Network Working Group
Internet Draft
Category: Standards Track
Expiration Date: September 2010

R. Aggarwal (Editor)
Juniper Networks

A. Isaac
Bloomberg

J. Uttaro
AT&T

R. Shekhar
Juniper Networks

March 26, 2010

BGP MPLS Based MAC VPN

draft-raggarwa-mac-vpn-00.txt

Status of this Memo

Copyright and License Notice

Abstract

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

Table of Contents

1. Specification of requirements

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

2. Contributors

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

Quaizar Vohra
Kireeti Kompella
Apurva Mehta
Juniper Networks

3. Introduction

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

There is a desire by Service Providers (SP) and data center providers to provide MPLS based bridged / LAN services or/and infrastructure such that they meet the requirements listed below. An example of such a service is a VPLS service offered by a SP. Another example is a MPLS based Virtual Bridged Network infrastructure in a data center. Here are the requirements:

- Minimal or no configuration required. MPLS implementations have reduced the amount of configuration over the years. There is a need for greater auto-configuration.

- Support of multiple active points of attachment for CEs which may be hosts, switches or routers. Current MPLS technologies such as VPLS, currently do not support this. This allows load-balancing among multiple paths active. MPLS technologies such as VPLS currently do not allow the same MAC to be learned from two different PEs and be active at the same time.

- Ability to span a VLAN across multiple racks in different geographic locations.

- Minimize or eliminate flooding of unknown unicast traffic.

- Allow hosts and Virtual Machines (VMs) in a data center to relocate without requiring renumbering. For instnace VMs may be moved for load or failure reasons.

- Ability to scale up to hundreds of thousands of hosts across multiple data centers, where connectivity is required between hosts in different data centers.

- Support for virtualization. This includes the ability to separate hosts and VMs working together from other such groups, and the ability to have overlapping IP and MAC addresses/

- Fast convergence

This document proposes a MPLS based technology, referred to as MPLS-based MAC VPN (MAC-VPN) for meeting the requirements described in this section. MAC-VPN requires extensions to existing IP/MPLS protocols as described in the next section. In addition to these extensions MAC-VPN uses several building blocks from existing MPLS technologies.

4. BGP MPLS Based MAC-VPN

This section describes the framework of MAC-VPN to meet the requirements described in the previous section.

An MAC-VPN comprises CEs that are connected to PEs or MPLS Edge Switches (MES) that comprise the edge of the MPLS infrastructure. A CE may be a host, a router or a switch. The MPLS Edge Switches provide layer 2 virtual bridge connectivity between the CEs. There may be multiple MAC VPNs in the provider's network. This document uses the terms MAC-VPN, MAC VPN inter-changeably. The instance of a MAC VPN on an MES is referred to as a MAC VPN Instance (MVI).

The MESes are connected by a MPLS LSP infrastructure which provides the benefits of MPLS such as fast-reroute, resiliency etc.

In a MAC VPN, learning between MESes occurs not in the data plane (as happens with traditional bridging) but in the control plane. Control plane learning offers much greater control over the learning process, such as restricting who learns what, and the ability to apply policies. Furthermore, the control plane chosen for this is BGP (very similar to IP VPNs (RFC 4364)), providing much greater scale, and the ability to "virtualize" or isolate groups of interacting agents (hosts, servers, Virtual Machines) from each other. In MAC VPNs MESes advertise the MAC addresses learned from the CEs that are connected to them, along with a MPLS label, to other MESes in the

control plane. Control plane learning enables load balancing and allows CEs to connect to multiple active points of attachment. It also improves convergence times in the event of certain network failures.

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

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

The policy attributes of a MAC VPN are very similar to an IP VPN. A MAC-VPN instance requires a Route-Distinguisher (RD) and a MAC-VPN requires one or more Route-Targets (RTs). A CE attaches to a MAC-VPN on a MES in a particular MVI on a VLAN or simply an ethernet interface. When the point of attachment is a VLAN there may be one or more VLANs in a particular MAC-VPN. Some deployment scenarios guarantee uniqueness of VLANs across MAC-VPNs: all points of attachment of a given MAC VPN use the same VLAN, and no other MAC VPN uses this VLAN. This document refers to this case as a "Single VLAN MAC-VPN" and describes simplified procedures to optimize for it.

The next section discusses how layer 2 connectivity is achieved between the CEs.

Section 8 describes how load balancing and link bonding is achieved for MAC-VPN. Section 10 describes procedures for handling MAC moves.

5. Ethernet Segment Identifier

If a CE is multi-homed to two or more MESes, the set of attachment circuits constitutes an "Ethernet segment". An Ethernet segment may appear to the CE as a Link Aggregation Group. Ethernet segments have an identifier, called the "Ethernet Segment Identifier" (ESI). A single-homed CE is considered to be attached to a Ethernet segment with ESI 0. Otherwise, an Ethernet segment MUST have a unique non-zero ESI. The ESI can be assigned using various mechanisms:

1. The ESI may be configured. For instance when MAC VPNs are used to provide a VPLS service the ESI is fairly analogous to the VE ID used for the procedures in [BGP-VPLS].

2. If LACP is used, between the MESes and CEs that are hosts, then the ESI is determined by LACP. This is the 48 bit virtual MAC address of the host for the LACP link bundle. As far as the host is concerned it would treat the multiple MESes that it is homed to as the same switch. This allows the host to aggregate links to different MESes in the same bundle.

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

4. In the case of indirectly connected hosts and a bridged LAN between the hosts and the MESes, the ESI is determined based on the Layer 2 bridge protocol as follows:

If STP is used then the value of the ESI is derived by listening to BPDUs on the ethernet segment. The MES does not run STP. However it does learn the Switch ID, MSTP ID and Root Bridge ID by listening to BPDUs. The ESI is as follows:

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

6. Determining Reachability to Unicast MAC Addresses

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

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

6.1. Local Learning

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

The MESes in a particular MAC-VPN MUST support local data plane learning using vanilla ethernet learning procedures. A MES must be capable of learning MAC addresses in the data plane when it receives packets such as the following from the CE network:

- DHCP requests

- gratuitous ARP request for its own MAC.

- ARP request for a peer.

Alternatively if a CE is a host then MESes MAY learn the MAC addresses of the host in the control plane using extensions to protocols such as LLDP that run between the MES and the hosts.

In the case where a CE is a host or a switched network connected to hosts, the MAC address is reachable via a given MES may move such that it becomes reachable via another MES. This is referred to as a "MAC Move". Procedures to support this are described in section 10.

6.2. Remote learning

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

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

6.2.1. BGP MAC-VPN MAC Address Advertisement

BGP is extended to advertise these MAC addresses using a new NLRI called the MAC-VPN-NLRI, with AFI (TBD) and a new SAFI of MAC-VPN (TBD).

The MAC-VPN-NLRI encodes the following elements when it is used for advertising MAC addresses:

a) Route-Distinguisher of the MAC-VPN instance that is advertising the NLRI. A RD MUST be assigned for a given MAC-VPN instance on a MES. This RD MUST be unique across all MAC-VPN instances on a MES. This can be accomplished by using a Type 1 RD [RFC4364]. The value field comprises an IP address of the MES (typically, the loopback address) followed by a number unique to the MES. This number may be generated by the MES, or, in the Single VLAN MAC-VPN case, may be the 12 bit VLAN ID, with the remaining 4 bits set to 0.

b) VLAN ID if the MAC address is learned over a VLAN from the CE, else this field is set to 0.

c) The Ethernet Segment Identifier described in the previous section.

d) The MAC address.

e) The advertisement may optionally carry one of the IP addresses associated with the MAC address. If used, this aids the implemntation of proxy ARP on MESes thereby reducing the flooding of broadcast packets.

f) A MAC-VPN MPLS label that is used by the MES to forward packets received from remote MESes. The forwarding procedures are specified in section 8. A MES MAY advertise the same MAC-VPN label for all MAC addresses in a given MAC-VPN instance. This label assignment methodology is referred to as a per MVI label assigment. Or a MES may advertise a unique MAC-VPN label per MAC address. Both of these methodologies have their tradeoffs. Per MVI label assignment requires the least number of MAC-VPN labels, but requires a MAC lookup in addition to a MPLS lookup on an egress MES for forwarding. On the other hand a unique label per MAC allows an egress MES to forward a packet that it receives from another MES, to the connected CE, after looking up only the MPLS labels and not having to do a MAC lookup.

The BGP advertisement also carries the following attributes:

a) One or more Route Target (RT) attributes MUST be carried.

RTs may be configured (as in IP VPNs), or may be derived automatically from the VLAN ID associated with the advertisement.

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

+ The Global Administrator field of the RT MUST be set to an IP address of the MES. This address SHOULD be common for all the MAC-VPN instances on the MES (e.,g., this address may be the MES's loopback address).

+ The Local Administrator field of the RT contains a 2 octets long number that encodes the VLAN-ID.

The above auto-configuration of the RT implies that a different RT is used for every VLAN in a MAC-VPN, if the MAC-VPN contains multiple VLANs. For the "Single VLAN MAC-VPN" this results in auto-deriving the RT from the VLAN for that MAC-VPN.

b) The advertisement may optionally carry the IP addresses associated with the MAC address, if the number of IP addresses is more than one and cannot be encoded in the NLRI. This aids the implemntation of proxy ARP on MESes thereby reducing the flooding of broadcast packets.

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

7. Designated Forwarder Election

If a CE that is a host or a router is multi-homed directly to more than one MES in a MAC-VPN, only one of the MESes is responsible for certain actions:

- Sending multicast and broadcast traffic to the CE. Note that this is the default behavior. Optional mechanisms, which will be specified later, will allow load balancing of multicast and broadcast traffic from MESes to CEs on multiple active points of attachment

- Flooding unknown unicast traffic (i.e. traffic for which a MES does not know the destination MAC address) to the CE, if the environment requires flooding of unknown unicast traffic.

Note that a CE always sends packets using a single link. For instance if the CE is a host then, as mentioned earlier, the host treats the multiple links that it uses to reach the MESes as a LAG or a bundle.

If a bridge network is multi-homed to more than one MES in a MAC-VPN via switches, only one of the MESes is responsible for certain actions:

+ - Forwarding packets to other MESes, out of the bridged LAN which is multi-homed to more than one MES. This is the case when the MAC-VPN cloud is inter-connecting bridged LAN islands. There are certain cases where this may not be the case. For instance this is not required if the topology is loop free.

+ - Sending multicast and broadcast traffic to the bridge network. Note that this is the default behavior. Optional mechanisms, which will be specified later, will allow load balancing of multicast and broadcast traffic from MESes on CEs on multiple active points of attachment

+ - Flooding unknown unicast traffic (i.e. traffic for which a MES does not know the destination MAC address) to the bridge network.

This particular MES is referred to as the designated forwarder (DF) MES, for the ethernet segment over which the host is multi-homed to two or more MESes. This ethernet segment may be a link bundle if the host or router is directly connected to the MESes. Or this ethernet segment may be a bridged LAN network, if the CEs are switches. The bridged LAN network may be running a protocol such as STP. The granularity of the DF election MUST be at least this ethernet segment. In this case the same MES MUST be elected as the DF for all CEs on the ethernet segment. The granularity of the DF election MAY be the combination of the ethernet segment and VLAN on that ethernet segment. In this case the same MES MUST be elected as the DF for all hosts on a VLAN on that ethernet segment.

The MESes perform a designated forwarder (DF) election, for an ethernet segment, or ethernet segment, vlan combination using BGP. The Ethernet Segment Identifier is assigned as described in section 4.

In order to perform DF election each MES advertises in BGP, a DF election route, using the MAC-VPN-NLRI, for each ethernet segment in a MAC-VPN. This route contains the following information elements

a) Route-Distinguisher of the MAC-VPN instance that is advertising the NLRI. This RD is the same as the one used in section 5.2.1.

b) Ethernet Segment Identifier

c) Optional VLAN ID which may be set to 0.

d) An upstream assigned MPLS label referred to as the "ESI label". The usage of this label is described in section 8.

This route also carries the following BGP attributes:

- P-Tunnel attribute which is specified in [VPLS-MCAST]. The usage of this attribute is described in section 11.

- One or more Route Target (RT) attributes. These RTs are assigned using the same procedure as the one described in section 5.

The DF election for a particular ESI and VLAN combination proceeds as follows. First a MES constructs a candidate list of MESes. This comprises all the DF routes with that particular {ESI, VLAN} tuple that a MES imports in a MAC-VPN instance, including the DF route generated by the MES itself, if any. The DF MES is chosen from this candidate list. Note that DF election is carried out by all the MESes that import the DF route.

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

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

8. Forwarding Unicast Packets

8.1. Processing of Unknown Unicast Packets

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

However, if a destination MAC address of a received packet is not known by the MES, the MES may have to flood the packet. Flooding must take into account "split horizon forwarding" as follows. The principles behind the following procedures are borrowed from the split horizon forwarding rules in VPLS solutions [RFC 4761, RFC 4762]. When a MES capable of flooding (say MESx) receives a broadcast Ethernet frame, or one with an unknown destination MAC address, it must flood the frame. f the frame arrived from an attached CE, MESx must send a copy of the frame to every other attached CE, as well as to all other MESs participating in the MAC VPN. If, on the other hand, the frame arrived from another MES (say MESy), MESx must send a copy of the packet only to attached CEs. MESx MUST NOT send the frame to other MESs, since MESy would have already done so. Split horizon forwarding rules apply to broadcast and multicast packets, as well as packets to an unknown MAC address.

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

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

If ingress replication is in use, the P-Tunnel attribute, carried in the DF routes for the MAC VPN, specifies the downstream label that the other MESes can use to send unknown unicast, multicast or broadcast traffic for the MAC VPN to this particular MES. Note that if a MES has multiple ethernet segments for the same MAC-VPN instance and ingress replication is in use, then the MES SHOULD advertise the same P-Tunnel attribute for each DF route for that MAC-VPN instance.

The procedures for using P2MP LSPs are very similar to VPLS procedures [VPLS-MCAST]. The P-Tunnel attribute used by a MES for sending unknown unicast, broadcast or multicast traffic for a particular ethernet segment, is advertised in the DF route as described in section 6. Note that if a MES has multiple ethernet segments for the same MAC-VPN instance then it SHOULD advertise the same P-Tunnel attribute for each DF route for that MAC-VPN instance. The P-Tunnel attribute specifies the P2MP LSP identifier. This is the equivalent of an Inclusive tree in [VPLS-MCAST]. Note that multiple MAC-VPNs can use the same P2MP LSP, using upstream labels [VPLS-MCAST]. When P2MP LSPs are used for flooding unknown unicast traffic, packet re-ordering is possible.

8.2. Forwarding packets received from a CE

When a MES receives a packet from a CE it must first look up the source MAC address of the packet. In certain environments the source MAC address may be used to authenticate the CE and determine that traffic from the host can be allowed into the network.

If the MES decides to forward the packet the destination MAC address of the packet must be looked up. If the MES has received MAC address advertisements from one or more other MESes, for this destination MAC address, it is considered as a known MAC address. Else the MAC address is considered as an unknown MAC address.

For known MAC addresses the MES forwards this packet to one of the remote MESes. The packet is encapsulated in the MAC-VPN MPLS label advertised by the remote MES, for that MAC address, and in the MPLS

LSP label stack to reach the remote MES.

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

- The MES floods the packet to other MESes. The MES first encapsulates the packet in the ESI MPLS label as described in section 4. If P2MP LSPs are being used the packet is sent on the P2MP LSP that the MES is the root of for that MAC-VPN, with all the other MESes as the leaves. The packet is encapsulated in the P2MP LSP label stack. If ingress replication is used the packet is replicated once for each remote MES with the bottom label of the stack being the MPLS label advertised by the remote MES in the P-Tunnel attribute.

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

- The MES drops the packet.

8.3. Forwarding packets received from a remote MES

When a MES receives a MPLS packet from a remote MES then, after processing the MPLS label stack, if the top MPLS label ends up being a P2MP LSP label associated with a MAC-VPN or the downstream label advertised in the P-Tunnel attribute and after performing the split horizon procedures described in section 8:

- If the MES is the designated forwarder of unknown unicast, broadcast or multicast traffic, the default behavior is for the MES to flood the packet to all the host interfaces. In other words the default behavior is for the MES to assume that the destination MAC address is unknown unicast, broadcast or multicast and it is not required to do a destination MAC address lookup. As an option the MES may do a destination MAC lookup to flood the packet to only a subset of the host interfaces.
- If the MES is not the designated forwarder, the default behavior is for it to drop the packet.

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

9. Split Horizon

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

In order to accomplish this each MES distributes to other MESes an "ESI MPLS label" in the DF route as described in section 6. This label is upstream assigned by the MES that advertises the DF route. This label MUST be programmed by the other MESes, that are connected to the ESI advertised in the route, in the context label space for the advertising MES. Further the forwarding entry for this label must result in discarding packets received with this label.

Further the MES that advertises the "ESI MPLS label" MUST program in its platform MPLS forwarding table a forwarding entry for this label which results in sending packets to the ESI.

Consider MES1 and MES2 that are multi-homed to CE1 on ES1. When MES1 sends a multicast, broadcast or unknown unicast packet, that it receives from CE1, it MUST first push onto the MPLS label stack the ESI label that it has assigned for the ESI. The resulting packet is further encapsulated in the MPLS label stack necessary to transmit the packet to the other MESes. Penultimate hop popping MUST be disabled on the P2P or P2MP LSPs used in the MPLS transport infrastructure for MAC VPN. When MES2 receives this packet it decapsulates the top MPLS label and forwards the packet using the context label space determined by the top label. If the next label is the ESI label assigned by MES1 then MES2 must drop the packet.

10. Load Balancing of Unicast Packets

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

10.1. Load balancing of traffic from a MES to remote CEs

Whenever a remote MES imports a MAC advertisement for a given ESI, in a MAC VPN instance, it MUST consider the MAC as reachahable via all the MESes from which it has imported DF routes for that ESI.

Consider a CE, CE1, that is dual homed to two MESes, MES1 and MES2 on a LAG interface, ES1, and is sending packets with MAC address MAC1. Based on MAC-VPN extensions described in sections 5 and 6, a remote MES say MES3 is able to learn that a MAC1 is reachable via MES1 and MES2. Both MES1 and MES2 may advertise MAC1 in BGP if they receive packets with MAC1 from CE1. If this is not the case and if MAC1 is advertised only by MES1, MES3 still considers MAC1 as reachable via both MES1 and MES2 as both MES1 and MES2 advertise a DF route for ES1.

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

The MPLS label stack to send packets to MES2 is the MPLS LSP stack to get to MES2 and the upstream assigned label in the DF route advertised by MES2 for ES1, if MES2 has not advertised MAC1 in BGP.

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

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

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

When MES1 or MES2 receive the packet destined for CE1 from MES3, if the packet is a unicast MAC packet it is forwarded to CE1. If it is a multicast or broadcast MAC packet then only one of MES1 or MES2 must forward the packet to the CE. Which of MES1 or MES2 forward this packet to the CE is determined by default based on which of the two is the DF. An alternate procedure to load balance multicast packets will be described in the future.

If the connectivity between the multi-homed CE and one of the MESes

that it is multi-homed to fails, the MES MUST withdraw the MAC address from BGP. This enables the remote MESes to remove the MPLS next-hop to this particular MES from the set of MPLS next-hops that can be used to forward traffic to the CE.

Load balancing requires that the MESes that the CE is multi-homed to are configured with different Route-Distinguishers (RDs).

10.2. Load balancing of traffic between a MES and a local CE

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

10.2.1. Data plane learning

Consider that the MESes perform data plane learning for local MAC addresses learned from local CEs. This enables the MES(s) to learn a particular MAC address and associate it with one or more interfaces. The MESes can now load balance traffic destined to that MAC address on the multiple interfaces.

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

10.2.2. Control plane learning

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

11. MAC Moves

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

Remote MESes must be able to distinguish a MAC move from the case where a MAC address on an ESI is reachable via two different MESes and load balancing is performed as described in section 9. This distinction can be made as follows. If a MAC is learned by a particular MES from multiple MESes, then the MES performs load balancing only amongst the set of MESes that advertised the MAC with the same ESI. If this is not the case then the MES chooses only one of the advertising MESes to reach the MAC as per BGP path selection.

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

This specification requires that when MES1 learns MAC1 from MES2, and MAC1 as learned by MES1 from the local CE, is not on the same ethernet segment as the one associated with MAC1 by MES2, MES1 must withdraw its own MAC address advertisement from BGP. Further if MES1 receives traffic destined for MAC1 it must send the traffic to MES2. This procedure reduces the duration of traffic loss associated with MAC moves.

12. Multicast

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

12.1. Ingress Replication

The MESes may use ingress replication for flooding unknown unicast, multicast or broadcast traffic as described in section 7.1. A given unknown unicast or broadcast packet must be sent to all the remote MESes. However a given multicast packet for a multicast flow may be sent to only a subset of the MESes. Specifically a given multicast flow may be sent to only those MESes that have receivers that are interested in the multicast flow. Determining which of the MESes have

receivers for a given multicast flow is done using explicit tracking described below.

12.2. P2MP LSPs

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

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

12.2.1. 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 MAC VPN instances on a given MES. A particular P-Multicast tree can be set up to carry the traffic originated by sites belonging to a single MAC VPN, or to carry the traffic originated by sites belonging to different MAC VPNs. The ability to carry the traffic of more than one MAC VPN on the same tree is termed 'Aggregation'. The tree needs to include every MES that is a member of any of the MAC VPNs that are using the tree. This implies that a MES may receive multicast traffic for a multicast stream even if it doesn't have any receivers that are interested in receiving traffic for that stream.

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

The procedures for signaling an Inclusive Tree are the same as those in [VPLS-MCAST] with the VPLS-AD route replaced with the DF route. The P-Tunnel attribute [VPLS-MCAST] for an Inclusive tree is advertised in the DF route as described in section 5. Note that a MES can "aggregate" multiple inclusive trees for different MAC-VPNs on the same P2MP LSP using upstream labels. The procedures for aggregation are the same as those described in [VPLS-MCAST], with VPLS A-D routes replaced by MAC-VPN DF routes.

12.2.2. Selective Trees

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

A SP can use both Inclusive P-Multicast trees and Selective P-Multicast trees or either of them for a given MAC VPN on a MES, based on local configuration.

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

A MAC-VPN MES advertises a selective tree using a MAC-VPN selective A-D route. The procedures are the same as those in [VPLS-MCAST] with S-PMSI A-D routes in [VPLS-MCAST] replaced by MAC-VPN selective A-D routes. The information elements of the MAC VPN selective A-D route are the same as those of the VPLS S-PMSI A-D route with the following difference. A MAC VPN selective A-D route may encode a MAC address in the Group field. The encoding details of the MAC VPN selective A-D route will be described in the next revision.

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

12.3. Explicit Tracking

[VPLS-MCAST] describes procedures for explicit tracking that rely on Leaf A-D routes. The same procedures are used for explicit tracking in this specification with VPLS Leaf A-D routes replaced with MAC-VPN Leaf A-D routes. These procedures allow a root MES to request multicast membership information for a given (S, G), from leaf MESs. Leaf MESs rely on IGMP snooping or PIM snooping between the MES and the CE to determine the multicast membership information. Note that the procedures in [VPLS-MCAST] do not describe how explicit tracking

is performed if the CEs are enabled with join suppression. The procedures for this case will be described in a future version.

13. Convergence

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

13.1. Transit Link and Node Failures between MESes

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

13.2. MES Failures

Consider a host host1 that is dual homed to MES1 and MES2. If MES1 fails, a remote MES, MES3, can discover this based on the failure of the BGP session. This failure detection can be in the sub-second range if BFD is used to detect BGP session failure. MES3 can update its forwarding state to start sending all traffic for host1 to only MES2. It is to be noted that this failure recovery is potentially faster than what would be possible if data plane learning were to be used. As in that case MES3 would have to rely on re-learning of MAC addresses via MES2.

13.2.1. Local Repair

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

13.3. MES to CE Network Failures

Deatils will be described in the future.

14. Acknowledgements

We would like to thank Yakov Rekhter, Kaushik Ghosh, Nischal Sheth and Amit Shukla for discussions that helped shape this document. We would also like to thank Han Nguyen for his comments and support of this work.

15. References

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

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

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

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

[VPLS-MULTIHOMING] "Multi-homing in BGP-based Virtual Private LAN Service", K. Kompella et.al., draft-kompella-l2vpn-vpls-multihoming-01.txt

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

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

16. Author's Address

Rahul Aggarwal
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US

Email: rahul@juniper.net

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

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

Ravi Shekhar
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 US