Internet Engineering Task Force (IETF)
Request for Comments: 7432
Category: Standards Track
ISSN: 2070-1721

A. Sajassi, Ed.
Cisco
R. Aggarwal
Arktan
N. Bitar
Verizon
A. Isaac
Bloomberg
J. Uttaro
AT&T
J. Drake
Juniper Networks
W. Henderickx
Alcatel-Lucent
February 2015

BGP MPLS-Based Ethernet VPN

Abstract

This document describes procedures for BGP MPLS-based Ethernet VPNs (EVPN). The procedures described here meet the requirements specified in RFC 7209 -- "Requirements for Ethernet VPN (EVPN)".

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7432.

Table of Contents

1. Introduction

Virtual Private LAN Service (VPLS), as defined in [RFC4664], [RFC4761], and [RFC4762], is a proven and widely deployed technology. However, the existing solution has a number of limitations when it comes to multihoming and redundancy, multicast optimization, provisioning simplicity, flow-based load balancing, and multipathing; these limitations are important considerations for Data Center (DC) deployments. [RFC7209] describes the motivation for a new solution to address these limitations. It also outlines a set of requirements that the new solution must address.

This document describes procedures for a BGP MPLS-based solution called Ethernet VPN (EVPN) to address the requirements specified in [RFC7209]. Please refer to [RFC7209] for the detailed requirements and motivation. EVPN requires extensions to existing IP/MPLS protocols as described in this document. In addition to these extensions, EVPN uses several building blocks from existing MPLS technologies.

2. 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].

3. Terminology

Broadcast Domain: In a bridged network, the broadcast domain corresponds to a Virtual LAN (VLAN), where a VLAN is typically represented by a single VLAN ID (VID) but can be represented by several VIDs where Shared VLAN Learning (SVL) is used per [802.1Q].

Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.

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

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

Ethernet Tag: An Ethernet tag identifies a particular broadcast domain, e.g., a VLAN. An EVPN instance consists of one or more broadcast domains.

LACP: Link Aggregation Control Protocol.

MP2MP: Multipoint to Multipoint.

MP2P: Multipoint to Point.

P2MP: Point to Multipoint.

P2P: Point to Point.

PE: Provider Edge device.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

4. BGP MPLS-Based EVPN Overview

This section provides an overview of EVPN. An EVPN instance comprises Customer Edge devices (CEs) that are connected to Provider Edge devices (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 EVPN instances in the provider's network.

The PEs may be connected by an MPLS Label Switched Path (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 (Generic Routing Encapsulation) tunneling or other IP tunneling can be used between the PEs. The detailed procedures in 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 Packet Switched Network (PSN) tunneling technology.

In an EVPN, MAC learning between PEs occurs not in the data plane (as happens with traditional bridging in VPLS [RFC4761] [RFC4762]) 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 [RFC4364]). This provides flexibility and the ability to preserve the "virtualization" or isolation of groups of interacting agents (hosts, servers, virtual machines) from each other. In EVPN, 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 Multiprotocol BGP (MP-BGP). Control-plane learning enables load balancing of traffic to and from CEs that are multihomed 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, the Link Layer Discovery Protocol (LLDP), IEEE 802.1aq, Address Resolution Protocol (ARP), management plane, or other protocols.

It is a local decision as to whether the Layer 2 forwarding table on a 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 EVPN are very similar to those of IP-VPN. An EVPN instance requires a Route Distinguisher (RD) that is unique per MAC-VRF and one or more globally unique Route Targets (RTs). A CE attaches to a MAC-VRF on a PE, on an Ethernet interface that may be configured for one or more Ethernet tags, e.g., VLAN IDs. Some deployment scenarios guarantee uniqueness of VLAN IDs across EVPN instances: all points of attachment for a given EVPN instance use the same VLAN ID, and no other EVPN instance uses this VLAN ID. This document refers to this case as a "Unique VLAN EVPN" and describes simplified procedures to optimize for it.

5. Ethernet Segment

As indicated in [RFC7209], each Ethernet segment needs a unique identifier in an EVPN. This section defines how such identifiers are assigned and how they are encoded for use in EVPN signaling. Later sections of this document describe the protocol mechanisms that utilize the identifiers.

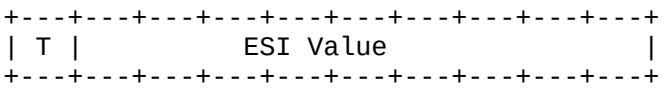
When a customer site is connected to one or more PEs via a set of Ethernet links, then this set of Ethernet links constitutes an "Ethernet segment". For a multihomed site, each Ethernet segment (ES) is identified by a unique non-zero identifier called an Ethernet Segment Identifier (ESI). An ESI is encoded as a 10-octet integer in line format with the most significant octet sent first. The following two ESI values are reserved:

- ESI 0 denotes a single-homed site.

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

In general, an Ethernet segment SHOULD have a non-reserved ESI that is unique network wide (i.e., across all EVPN instances on all the PEs). If the CE(s) constituting an Ethernet segment is (are) managed by the network operator, then ESI uniqueness should be guaranteed; 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 Designated Forwarder (DF) election.

In a network with managed and non-managed CEs, the ESI has the following format:



Where:

T (ESI Type) is a 1-octet field (most significant octet) that specifies the format of the remaining 9 octets (ESI Value). The following six ESI types can be used:

- Type 0 (T=0x00) - This type indicates an arbitrary 9-octet ESI value, which is managed and configured by the operator.

- Type 1 (T=0x01) - When IEEE 802.1AX LACP is used between the PEs and CEs, this ESI type indicates an auto-generated ESI value determined from LACP by concatenating the following parameters:

  + CE LACP System MAC address (6 octets). The CE LACP System MAC address MUST be encoded in the high-order 6 octets of the ESI Value field.

  + CE LACP Port Key (2 octets). The CE LACP port key MUST be encoded in the 2 octets next to the System MAC address.

  + The remaining octet will be set to 0x00.

  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.

- Type 2 (T=0x02) - This type is used in the case of indirectly connected hosts via a bridged LAN between the CEs and the PEs. The ESI Value is auto-generated and determined based on the Layer 2 bridge protocol as follows: If the Multiple Spanning Tree Protocol (MSTP) is used in the bridged LAN, then the value of the ESI is derived by listening to Bridge PDUs (BPDUs) on the Ethernet segment. To achieve this, the PE is not required to run MSTP. 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 Value is constructed as follows:

  + Root Bridge MAC address (6 octets). The Root Bridge MAC address MUST be encoded in the high-order 6 octets of the ESI Value field.

  + Root Bridge Priority (2 octets). The CE Root Bridge Priority MUST be encoded in the 2 octets next to the Root Bridge MAC address.

  + The remaining octet will be set to 0x00.

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

- Type 3 (T=0x03) - This type indicates a MAC-based ESI Value that can be auto-generated or configured by the operator. The ESI Value is constructed as follows:

  + System MAC address (6 octets). The PE MAC address MUST be encoded in the high-order 6 octets of the ESI Value field.

  + Local Discriminator value (3 octets). The Local Discriminator value MUST be encoded in the low-order 3 octets of the ESI Value.

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

- Type 4 (T=0x04) - This type indicates a router-ID ESI Value that can be auto-generated or configured by the operator. The ESI Value is constructed as follows:

  + Router ID (4 octets). The system router ID MUST be encoded in the high-order 4 octets of the ESI Value field.

  + Local Discriminator value (4 octets). The Local Discriminator value MUST be encoded in the 4 octets next to the IP address.

  + The low-order octet of the ESI Value will be set to 0x00.

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

- Type 5 (T=0x05) - This type indicates an Autonomous System (AS)-based ESI Value that can be auto-generated or configured by the operator. The ESI Value is constructed as follows:

  + AS number (4 octets). This is an AS number owned by the system and MUST be encoded in the high-order 4 octets of the ESI Value field. If a 2-octet AS number is used, the high-order extra 2 octets will be 0x0000.

  + Local Discriminator value (4 octets). The Local Discriminator value MUST be encoded in the 4 octets next to the AS number.

  + The low-order octet of the ESI Value will be set to 0x00.

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

6. Ethernet Tag ID

An Ethernet Tag ID is a 32-bit field containing either a 12-bit or 24-bit identifier that identifies a particular broadcast domain (e.g., a VLAN) in an EVPN instance. The 12-bit identifier is called the VLAN ID (VID). An EVPN instance consists of one or more broadcast domains (one or more VLANs). VLANs are assigned to a given EVPN instance by the provider of the EVPN service. A given VLAN can itself be represented by multiple VIDs. In such cases, the PEs participating in that VLAN for a given EVPN instance are responsible for performing VLAN ID translation to/from locally attached CE devices.

If a VLAN is represented by a single VID across all PE devices participating in that VLAN for that EVPN instance, then there is no need for VID translation at the PEs. Furthermore, some deployment scenarios guarantee uniqueness of VIDs across all EVPN instances; all points of attachment for a given EVPN instance use the same VID, and no other EVPN instances use that VID. This allows the RT(s) for each EVPN instance to be derived automatically from the corresponding VID, as described in Section 7.10.1.

The following subsections discuss the relationship between broadcast domains (e.g., VLANs), Ethernet Tag IDs (e.g., VIDs), and MAC-VRFs as well as the setting of the Ethernet Tag ID, in the various EVPN BGP routes (defined in Section 8), for the different types of service interfaces described in [RFC7209].

The following Ethernet Tag ID value is reserved:

- Ethernet Tag ID {0xFFFFFFFF} is known as MAX-ET.

6.1. VLAN-Based Service Interface

With this service interface, an EVPN instance consists of only a single broadcast domain (e.g., a single VLAN). Therefore, there is a one-to-one mapping between a VID on this interface and a MAC-VRF. Since a MAC-VRF corresponds to a single VLAN, it consists of a single bridge table corresponding to that VLAN. If the VLAN is represented by multiple VIDs (e.g., a different VID per Ethernet segment per PE), then each PE needs to perform VID translation for frames destined to its Ethernet segment(s). In such scenarios, the Ethernet frames transported over an MPLS/IP network SHOULD remain tagged with the originating VID, and a VID translation MUST be supported in the data path and MUST be performed on the disposition PE. The Ethernet Tag ID in all EVPN routes MUST be set to 0.

6.2. VLAN Bundle Service Interface

With this service interface, an EVPN instance corresponds to multiple broadcast domains (e.g., multiple VLANs); however, only a single bridge table is maintained per MAC-VRF, which means multiple VLANs share the same bridge table. This implies that MAC addresses MUST be unique across all VLANs for that EVI in order for this service to work. In other words, there is a many-to-one mapping between VLANs and a MAC-VRF, and the MAC-VRF consists of a single bridge table. Furthermore, a single VLAN must be represented by a single VID -- e.g., no VID translation is allowed for this service interface type. The MPLS-encapsulated frames MUST remain tagged with the originating VID. Tag translation is NOT permitted. The Ethernet Tag ID in all EVPN routes MUST be set to 0.

6.2.1. 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 6.2.

6.3. VLAN-Aware Bundle Service Interface

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

Broadcast, unknown unicast, or multicast (BUM) traffic is sent only to the CEs in a given broadcast domain; however, the broadcast domains within an EVI either MAY each have their own P-Tunnel or MAY share P-Tunnels -- e.g., all of the broadcast domains in an EVI MAY share a single P-Tunnel.

In the case where a single VLAN is represented by a single VID and thus no VID translation is required, an MPLS-encapsulated packet MUST carry that VID. The Ethernet Tag ID in all EVPN routes MUST be set to that VID. The advertising PE MAY advertise the MPLS Label1 in the MAC/IP Advertisement route representing ONLY the EVI or representing both the Ethernet Tag ID and the EVI. This decision is only a local matter by the advertising PE (which is also the disposition PE) and doesn’t affect any other PEs.

In the case where a single VLAN is represented by different VIDs on different CEs and thus VID translation is required, a normalized Ethernet Tag ID (VID) MUST be carried in the EVPN BGP routes. Furthermore, the advertising PE advertises the MPLS Label1 in the MAC/IP Advertisement route representing both the Ethernet Tag ID and the EVI, so that upon receiving an MPLS-encapsulated packet, it can identify the corresponding bridge table from the MPLS EVPN label and perform Ethernet Tag ID translation ONLY at the disposition PE -- i.e., the Ethernet frames transported over the MPLS/IP network MUST remain tagged with the originating VID, and VID translation is performed on the disposition PE. The Ethernet Tag ID in all EVPN routes MUST be set to the normalized Ethernet Tag ID assigned by the EVPN provider.

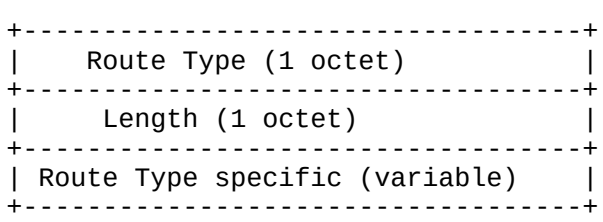
6.3.1. Port-Based VLAN-Aware Service Interface

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 are mapped to a single bundle but without any VID translation. The procedures are a subset of those described in Section 6.3.

7. BGP EVPN Routes

This document defines a new BGP Network Layer Reachability Information (NLRI) called the EVPN NLRI.

The format of the EVPN NLRI is as follows:



The Route Type field defines the encoding of the rest of the EVPN NLRI (Route Type specific EVPN NLRI).

The Length field indicates the length in octets of the Route Type specific field of the EVPN NLRI.

This document defines the following Route Types:

+ 1 - Ethernet Auto-Discovery (A-D) route
+ 2 - MAC/IP Advertisement route
+ 3 - Inclusive Multicast Ethernet Tag route
+ 4 - Ethernet Segment route

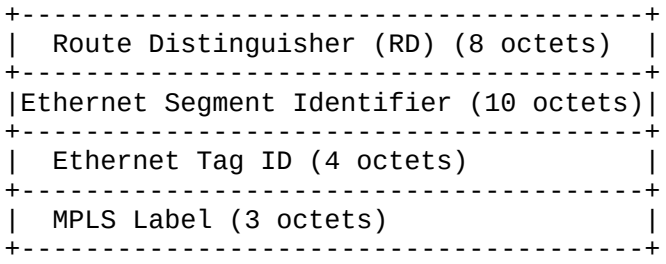
The detailed encoding and procedures for these route types are described in subsequent sections.

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

In order for two BGP speakers to exchange labeled EVPN NLRI, they must use BGP Capabilities Advertisements 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 (EVPN).

7.1. Ethernet Auto-discovery Route

An Ethernet A-D route type specific EVPN NLRI consists of the following:

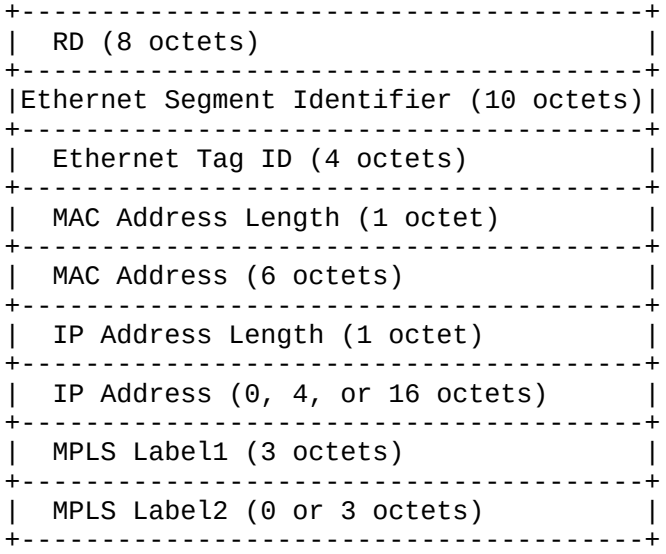


For the purpose of BGP route key processing, only the Ethernet Segment Identifier and the Ethernet Tag ID are considered to be part of the prefix in the NLRI. The MPLS Label field is to be treated as a route attribute as opposed to being part of the route.

For procedures and usage of this route, please see Sections 8.2 ("Fast Convergence") and 8.4 ("Aliasing and Backup Path").

7.2. MAC/IP Advertisement Route

A MAC/IP Advertisement route type specific EVPN NLRI consists of the following:

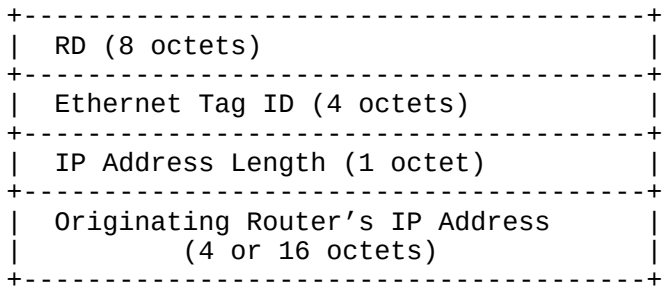


For the purpose of BGP route key processing, only the Ethernet Tag ID, MAC Address Length, MAC Address, IP Address Length, and IP Address fields are considered to be part of the prefix in the NLRI. The Ethernet Segment Identifier, MPLS Label1, and MPLS Label2 fields are to be treated as route attributes as opposed to being part of the "route". Both the IP and MAC address lengths are in bits.

For procedures and usage of this route, please see Sections 9 ("Determining Reachability to Unicast MAC Addresses") and 14 ("Load Balancing of Unicast Packets").

7.3. Inclusive Multicast Ethernet Tag Route

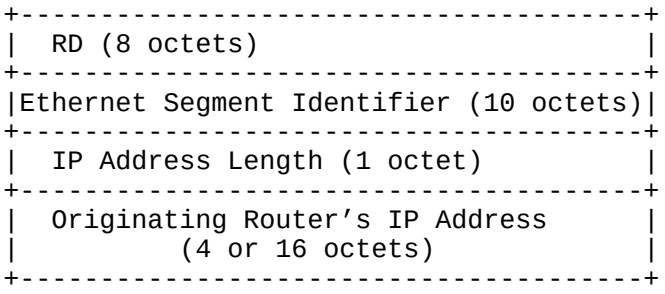
An Inclusive Multicast Ethernet Tag route type specific EVPN NLRI consists of the following:



For procedures and usage of this route, please see Sections 11 ("Handling of Multi-destination Traffic"), 12 ("Processing of Unknown Unicast Packets"), and 16 ("Multicast and Broadcast"). The IP address length is in bits. For the purpose of BGP route key processing, only the Ethernet Tag ID, IP Address Length, and Originating Router’s IP Address fields are considered to be part of the prefix in the NLRI.

7.4. Ethernet Segment Route

An Ethernet Segment route type specific EVPN NLRI consists of the following:

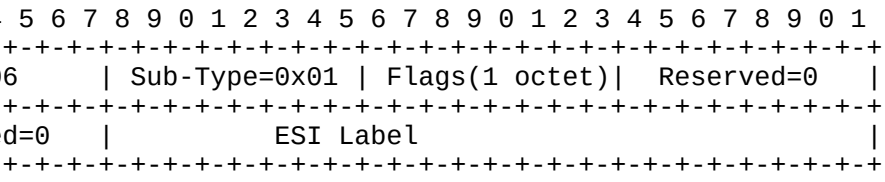


For procedures and usage of this route, please see Section 8.5 ("Designated Forwarder Election"). The IP address length is in bits. For the purpose of BGP route key processing, only the Ethernet Segment ID, IP Address Length, and Originating Router's IP Address fields are considered to be part of the prefix in the NLRI.

7.5. ESI Label Extended Community

This Extended Community is a new transitive Extended Community having a Type field value of 0x06 and the Sub-Type 0x01. It may be advertised along with Ethernet Auto-discovery routes, and it enables split-horizon procedures for multihomed sites as described in Section 8.3 ("Split Horizon"). The ESI Label field represents an ES by the advertising PE, and it is used in split-horizon filtering by other PEs that are connected to the same multihomed Ethernet segment.

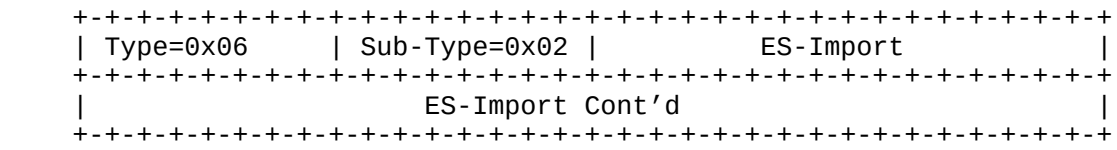
Each ESI Label extended community is encoded as an 8-octet value, as follows:



The low-order bit of the Flags octet is defined as the "Single-Active" bit. A value of 0 means that the multihomed site is operating in All-Active redundancy mode, and a value of 1 means that the multihomed site is operating in Single-Active redundancy mode.

7.6. 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 multihomed site to import the Ethernet Segment routes. The value is derived automatically for the ESI Types 1, 2, and 3, by encoding the high-order 6-octet portion of the 9-octet ESI Value, which corresponds to a MAC address, in the ES-Import Route Target. The format of this Extended Community is as follows:



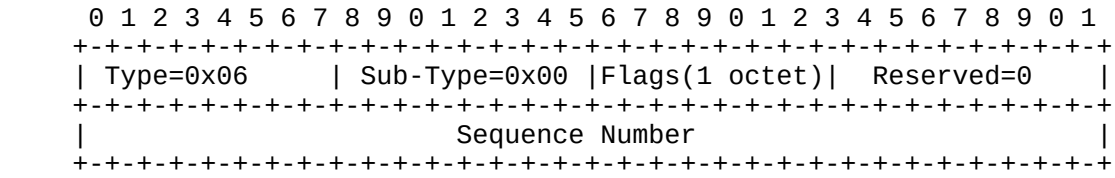
This document expands the definition of the Route Target extended community to allow the value of the high-order octet (Type field) to be 0x06 (in addition to the values specified in [RFC4360]). The low-order octet (Sub-Type field) value 0x02 indicates that this Extended Community is of type "Route Target". The new Type field value 0x06 indicates that the structure of this RT is a 6-octet value (e.g., a MAC address). A BGP speaker that implements RT Constraint [RFC4684] MUST apply the RT Constraint procedures to the ES-Import RT as well.

For procedures and usage of this attribute, please see Section 8.1 ("Multihomed Ethernet Segment Auto-discovery").

7.7. MAC Mobility Extended Community

This Extended Community is a new transitive Extended Community having a Type field value of 0x06 and the Sub-Type 0x00. It may be advertised along with MAC/IP Advertisement routes. The procedures for using this Extended Community are described in Section 15 ("MAC Mobility").

The MAC Mobility extended community is encoded as an 8-octet value, as follows:



The low-order bit of the Flags octet is defined as the "Sticky/static" flag and may be set to 1. A value of 1 means that the MAC address is static and cannot move. The sequence number is used to ensure that PEs retain the correct MAC/IP Advertisement route when multiple updates occur for the same MAC address.

7.8. Default Gateway Extended Community

The Default Gateway community is an Extended Community of an Opaque Type (see Section 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 0x0d (Default Gateway) as assigned by IANA. The Value field of this community is reserved (set to 0 by the senders, ignored by the receivers). For procedures and usage of this attribute, please see Section 10.1 ("Default Gateway").

7.9. Route Distinguisher Assignment per MAC-VRF

The Route Distinguisher (RD) MUST be set to the RD of the MAC-VRF that is advertising the NLRI. An RD MUST be assigned for a given MAC-VRF on a PE. This RD MUST be unique across all MAC-VRFs on a 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 EVPN case, the low-order 12 bits may be the 12-bit VLAN ID, with the remaining high-order 4 bits set to 0.

7.10. Route Targets

The EVPN route MAY carry one or more Route Target (RT) attributes. RTs may be configured (as in IP VPNs) or may be derived automatically.

If a PE uses RT Constraint, the PE advertises all such RTs using RT Constraints per [RFC4684]. The use of RT Constraints allows each EVPN route to reach only those PEs that are configured to import at least one RT from the set of RTs carried in the EVPN route.

7.10.1. Auto-derivation from the Ethernet Tag ID

For the "Unique VLAN EVPN" scenario, it is highly desirable to auto-derive the RT from the Ethernet Tag ID (VLAN ID) for that EVPN instance. The procedure for performing such auto-derivation is as follows:

+ The Global Administrator field of the RT MUST be set to the Autonomous System (AS) number with which the PE is associated.

+ The 12-bit VLAN ID MUST be encoded in the lowest 12 bits of the Local Administrator field, with the remaining bits set to zero.

8. Multihoming Functions

This section discusses the functions, procedures, and associated BGP routes used to support multihoming in EVPN. This covers both multihomed device (MHD) and multihomed network (MHN) scenarios.

8.1. Multihomed 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.

8.1.1. 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 PE (typically, the loopback address) followed by a number unique to the PE.

The Ethernet Segment Identifier (ESI) MUST be set to the 10-octet value described in Section 5.

The BGP advertisement that advertises the Ethernet Segment route MUST also carry an ES-Import Route Target, as defined in Section 7.6.

The Ethernet Segment route filtering MUST be done such that the Ethernet Segment route is imported only by the PEs that are multihomed 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 Route Target, constructed from the ESI.

8.2. Fast Convergence

In EVPN, MAC address reachability is learned 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/IP Advertisement routes that must be withdrawn by the PE encountering a failure. For highly scaled environments, this scheme yields slow convergence.

To alleviate this, EVPN 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 a set of one or more Ethernet A-D per ES routes for each locally attached Ethernet segment (refer to Section 8.2.1 below for details on how these routes are constructed). A PE may need to advertise more than one Ethernet A-D per ES route for a given ES because the ES may be in a multiplicity of EVIs and the RTs for all of these EVIs may not fit into a single route. Advertising a set of Ethernet A-D per ES routes for the ES allows each route to contain a subset of the complete set of RTs. Each Ethernet A-D per ES route is differentiated from the other routes in the set by a different Route Distinguisher (RD).

Upon a failure in connectivity to the attached segment, the PE withdraws the corresponding set of Ethernet A-D per ES routes. 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 its next-hop adjacencies accordingly.

8.2.1. Constructing Ethernet A-D per Ethernet Segment Route

This section describes the procedures used to construct the Ethernet A-D per ES route, which is used for fast convergence (as discussed above) and for advertising the ESI label used for split-horizon filtering (as discussed in Section 8.3). Support of this route is REQUIRED.

The 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 a number unique to the PE.

The Ethernet Segment Identifier MUST be a 10-octet entity as described in Section 5 ("Ethernet Segment"). The Ethernet A-D route is not needed when the Segment Identifier is set to 0 (e.g., single-homed scenarios).

The Ethernet Tag ID MUST be set to MAX-ET.

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 redundancy mode is desired, then the "Single-Active" 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 the ESI label and MUST have the same value in each Ethernet A-D per ES route advertised for the ES. 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 8.3.

If Single-Active redundancy mode is desired, then the "Single-Active" bit in the flags of the ESI Label extended community MUST be set to 1 and the ESI label SHOULD be set to a valid MPLS label value.

8.2.1.1. Ethernet A-D Route Targets

Each Ethernet A-D per ES route MUST carry one or more Route Target (RT) attributes. The set of Ethernet A-D routes per ES MUST carry the entire set of RTs for all the EVPN instances to which the Ethernet segment belongs.

8.3. Split Horizon

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

When a set of PEs are operating in Single-Active redundancy mode, the use of this split-horizon filtering mechanism is highly recommended because it prevents transient loops at the time of failure or recovery that would impact the Ethernet segment -- e.g., when two PEs think that both are DFs for that segment before the DF election procedure settles down.

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 EVPN network). This label is referred to as the ESI label and MUST be distributed by all PEs when operating in All-Active redundancy mode using a set of Ethernet A-D per ES routes, per Section 8.2.1 above. The ESI label SHOULD be distributed by all PEs when operating in Single-Active redundancy mode using a set of Ethernet A-D per ES routes. These routes are imported by the PEs connected to the Ethernet segment and also by the PEs that have at least one EVPN instance in common with the Ethernet segment in the route. As described in Section 8.1.1, the route MUST carry an ESI Label extended community with a valid ESI label. The disposition PE relies on the value of the ESI label to determine whether or not a BUM frame is allowed to egress a specific Ethernet segment.

8.3.1. 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 EVPN network.

8.3.1.1. Ingress Replication

Each PE that operates in All-Active or Single-Active redundancy mode and that uses ingress replication to receive BUM traffic advertises a downstream assigned ESI label in the set of Ethernet A-D per ES routes for its attached ES. This label MUST be programmed in the platform label space by the advertising PE, and the forwarding entry

for this label must result in NOT forwarding packets received with this label onto the Ethernet segment for which the label was distributed.

The rules for the inclusion of the ESI label in a BUM packet by the ingress PE operating in All-Active redundancy mode are as follows:

- A non-DF ingress PE MUST include the ESI label distributed by the DF egress PE in the copy of a BUM packet sent to it.

- An ingress PE (DF or non-DF) SHOULD include the ESI label distributed by each non-DF egress PE in the copy of a BUM packet sent to it.

The rule for the inclusion of the ESI label in a BUM packet by the ingress PE operating in Single-Active redundancy mode is as follows:

- An ingress DF PE SHOULD include the ESI label distributed by the egress PE in the copy of a BUM packet sent to it.

In both All-Active and Single-Active redundancy mode, an ingress PE MUST NOT include an ESI label in the copy of a BUM packet sent to an egress PE that is not attached to the ES through which the BUM packet entered the EVI.

As an example, consider PE1 and PE2, which are multihomed to CE1 on ES1 and operating in All-Active multihoming 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 corresponding to an EVPN instance. 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 onto the MPLS label stack 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, from the top MPLS label, the set of ESIs to which it will replicate the packet 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 that 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 packet for VLAN1 from CE1, then it SHOULD encapsulate the packet with an ESI label received from PE1 when sending it to PE1 in order to avoid any transient loops during a failure scenario that would impact ES1 (e.g., port or link failure).

8.3.1.2. P2MP MPLS LSPs

The non-DF PEs that operate in All-Active redundancy mode and that use P2MP LSPs to send BUM traffic advertise an upstream assigned ESI label in the set of Ethernet A-D per ES routes for their common attached ES. 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 for which the label was distributed. 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 label pop with no other associated action.

The DF PE that operates in Single-Active redundancy mode and that uses P2MP LSPs to send BUM traffic should advertise an upstream assigned ESI label in the set of Ethernet A-D per ES routes for its attached ES, just as described in the previous paragraph.

As an example, consider PE1 and PE2, which are multihomed to CE1 on ES1 and operating in All-Active multihoming mode. Also, consider that PE3 belongs to one of the EVPN instances of ES1. Further, assume that PE1, which is the non-DF, is 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 on which the packet was received. 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 EVPN. When PE2 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 PE1 to ES1, then PE2 MUST NOT forward the packet onto ES1. When PE3 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 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 that EVPN instance. It should be noted that when PE2 sends a BUM frame over a P2MP LSP, it should encapsulate the frame with an ESI label even though it is the DF for that VLAN, in order to avoid any transient loops during a failure scenario that would impact ES1 (e.g., port or link failure).

8.4. Aliasing and Backup Path

In the case where a CE is multihomed to multiple PE nodes, using a Link Aggregation Group (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/IP Advertisement routes for these addresses from a single PE, even though multiple PEs are connected to the multihomed segment. As a result, the remote PEs are not able to effectively load balance traffic among the PE nodes connected to the multihomed Ethernet segment. This could be the case, for example, when the PEs perform data-plane 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 address this issue, EVPN introduces the concept of 'aliasing', which is the ability of a PE to signal that it has reachability to an EVPN instance on a given ES even when it has learned no MAC addresses from that EVI/ES. The Ethernet A-D per EVI route is used for this purpose. A remote PE that receives a MAC/IP Advertisement route with a non-reserved ESI SHOULD consider the advertised MAC address to be reachable via all PEs that have advertised reachability to that MAC address's EVI/ES via the combination of an Ethernet A-D per EVI route for that EVI/ES (and Ethernet tag, if applicable) AND Ethernet A-D per ES routes for that ES with the "Single-Active" bit in the flags of the ESI Label extended community set to 0.

Note that the Ethernet A-D per EVI route may be received by a remote PE before it receives the set of Ethernet A-D per ES routes. Therefore, in order to handle corner cases and race conditions, the Ethernet A-D per EVI route MUST NOT be used for traffic forwarding by a remote PE until it also receives the associated set of Ethernet A-D per ES routes.

The backup path is a closely related function, but it is used in Single-Active redundancy mode. In this case, a PE also advertises that it has reachability to a given EVI/ES using the same combination of Ethernet A-D per EVI route and Ethernet A-D per ES route as discussed above, but with the "Single-Active" bit in the flags of the ESI Label extended community set to 1. A remote PE that receives a MAC/IP Advertisement route with a non-reserved ESI SHOULD consider the advertised MAC address to be reachable via any PE that has advertised this combination of Ethernet A-D routes, and it SHOULD install a backup path for that MAC address.

8.4.1. Constructing Ethernet A-D per EVPN Instance Route

This section describes the procedures used to construct the Ethernet A-D per EVPN instance (EVI) route, which is used for aliasing (as discussed above). Support of this route is OPTIONAL.

The Route Distinguisher (RD) MUST be set per Section 7.9.

The Ethernet Segment Identifier MUST be a 10-octet entity as described in Section 5 ("Ethernet Segment"). The Ethernet A-D route is not needed 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 EVPN. 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 per <ESI, Ethernet Tag ID> tuple per MAC-VRF. This is applicable when the PE uses MPLS-based disposition with VID translation or may be applicable when the PE uses MAC-based disposition with VID translation.

+ One Ethernet A-D route for each <ESI> per MAC-VRF (where the Ethernet Tag ID is set to 0). This is applicable when the PE uses MAC-based disposition or MPLS-based disposition without VID translation.

The usage of the MPLS label is described in Section 14 ("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.

The Ethernet A-D route MUST carry one or more Route Target (RT) attributes, per Section 7.10.

8.5. Designated Forwarder Election

Consider a CE that is a host or a router that is multihomed directly to more than one PE in an EVPN instance 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 a 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 <ES, VLAN> or <ES, VLAN bundle> for multicast, broadcast, and unknown unicast traffic, is the default behavior in this specification.

Note that a CE always sends packets belonging to a specific flow using a single link towards a 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 multihomed to more than one PE in an EVPN network via switches, then the support of All-Active redundancy mode requires the bridged network to be connected to two or more PEs using a LAG.

If a bridged network does not connect to the PEs using a LAG, then only one of the links between the bridged network and the PEs must be the active link for a given <ES, VLAN> or <ES, VLAN bundle>. In this case, the set of Ethernet A-D per ES routes advertised by each PE MUST have the "Single-Active" bit in the flags of the ESI Label extended community set to 1.

The default procedure for DF election at the granularity of <ES, VLAN> for VLAN-based service or <ES, VLAN bundle> for VLAN-(aware) bundle service is referred to as "service carving". With service carving, it is possible to elect multiple DFs per Ethernet segment (one per VLAN or VLAN bundle) in order to perform load balancing of multi-destination traffic destined to a given segment. The load-balancing procedures carve up the VLAN space per ES among the PE

nodes evenly, in such a way that every PE is the DF for a disjoint set of VLANs or VLAN bundles for that ES. 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 should be the 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. Each IP address in this list is extracted from the "Originating Router's IP address" field of the advertised Ethernet Segment route. 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 EVPN instance on the Ethernet segment, using the following rule:

   Assuming a redundancy group of N PE nodes, for VLAN-based service, the PE with ordinal i is the DF for an <ES, VLAN V> when (V mod N) = i. In the case of VLAN-(aware) bundle service, then the numerically lowest VLAN value in that bundle on that ES MUST be used in the modulo function.

   It should be noted that using the "Originating Router's IP address" field in the Ethernet Segment route to get the PE IP address needed for the ordered list allows for a CE to be multihomed across different ASes if such a need ever arises.

4. The PE that is elected as a DF for a given <ES, VLAN> or <ES, VLAN bundle> will unblock multi-destination traffic for that VLAN or VLAN bundle on the corresponding ES. 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 in the egress direction towards that <ES, VLAN> or <ES, VLAN bundle>.

   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 redundancy group. For PE node failure, or upon PE commissioning or decommissioning, the PEs re-trigger the service carving. In the case of Single-Active

multihoming, when a service moves from one PE in the redundancy group to another PE as a result of re-carving, the PE, which ends up being the elected DF for the service, SHOULD trigger a MAC address flush notification towards the associated Ethernet segment. This can be done, for example, using the IEEE 802.1ak Multiple VLAN Registration Protocol (MVRP) 'new' declaration.

8.6. Interoperability with Single-Homing PEs

Let's refer to PEs that only support single-homed CE devices as single-homing PEs. For single-homing PEs, all the above multihoming procedures can be omitted; however, to allow for single-homing PEs to fully interoperate with multihoming PEs, some of the multihoming procedures described above SHOULD be supported even by single-homing PEs:

- procedures related to processing Ethernet A-D routes for the purpose of fast convergence (Section 8.2 ("Fast Convergence")), to let single-homing PEs benefit from fast convergence

- procedures related to processing Ethernet A-D routes for the purpose of aliasing (Section 8.4 ("Aliasing and Backup Path")), to let single-homing PEs benefit from load balancing

- procedures related to processing Ethernet A-D routes for the purpose of a backup path (Section 8.4 ("Aliasing and Backup Path")), to let single-homing PEs benefit from the corresponding convergence improvement

9. 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":

9.1. 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 EVPN instance MUST support local data-plane learning using standard IEEE Ethernet learning procedures. A 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

- An ARP Request for its own MAC

- An 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 "MAC Mobility". Procedures to support this are described in Section 15 ("MAC Mobility").

9.2. 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 "remote" MAC addresses.

This document requires a 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 that EVPN instance, using MP-BGP and, specifically, the MAC/IP Advertisement route.

9.2.1. Constructing MAC/IP Address Advertisement

BGP is extended to advertise these MAC addresses using the MAC/IP Advertisement route type in the EVPN NLRI.

The RD MUST be set per Section 7.9.

The Ethernet Segment Identifier is set to the 10-octet ESI described in Section 5 ("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 tables in the MAC-VRF (i.e., the PE needs to support VLAN-aware bundle service for that EVI).

When the Ethernet Tag ID in the NLRI is set to a non-zero value for a particular broadcast domain, then this Ethernet Tag ID may be either the CE's Ethernet tag value (e.g., CE VLAN ID) or the EVPN provider's Ethernet tag value (e.g., provider VLAN ID). The latter would be the case if the CE Ethernet tags (e.g., CE VLAN ID) for a particular broadcast domain are different on different CEs.

The MAC Address Length field is in bits, and it is set to 48. MAC address length values other than 48 bits are outside the scope of this document. The encoding of a MAC address MUST be the 6-octet MAC address specified by [802.1Q] and [802.1D-REV].

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, it is then encoded in this route. When an IP address is present, the IP Address Length field is in bits, and it is set to 32 or 128 bits. Other IP Address Length values are outside the scope of this document. The encoding of an IP address MUST be either 4 octets for IPv4 or 16 octets for IPv6. The Length field of the EVPN NLRI (which is in octets and is described in Section 7) is sufficient to determine whether an IP address is encoded in this route and, if so, whether the encoded IP address is IPv4 or IPv6.

The MPLS Label1 field is encoded as 3 octets, where the high-order 20 bits contain the label value. The MPLS Label1 MUST be downstream assigned, and it is associated with the MAC address being advertised by the advertising PE. The advertising PE uses this label when it receives an MPLS-encapsulated packet to perform forwarding based on the destination MAC address toward the CE. The forwarding procedures are specified in Sections 13 and 14.

A PE may advertise the same single EVPN label for all MAC addresses in a given MAC-VRF. This label assignment is referred to as a per MAC-VRF label assignment. Alternatively, a PE may advertise a unique EVPN label per <MAC-VRF, Ethernet tag> combination. This label assignment is referred to as a per <MAC-VRF, Ethernet tag> label assignment. As a third option, a PE may advertise a unique EVPN label per <ESI, Ethernet tag> combination. This label assignment is referred to as a per <ESI, Ethernet tag> label assignment. As a fourth option, a PE may advertise a unique EVPN label per MAC address. This label assignment is referred to as a per MAC label assignment. All of these label assignment methods have their trade-offs. The choice of a particular label assignment methodology is purely local to the PE that originates the route.

An assignment per MAC-VRF label requires the least number of EVPN 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 MPLS Label2 field is an optional field. If it is present, then it is encoded as 3 octets, where the high-order 20 bits contain the label value.

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/IP 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 7.10.1.

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

9.2.2. Route Resolution

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

If the Ethernet Segment Identifier field in a received MAC/IP 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/IP Advertisement route is set to a non-reserved ESI, then if the receiving PE decides to install forwarding state for the associated MAC address, it MUST be when both the MAC/IP Advertisement route AND the associated set of Ethernet A-D per ES routes have been received. The dependency of MAC route installation on Ethernet A-D per ES routes is to ensure that MAC routes don't get accidentally installed during a mass withdraw period.

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

T1 When the MAC/IP Advertisement route from PE1 and the set of Ethernet A-D per ES routes and Ethernet A-D per EVI routes 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 set of Ethernet A-D per ES routes, then PE3 forwards traffic destined to M1 to PE2 only.

T2′ If after T1 PE2 withdraws its set of Ethernet A-D per ES routes, then PE3 forwards traffic destined to M1 to PE1 only.

T2′′ If after T1 PE1 withdraws its MAC/IP Advertisement route, then PE3 treats traffic to M1 as unknown unicast.

T3 PE2 also advertises a MAC route for M1, and then PE1 withdraws its MAC route for M1. PE3 continues forwarding traffic destined to M1 to both PE1 and PE2. In other words, despite M1 withdrawal by PE1, PE3 forwards the traffic destined to M1 to both PE1 and PE2. This is because a flow from the CE, resulting in M1 traffic getting hashed to PE1, can get terminated, resulting in M1 being aged out in PE1; however, M1 can be reachable by both PE1 and PE2.

10. ARP and ND

The IP Address field in the MAC/IP Advertisement route may optionally carry one of the IP addresses associated with the MAC address. This provides an option that 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 EVPN network. A 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 a 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/IP Advertisement route. The IP address may be an IPv4 address encoded using 4 octets or an IPv6 address encoded using 16 octets. For ARP and ND purposes, the IP Address Length field MUST be set to 32 for an IPv4 address or 128 for an IPv6 address.

If there are multiple IP addresses associated with a MAC address, then multiple MAC/IP 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 same MAC address for dual-IP-stack scenarios. When the IP address is dissociated with the MAC address, then the MAC/IP Advertisement route with that particular IP address MUST be withdrawn.

Note that a MAC-only route can be advertised along with, but independent from, a MAC/IP route for scenarios where the MAC learning over an access network/node is done in the data plane and independent from ARP snooping that generates a MAC/IP route. In such scenarios, when the ARP entry times out and causes the MAC/IP to be withdrawn, then the MAC information will not be lost. In scenarios where the host MAC/IP is learned via the management or control plane, then the sender PE may only generate and advertise the MAC/IP route. If the receiving PE receives both the MAC-only route and the MAC/IP route, then when it receives a withdraw message for the MAC/IP route, it MUST delete the corresponding entry from the ARP table but not the MAC entry from the MAC-VRF table, unless it receives a withdraw message for the MAC-only route.

When a 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 by responding to the ARP Request.

10.1. Default Gateway

When a PE needs to perform inter-subnet forwarding where each subnet is represented by a different broadcast domain (e.g., a different VLAN), the inter-subnet forwarding is performed at Layer 3, and the PE that performs such a function is called the default gateway for the EVPN instance. In this case, when the PE receives an ARP Request for the IP address configured as the default gateway address, the PE originates an ARP Reply.

Each PE that acts as a default gateway for a given EVPN instance MAY advertise in the EVPN control plane its default gateway MAC address using the MAC/IP Advertisement route, and each such PE indicates that such a route is associated with the default gateway. This is accomplished by requiring the route to carry the Default Gateway extended community defined in Section 7.8 ("Default Gateway Extended Community"). The ESI field is set to zero when advertising the MAC route with the Default Gateway extended community.

The IP Address field of the MAC/IP Advertisement route is set to the default gateway IP address for that subnet (e.g., an EVPN instance). For a given subnet (e.g., a VLAN or EVPN instance), the default gateway IP address is the same across all the participant PEs. The inclusion of this IP address enables the receiving PE to check its configured default gateway IP address against the one received in the MAC/IP Advertisement route for that subnet (or EVPN instance), and if there is a discrepancy, then the PE SHOULD notify the operator and log an error message.

Unless it is known a priori (by means outside of this document) that all PEs of a given EVPN instance act as a default gateway for that EVPN instance, the MPLS label MUST be set to a valid downstream assigned label.

Furthermore, even if all PEs of a given EVPN instance do act as a default gateway for that EVPN instance, but only some, but not all, of these PEs have sufficient (routing) information to provide inter-subnet routing for all the inter-subnet traffic originated within the subnet associated with the EVPN instance, then when such a PE advertises in the EVPN control plane its default gateway MAC address using the MAC/IP Advertisement route and indicates that such a route is associated with the default gateway, the route MUST carry a valid downstream assigned label.

If all PEs of a given EVPN instance act as a default gateway for that EVPN instance, and the same default gateway MAC address is used across all gateway devices, then no such advertisement is needed. However, if each default gateway uses a different MAC address, then each default gateway needs to be aware of other gateways' MAC addresses and thus the need for such an advertisement. This is called MAC address aliasing, since a single default gateway can be represented by multiple MAC addresses.

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.

Each PE that acts as a default gateway for a given EVPN instance 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.

11. 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 (VLAN) in an EVPN instance to all the other PEs that span that Ethernet tag (VLAN) in that EVPN instance. In certain scenarios, as described in Section 12 ("Processing of Unknown Unicast Packets"), a given PE may also need to flood unknown unicast traffic to other PEs.

The PEs in a particular EVPN instance 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 its usage in further detail.

11.1. Constructing Inclusive Multicast Ethernet Tag Route

The RD MUST be set per Section 7.9.

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 field value MUST be set to an IP address of the PE that should be common for all the EVIs on the PE (e.g., this address may be the PE’s loopback address). The IP Address Length field is in bits.

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 Inclusive Multicast Ethernet Tag route MUST also carry one or more Route Target (RT) attributes. The assignment of RTs as described in Section 7.10 MUST be followed.

11.2. 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 Provider Multicast Service Interface (PMSI) Tunnel attribute as specified in [RFC6514].

Depending on the technology used for the P-tunnel for the EVPN instance 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 EVPN, 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 EVPN instances (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 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 EVIs 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 EVPN, the route MUST include the PMSI Tunnel attribute with the Tunnel Type set to Ingress Replication and the 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 EVPN traffic received over an 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.

12. 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 learned 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. When flooding, one 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] [RFC4762]. When a PE capable of flooding (say PEx) receives an unknown destination MAC address, it floods the frame. If the frame arrived from an attached CE, PEx must send a copy of that frame on every Ethernet segment (belonging to that EVI) for which it is the DF, other than the Ethernet segment on which it received the frame. In addition, the PE must flood the frame to all other PEs participating in that EVPN instance. If, on the other hand, the frame arrived from another PE (say PEy), PEx must send a copy of the packet on each Ethernet segment (belonging to that EVI) for which it is the DF. PEx MUST NOT send the frame to other PEs, since PEy would have already done so. Split-horizon forwarding rules apply to unknown MAC addresses.

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 EVPN instance may use ingress replication using RSVP-TE P2P LSPs or LDP MP2P LSPs for sending unknown unicast traffic to other PEs. Or, they may use RSVP-TE P2MP or LDP P2MP for sending such traffic to other PEs.

12.1. Ingress Replication

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

12.2. P2MP MPLS LSPs

The procedures for using P2MP LSPs are very similar to the VPLS procedures described in [RFC7117]. The P-tunnel attribute used by a PE for sending unknown unicast, broadcast, or multicast traffic for a particular EVPN instance is advertised in the Inclusive Multicast Ethernet Tag route as described in Section 11 ("Handling of Multi-destination Traffic").

The P-tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree as described in [RFC7117]. Note that multiple Ethernet tags, which may be in different EVPN instances, may use the same P2MP LSP, using upstream labels [RFC7117]. This is the equivalent of an Aggregate Inclusive tree [RFC7117]. When P2MP LSPs are used for flooding unknown unicast traffic, packet reordering 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.

13. Forwarding Unicast Packets

This section describes procedures for forwarding unicast packets by PEs, where such packets are received from either directly connected CEs or some other PEs.

13.1. Forwarding Packets Received from a CE

When a PE receives a packet from a CE, on a given Ethernet Tag ID, it must first look up the source MAC address of the packet. In certain environments that enable MAC security, the source MAC address MAY be used to validate the host identity 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 has learned it from locally connected CEs, the MAC address is considered a known MAC address. Otherwise, it is considered 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 EVPN 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 8.3. If ingress replication is used, the packet MUST be replicated to each remote PE, with the VPN 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 a <MAC-VRF> or <MAC-VRF, Ethernet tag> combination.

  The Ethernet tag in the route may 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 of which the PE is the root, for the Ethernet tag in the EVPN instance. If the same P2MP LSP is used for all Ethernet tags, then all the PEs in the EVPN instance MUST be the leaves of the P2MP LSP. If a distinct P2MP LSP is used for a given Ethernet tag in the EVPN instance, 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.

13.2. Forwarding Packets Received from a Remote PE

This section describes the procedures for forwarding known and unknown unicast packets received from a remote PE.

13.2.1. Unknown Unicast Forwarding

When a 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 EVPN instance or -- in the case of ingress replication -- the downstream label advertised in the P-tunnel attribute, and after performing the split-horizon procedures described in Section 8.3:

- If the PE is the designated forwarder of BUM 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 for BUM traffic 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 a BUM 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.

13.2.2. Known Unicast Forwarding

If the top MPLS label ends up being an EVPN 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.

14. Load Balancing of Unicast Packets

This section specifies the load-balancing procedures for sending known unicast packets to a multihomed CE.

14.1. Load Balancing of Traffic from a PE to Remote CEs

Whenever a remote PE imports a MAC/IP Advertisement route for a given <ESI, Ethernet tag> in a MAC-VRF, it MUST examine all imported Ethernet A-D routes for that ESI in order to determine the load-balancing characteristics of the Ethernet segment.

14.1.1. Single-Active Redundancy Mode

For a given ES, if the remote PE has imported the set of Ethernet A-D per ES routes from at least one PE, where the "Single-Active" flag in the ESI Label extended community is set, then the remote PE MUST deduce that the ES is operating in Single-Active redundancy mode. As such, the MAC address will be reachable only via the PE announcing the associated MAC/IP Advertisement route -- this is referred to as the primary PE. The other PEs advertising the set of Ethernet A-D per ES routes for the same ES provide backup paths for that ES, in case the primary PE encounters a failure, and are referred to as backup PEs. It should be noted that the primary PE for a given <ES, VLAN> (or <ES, VLAN bundle>) is the DF for that <ES, VLAN> (or <ES, VLAN bundle>).

If the primary PE encounters a failure, it MAY withdraw its set of Ethernet A-D per ES routes for the affected ES prior to withdrawing its set of MAC/IP Advertisement routes.

If there is only one backup PE for a given ES, the remote PE MAY use the primary PE's withdrawal of its set of Ethernet A-D per ES routes 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 ES, it will start sending MAC/IP Advertisement routes while the failed PE withdraws its routes. This mechanism minimizes the flooding of traffic during fail-over events.

If there is more than one backup PE for a given ES, the remote PE MUST use the primary PE's withdrawal of its set of Ethernet A-D per ES routes as a trigger to start flooding traffic for the associated MAC addresses (as long as flooding of unknown unicast packets is administratively allowed), as it is not possible to select a single backup PE.

14.1.2. All-Active Redundancy Mode

For a given ES, if the remote PE has imported the set of Ethernet A-D per ES routes from one or more PEs and none of them have the "Single-Active" flag in the ESI Label extended community set, then the remote PE MUST deduce that the ES is operating in All-Active redundancy mode. A remote PE that receives a MAC/IP Advertisement route with a non-reserved ESI SHOULD consider the advertised MAC address to be reachable via all PEs that have advertised reachability to that MAC address's EVI/ES via the combination of an Ethernet A-D per EVI route for that EVI/ES (and Ethernet tag, if applicable) AND an Ethernet A-D per ES route for that ES. The remote PE MUST use

received MAC/IP Advertisement routes and Ethernet A-D per EVI/per ES routes to construct the set of next hops for the advertised MAC address.

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 for that PE, then the next hop and the MPLS label stack are 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 same ESI and Ethernet tag as the ones 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 a set of Ethernet A-D per ES routes for that ES AND an Ethernet A-D route per EVI exist, only then must the label from that latter route 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 source MAC address MAC1 on VLAN1 (mapped to EVI1). 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 set of Ethernet A-D per ES routes for ES1 as well as an Ethernet A-D per EVI route for <EVI1, ES1>.

The MPLS label stack to send the packets to PE1 is the MPLS LSP stack to get to PE1 (at the top of the stack) followed by the EVPN 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 (at the top of the stack) followed by 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 receives the packet destined for CE1 from PE3, if the packet is a known unicast, it is forwarded to CE1. If it is a BUM packet, then only one of PE1 or PE2 must forward the packet to the CE. Whether PE1 or PE2 forwards this packet to the CE is determined based on which of the two is the DF.

14.2. Load Balancing of Traffic between a 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 a LAG. The PE(s) and the CE can load balance traffic onto these interfaces using one of the following mechanisms.

14.2.1. 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 multipathing. 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.

14.2.2. Control-Plane Learning

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

15. 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 multihoming 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/IP 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 EVPN instance 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/IP 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 EVPN instance receive all of these correctly through the intervening BGP infrastructure, introducing a sequence number into the MAC Mobility extended community attribute is necessary.

In order to process mobility events correctly, an implementation MUST handle scenarios in which sequence number wraparound occurs.

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/IP Advertisement route with a different Ethernet segment identifier advertises the MAC address in a MAC/IP Advertisement route tagged with a MAC Mobility extended community attribute with a sequence number one greater than the sequence

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

- A PE detecting a locally attached MAC address for which it had previously received a MAC/IP Advertisement route with the same non-zero Ethernet segment identifier advertises it with:

  1. no MAC Mobility extended community attribute, if the received route did not carry said attribute.

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

- A PE detecting a locally attached MAC address for which it had previously received a MAC/IP Advertisement route with the same zero Ethernet segment identifier (single-homed scenarios) advertises it with a MAC Mobility extended community attribute with the sequence number set properly. In the case of single-homed scenarios, there is no need for ESI comparison. ESI comparison is done for multihoming in order to prevent false detection of MAC moves among the PEs attached to the same multihomed site.

A PE receiving a MAC/IP 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/IP Advertisement route. If two (or more) PEs advertise the same MAC address with the same sequence number but different Ethernet segment identifiers, a PE that receives these routes selects the route advertised by the PE with the lowest IP address as the best route. If the PE is the originator of the MAC route and it receives the same MAC address with the same sequence number that it generated, it will compare its own IP address with the IP address of the remote PE and will select the lowest IP. If its own route is not the best one, it will withdraw the route.

15.1. 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 misconfigured with the same (duplicate) MAC address. In such a 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 a situation and avoid incrementing the sequence number (in the MAC Mobility extended community attribute) to infinity. In order to remedy such a situation, a PE that detects a MAC mobility event via local learning starts an M-second timer (with a default value of M = 180), and if it detects N MAC moves before the timer expires (with a default value of N = 5), it concludes that a duplicate-MAC situation has occurred. The PE MUST alert the operator and stop sending and processing any BGP MAC/IP Advertisement routes for that MAC address until a corrective action is taken by the operator. The values of M and N MUST be configurable to allow for flexibility in operator control. Note that the other PEs in the EVPN instance will forward the traffic for the duplicate MAC address to one of the PEs advertising the duplicate MAC address.

15.2. Sticky MAC Addresses

There are scenarios in which it is desired to configure some MAC addresses as static so that they are not subjected to MAC moves. In such scenarios, these MAC addresses are advertised with a MAC Mobility extended community where the static flag is set to 1 and the sequence number is set to zero. If a PE receives such advertisements and later learns the same MAC address(es) via local learning, then the PE MUST alert the operator.

16. Multicast and Broadcast

The PEs in a particular EVPN instance may use ingress replication or P2MP LSPs to send multicast traffic to other PEs.

16.1. Ingress Replication

The PEs may use ingress replication for flooding BUM traffic as described in Section 11 ("Handling of Multi-destination Traffic"). A given 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 per [RFC7117].

16.2. P2MP LSPs

A PE may use an "Inclusive" tree for sending a BUM packet. This terminology is borrowed from [RFC7117].

A variety of transport technologies may be used in the service provider (SP) network. For Inclusive P-multicast trees, these transport technologies include point-to-multipoint LSPs created by RSVP-TE or Multipoint LDP (mLDP).

16.2.1. 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 EVPN instances on a given PE. A particular P-multicast tree can be set up to carry the traffic originated by sites belonging to a single EVPN instance, or to carry the traffic originated by sites belonging to several EVPN instances. The ability to carry the traffic of more than one EVPN instance on the same tree is termed ’Aggregation’, and the tree is called an Aggregate Inclusive P-multicast tree or Aggregate Inclusive tree for short. The Aggregate Inclusive tree needs to include every PE that is a member of any of the EVPN instances that are using the tree. This implies that a PE may receive BUM traffic even if it doesn’t have any receivers that are interested in receiving that traffic.

An Inclusive or Aggregate Inclusive tree as defined in this document is a P2MP tree. A P2MP tree is used to carry traffic only for EVPN 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 [RFC7117], with the VPLS A-D route replaced with the Inclusive Multicast Ethernet Tag route. The P-tunnel attribute [RFC7117] for an Inclusive tree is advertised with the Inclusive Multicast Ethernet Tag route as described in Section 11 ("Handling of Multi-destination Traffic"). Note that for an Aggregate Inclusive tree, a PE can "aggregate" multiple EVPN instances on the same P2MP LSP using upstream labels. The procedures for aggregation are the same as those described in [RFC7117], with VPLS A-D routes replaced by EVPN Inclusive Multicast Ethernet Tag routes.

17. Convergence

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

17.1. Transit Link and Node Failures between PEs

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

17.2. PE Failures

Consider a host CE1 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 Bidirectional Forwarding Detection (BFD) is used to detect BGP session failures. PE3 can update its forwarding state to start sending all traffic for CE1 to only PE2.

17.3. PE-to-CE Network Failures

If the connectivity between the multihomed CE and one of the PEs to which it is attached fails, the PE MUST withdraw the set of Ethernet A-D per ES routes that had been previously advertised for that ES. 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. When the MAC entry on the PE ages out, the PE MUST withdraw the MAC address from BGP.

When an Ethernet tag is decommissioned on an Ethernet segment, then the PE MUST withdraw the Ethernet A-D per EVI route(s) announced for the <ESI, Ethernet tags> that are impacted by the decommissioning. In addition, the PE MUST also withdraw the MAC/IP Advertisement routes that are impacted by the decommissioning.

The Ethernet A-D per ES routes should be used by an implementation to optimize the withdrawal of MAC/IP Advertisement routes. When a PE receives a withdrawal of a particular Ethernet A-D route from an advertising PE, it SHOULD consider all the MAC/IP Advertisement routes that are learned from the same ESI 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.

18. Frame Ordering

In a MAC address, if the value of the first nibble (bits 8 through 5) of the most significant octet of the destination MAC address (which follows the last MPLS label) happens to be 0x4 or 0x6, then the Ethernet frame can be misinterpreted as an IPv4 or IPv6 packet by intermediate P nodes performing ECMP based on deep packet inspection, thus resulting in load balancing packets belonging to the same flow on different ECMP paths and subjecting those packets to different delays. Therefore, packets belonging to the same flow can arrive at the destination out of order. This out-of-order delivery can happen during steady state in the absence of any failures, resulting in significant impact on network operations.

In order to avoid any such misordering, the following rules are applied:

- If a network uses deep packet inspection for its ECMP, then the "Preferred PW MPLS Control Word" [RFC4385] SHOULD be used with the value 0 (e.g., a 4-octet field with a value of zero) when sending EVPN-encapsulated packets over an MP2P LSP.

- If a network uses entropy labels [RFC6790], then the control word SHOULD NOT be used when sending EVPN-encapsulated packets over an MP2P LSP.

- When sending EVPN-encapsulated packets over a P2MP LSP or P2P LSP, then the control word SHOULD NOT be used.

19. Security Considerations

Security considerations discussed in [RFC4761] and [RFC4762] apply to this document for MAC learning in the data plane over an Attachment Circuit (AC) and for flooding of unknown unicast and ARP messages over the MPLS/IP core. Security considerations discussed in [RFC4364] apply to this document for MAC learning in the control plane over the MPLS/IP core. This section describes additional considerations.

As mentioned in [RFC4761], there are two aspects to achieving data privacy and protecting against denial-of-service attacks in a VPN: securing the control plane and protecting the forwarding path. Compromise of the control plane could result in a PE sending customer data belonging to some EVPN to another EVPN, or black-holing EVPN customer data, or even sending it to an eavesdropper, none of which are acceptable from a data privacy point of view. In addition, compromise of the control plane could provide opportunities for

unauthorized EVPN data usage (e.g., exploiting traffic replication within a multicast tree to amplify a denial-of-service attack based on sending large amounts of traffic).

The mechanisms in this document use BGP for the control plane. Hence, techniques such as those discussed in [RFC5925] help authenticate BGP messages, making it harder to spoof updates (which can be used to divert EVPN traffic to the wrong EVPN instance) or withdrawals (denial-of-service attacks). In the multi-AS backbone options (b) and (c) [RFC4364], this also means protecting the inter-AS BGP sessions between the Autonomous System Border Routers (ASBRs), the PEs, or the Route Reflectors.

Further discussion of security considerations for BGP may be found in the BGP specification itself [RFC4271] and in the security analysis for BGP [RFC4272]. The original discussion of the use of the TCP MD5 signature option to protect BGP sessions is found in [RFC5925], while [RFC6952] includes an analysis of BGP keying and authentication issues.

Note that [RFC5925] will not help in keeping MPLS labels private -- knowing the labels, one can eavesdrop on EVPN traffic. Such eavesdropping additionally requires access to the data path within an SP network. Users of VPN services are expected to take appropriate precautions (such as encryption) to protect the data exchanged over a VPN.

One of the requirements for protecting the data plane is that the MPLS labels be accepted only from valid interfaces. For a PE, valid interfaces comprise links from other routers in the PE’s own AS. For an ASBR, valid interfaces comprise links from other routers in the ASBR’s own AS, and links from other ASBRs in ASes that have instances of a given EVPN. It is especially important in the case of multi-AS EVPN instances that one accept EVPN packets only from valid interfaces.

It is also important to help limit malicious traffic into a network for an impostor MAC address. The mechanism described in Section 15.1 shows how duplicate MAC addresses can be detected and continuous false MAC mobility can be prevented. The mechanism described in Section 15.2 shows how MAC addresses can be pinned to a given Ethernet segment, such that if they appear behind any other Ethernet segments, the traffic for those MAC addresses can be prevented from entering the EVPN network from the other Ethernet segments.

20. IANA Considerations

This document defines a new NLRI, called "EVPN", to be carried in BGP using multiprotocol extensions. This NLRI uses the existing AFI of 25 (L2VPN). IANA has assigned BGP EVPNs a SAFI value of 70.

IANA has allocated the following EVPN Extended Community sub-types in [RFC7153], and this document is the only reference for them.

0x00	MAC Mobility	[RFC7432]
0x01	ESI Label	[RFC7432]
0x02	ES-Import Route Target	[RFC7432]

This document creates a registry called "EVPN Route Types". New registrations will be made through the "RFC Required" procedure defined in [RFC5226]. The registry has a maximum value of 255. Initial registrations are as follows:

0	Reserved	[RFC7432]
1	Ethernet Auto-discovery	[RFC7432]
2	MAC/IP Advertisement	[RFC7432]
3	Inclusive Multicast Ethernet Tag	[RFC7432]
4	Ethernet Segment	[RFC7432]

21. References

21.1. Normative References

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

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007, <http://www.rfc-editor.org/info/rfc4760>.

[RFC4761] Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007, <http://www.rfc-editor.org/info/rfc4761>.

[RFC4762] Lasserre, M., Ed., and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007, <http://www.rfc-editor.org/info/rfc4762>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, March 2014, <http://www.rfc-editor.org/info/rfc7153>.

21.2. Informative References

[802.1D-REV] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges", IEEE Std. 802.1D, June 2004.

[802.1Q] "IEEE Standard for Local and metropolitan area networks - Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q(tm), 2014 Edition, November 2014.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006, <http://www.rfc-editor.org/info/rfc4272>.

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

[RFC4664] Andersson, L., Ed., and E. Rosen, Ed., "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006, <http://www.rfc-editor.org/info/rfc4664>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010, <http://www.rfc-editor.org/info/rfc5925>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012, <http://www.rfc-editor.org/info/rfc6514>.

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

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, May 2013, <http://www.rfc-editor.org/info/rfc6952>.

[RFC7117] Aggarwal, R., Ed., Kamite, Y., Fang, L., Rekhter, Y., and C. Kodeboniya, "Multicast in Virtual Private LAN Service (VPLS)", RFC 7117, February 2014, <http://www.rfc-editor.org/info/rfc7117>.

[RFC7209] Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, N., Henderickx, W., and A. Isaac, "Requirements for Ethernet VPN (EVPN)", RFC 7209, May 2014, <http://www.rfc-editor.org/info/rfc7209>.

Acknowledgements

Special thanks to Yakov Rekhter for reviewing this document several times and providing valuable comments, and for his very engaging discussions on several topics of this document that helped shape this document. We would also like to thank 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. We would like to thank Jorge Rabadan for his contribution to Section 5 of this document. We would like to thank Thomas Morin for his review of this document and his contribution of Section 8.6. Many thanks to Jakob Heitz for his help to improve several sections of this document.

We would also like to thank Clarence Filsfils, Dennis Cai, Quaizar Vohra, Kireeti Kompella, and Apurva Mehta for their contributions to this document.

Last but not least, special thanks to Giles Heron (our WG chair) for his detailed review of this document in preparation for WG Last Call and for making many valuable suggestions.

Contributors

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

Keyur Patel
Samer Salam
Sami Boutros
Cisco

Yakov Rekhter
Ravi Shekhar
Juniper Networks

Florin Balus
Nuage Networks

Authors' Addresses

Ali Sajassi (editor)
Cisco
EMail: sajassi@cisco.com

Rahul Aggarwal
Arktan
EMail: raggarwa_1@yahoo.com

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

Aldrin Isaac
Bloomberg
EMail: aisaac71@bloomberg.net

James Uttaro
AT&T
EMail: uttaro@att.com

John Drake
Juniper Networks
EMail: jdrake@juniper.net

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