stack required to transmit the packet to MES2. When MES2 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 MES2 then MES2 MUST NOT forward the packet onto ESI1.

17.2. ESI MPLS Label: P2MP MPLS LSPs

An MES that is using P2MP LSPs for sending broadcast, multicast or unknown unicast traffic, distributes to other MESes, that belong to the Ethernet segment or have an E-VPN in common with the Ethernet Segment, an upstream assigned "ESI MPLS label" in the Ethernet A-D route. This label is upstream assigned by the MES that advertises the route. This label MUST be programmed by the other MESes, that are connected to the ESI advertised in the route, in the context label space for the advertising MES. 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 MESes, that import the route but are not connected to the ESI advertised in the route, in the context label space for the advertising MES. Further the forwarding entry for this label must be a POP with no other associated action.

Consider MES1 and MES2 that are multi-homed to CE1 on ES1. Also consider MES3 that is in the same E-VPN as one of the E-VPNs to which ES1 belongs. Further assume that MES1 is using P2MP MPLS LSPs to send broadcast, multicast or uknown unicast packets. When MES1 sends a multicast, broadcast or unknown unicast 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 MESes. Penultimate hop popping MUST be disabled on the P2MP LSPs used in the MPLS transport infrastructure for E-VPN. When MES2 receives this packet it decapsulates 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 MES1 then MES2 MUST NOT forward the packet onto ESI1. When MES3 receives this packet it decapsulates 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 MES1 then MES3 MUST pop the label.

17.3. ESI MPLS Label: MP2MP LSPs

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

18. Load Balancing of Unicast Packets

This section specifies how load balancing is achieved to/from a CE that has more than one interface that is directly connected to one or more MESes. The CE may be a host or a router or it may be a switched network that is connected via LAG to the MESes.

18.1. Load balancing of traffic from an MES to remote CEs

Whenever a remote MES imports a MAC advertisement for a given <ESI, Ethernet Tag> in an E-VPN instance, it MUST consider the MAC as reachahable via all the MESes from which it has imported Ethernet A-D routes for that <ESI, Ethernet Tag>. Let us call this the initial Ethernet A-D route set for the given ESI.

For the given ESI the remote MES has imported a per Ethernet Segment Ethernet A-D route, from at least one MES, where the "Active-Standby" flag in the ESI MPLS Label Extended Community is set, then the remote MES MUST first use the procedures in the section "Designated Forwarder Election" to pick a Designated Forwarder. The eligible set of Ethernet A-D routes used in the procedures below must comprise this single Ethernet A-D route from the DF.

If for the given ESI none of the per Ethernet Segment Ethernet A-D

routse, imported by the remote MES, have the "Active-Standby" flag set in the ESI MPLS Label Extended Community, then the eligble set of Ethernet A-D routes is set to the initial Ethernet A-D route set.

The remote MES MUST use the MAC advertisement and eligible Ethernet A-D routes to constuct 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 MES to forward the packet. This label stack is determined as follows. If the next-hop is constructed as a result of a MAC route which has a valid MPLS label stack, then this label stack MUST be used. However if the MAC route doesn't exist or if it doesn't have a valid MPLS label stack then the next-hop and MPLS label stack is constructed as a result of one or more corresponding Ethernet A-D routes as follows. Note that the following description applies to determining the label stack for a particular next-hop to reach a given MES, from which the remote MES has received and imported one or more 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 MES.

If there is a corresponding Ethernet A-D route for that <ESI, Ethernet Tag> then that label stack MUST be used. If such an Ethernet Tag A-D route doesn't exist but Ethernet A-D routes exist for <ESI, Ethernet Tag = 0> and <ESI = 0, Ethernet Tag> then the label stack must be constructed by using the labels from these two routes. If this is not the case but an Ethernet A-D route exists for <ESI, Ethernet Tag = 0> then the label from that route must be used. Finally if this is also not the case but an Ethernet A-D route exists for <ESI = 0, Ethernet Tag = 0> then the label from that route must be used.

The following example explains the above when Ethernet A-D routes are advertised per <ESI, Ethernet Tag>.

Consider a CE, CE1, that is dual homed to two MESes, MES1 and MES2 on a LAG interface, ES1, and is sending packets with MAC address MAC1 on VLAN1. Based on E-VPN extensions described in sections "Determining Reachability of Unicast Addresses" and "Auto-Discovery of Ethernet Tags on Ethernet Segments", a remote MES say MES3 is able to learn that a MAC1 is reachable via MES1 and MES2. Both MES1 and MES2 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 MES1, MES3 still considers MAC1 as reachable via both MES1 and MES2 as both MES1 and MES2 advertise a Ethernet A-D route for <ESI1, VLAN1>.

The MPLS label stack to send the packets to MES1 is the MPLS LSP stack to get to MES1 and the E-VPN label advertised by MES1 for CE1's

MAC.

The MPLS label stack to send packets to MES2 is the MPLS LSP stack to get to MES2 and the MPLS label in the Ethernet A-D route advertised by MES2 for <ES1, VLAN1>, if MES2 has not advertised MAC1 in BGP.

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

The remote MES, MES3, can now load balance the traffic it receives from its CEs, destined for CE1, between MES1 and MES2. MES3 may use the IP flow information for it to hash into one of the MPLS next-hops for load balancing for IP traffic. Or MES3 may rely on the source and destination MAC addresses for load balancing.

Note that once MES3 decides to send a particular packet to MES1 or MES2 it can pick from more than path to reach the particular remote MES using regular MPLS procedures. For instance if the tunneling technology is based on RSVP-TE LSPs, and MES3 decides to send a particular packet to MES1 then MES3 can choose from multiple RSVP-TE LSPs that have MES1 as their destination.

When MES1 or MES2 receive the packet destined for CE1 from MES3, 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 MES1 or MES2 must forward the packet to the CE. Which of MES1 or MES2 forward this packet to the CE is determined by default based on which of the two is the DF. An alternate procedure to load balance multicast packets will be described in the future.

If the connectivity between the multi-homed CE and one of the MESes that it is multi-homed to fails, the MES MUST withdraw the MAC address from BGP. In addition the MES MUST withdraw the Ethernet Tag A-D routes, that had been previously advertised, for the Ethernet Segment to the CE. Note that to aid convergence the Ethernet Tag A-D routes MAY be withdrawn before the MAC routes. This enables the remote MESes to remove the MPLS next-hop to this particular MES 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 MES to CE failures please section "MES to CE Network Failures".

18.2. Load balancing of traffic between an MES and a local CE

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

18.2.1. Data plane learning

Consider that the MESes perform data plane learning for local MAC addresses learned from local CEs. This enables the MES(s) to learn a particular MAC address and associate it with one or more interfaces, if the technology between the MES and the CE supports multi-pathing. The MESes 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.

18.2.2. 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 MES(s) to learn the host's MAC address and associate it with one or more interfaces. The MESes 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 MES that receives the traffic employs E-VPN forwarding procedures to forward the traffic.

19. MAC Moves

In the case where a CE is a host or a switched network connected to hosts, the MAC address that is reachable via a given MES on a particular ESI may move such that it becomes reachable via another MES on another ESI. This is referred to as a "MAC Move".

Remote MESes must be able to distinguish a MAC move from the case where a MAC address on an ESI is reachable via two different MESes and load balancing is performed as described in section "Load Balancing of Unicast Packets". This distinction can be made as follows. If a MAC is learned by a particular MES from multiple MESes, then the MES performs load balancing only amongst the set of MESes that advertised the MAC with the same ESI. If this is not the case then the MES chooses only one of the advertising MESes to reach the MAC as per BGP path selection.

There can be traffic loss during a MAC move. Consider MAC1 that is advertised by MES1 and learned from CE1 on ESI1. If MAC1 now moves behind MES2, on ESI2, MES2 advertises the MAC in BGP. Until a remote MES, MES3, determines that the best path is via MES2, it will continue to send traffic destined for MAC1 to MES1. This will not occur deterministially until MES1 withdraws the advertisement for MAC1.

One recommended optimization to reduce the traffic loss during MAC

moves is the following option. When an MES sees a MAC update from a locally attached CE on an ESI, which is different from the ESI on which the MES has currently learned the MAC, the corresponding entry in the local bridge forwarding table SHOULD be immediately purged causing the MES to withdraw its own E-VPN MAC advertisement route and replace it with the update.

A future version of this specification will describe other optimized procedures to minimize traffic loss during MAC moves.

20. Multicast

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

20.1. Ingress Replication

The MESes 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 MESes. However a given multicast packet for a multicast flow may be sent to only a subset of the MESes. Specifically a given multicast flow may be sent to only those MESes that have receivers that are interested in the multicast flow. Determining which of the MESes have receivers for a given multicast flow is done using explicit tracking described below.

20.2. P2MP LSPs

A MES 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 MES-MES tunnels (using MPLS or IP/GRE encapsulation) and P2MP LSPs are supported, and the supported P2MP LSP signaling protocols are RSVP-TE, and mLDP.

20.3. 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 MESes in the E-VPN. Upon receiving the Inclusive Multicast Tag routes with a PMSI Tunnel attribute that contains the MP2MP Tunnel identifier, the receiving MESes initiate the setup of the MP2MP tunnel towards the

root using the procedures in [MLDP].

20.3.1. 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 E-VPN instances on a given MES. 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 MES that is a member of any of the E-VPNs that are using the tree. This implies that an MES 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 MES 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 Ethernet A-D route as described in section "Handling of Multi-Destination Traffic". Note that an MES can "aggregate" multiple inclusive trees for different E-VPNs 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 Ethernet A-D routes.

20.3.2. Selective Trees

A Selective P-Multicast tree is used by an MES to send IP multicast traffic for one or more specific IP multicast streams, originated by CEs connected to the MES, that belong to the same or different E-VPNs, to a subset of the MESs that belong to those E-VPNs. Each of the MESs 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 MES 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 MES routers that have receivers in these streams. This avoids flooding other MES routers in the E-VPN.

A SP can use both Inclusive P-Multicast trees and Selective P-

Multicast trees or either of them for a given E-VPN on an MES, based on local configuration.

The granularity of a selective tree is <RD, MES, 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 MES 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].

20.4. 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 MES to request multicast membership information for a given (S, G), from leaf MESs. Leaf MESs rely on IGMP snooping or PIM snooping between the MES 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.

21. Convergence

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

21.1. Transit Link and Node Failures between MESes

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 MESes.

21.2. MES Failures

Consider a host host1 that is dual homed to MES1 and MES2. If MES1 fails, a remote MES, MES3, 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. MES3 can update its forwarding state to start sending all traffic for host1 to only MES2. 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 MES3 would have to rely on re-learning of MAC addresses via MES2.

21.2.1. Local Repair

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

21.3. MES to CE Network Failures

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

The Ethernet A-D routes should be used by an implementation to optimize the withdrawal of MAC advertisement routes. When an MES receives a withdrawal of a particular Ethernet A-D route from an MES 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 MES, as having been withdrawn. This optimizes the network convergence times in the event of MES to CE failures.

22. 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 MESes connected to a given CE should synchronize [802.1AX] LACP state amongst each other. This ensures that the MESes 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 MESes 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 MESes 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 MESes 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.

23. Acknowledgements

We would like to thank Yakov Rekhter, Pedro Marques, Kaushik Ghosh, Nischal Sheth, Robert Raszuk and Amit Shukla 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 for his review.

24. References

[E-VPN-REQ] A. Sajassi, R. Aggarwal et. al., "Requirements for Ethernet VPN", draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt

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

[VPLS-MCAST] "Multicast in VPLS". R. Aggarwal et.al., draft-ietf-l2vpn-vpls-mcast-04.txt

[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.

[VPLS-MULTIHOMING] "BGP based Multi-homing in Virtual Private LAN Service", K. Kompella et. al., draft-ietf-l2vpn-vpls-multihoming-00.txt

[PIM-SNOOPING] "PIM Snooping over VPLS", V. Hemige et. al., draft-ietf-l2vpn-vpls-pim-snooping-01

[IGMP-SNOOPING] "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", M. Christensen et. al., RFC4541,

[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

25. Author's Address

Rahul Aggarwal
Email: raggarwa_1@yahoo.com

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

Email: sajassi@cisco.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

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US
Email: jdrake@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