Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 31, 2014

P. Dutta
F. Balus
Alcatel-Lucent
O. Stokes
Extreme Networks
G. Calvinac
France Telecom

November 27, 2013

LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS
draft-ietf-l2vpn-vpls-ldp-mac-opt-09

Abstract

RFC4762 describes a mechanism to remove or unlearn MAC addresses that have been dynamically learned in a VPLS Instance for faster convergence on topology change. The procedure also removes MAC addresses in the VPLS that do not require relearning due to such topology change. This document defines an enhancement to the MAC Address Withdrawal procedure with empty MAC List from RFC4762, which enables a Provider Edge(PE) device to remove only the MAC addresses that need to be relearned. Additional extensions to RFC4762 MAC Withdrawal procedures are specified to provide optimized MAC flushing for the PBB-VPLS specified in working group RFC7041.

Requirements Language

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

Status of this Memo

Copyright Notice

Table of Contents

1. Terminology

This document uses the terminology defined in [RFC7041], [RFC5036], [RFC4447] and [RFC4762].

Throughout this document VPLS means the emulated bridged LAN service offered to a customer. H-VPLS means the hierarchical connectivity or layout of MTU-s and PE-rs devices offering the VPLS [RFC4762].

The terms "Spoke Node" and "MTU-s" in H-VPLS are used interchangeably.

"Spoke PW" means the PW (Pseudowire) that provides connectivity between MTU-s and PE-rs nodes.

"Mesh PW" means the PW that provides connectivity between PE-rs nodes in a VPLS full mesh core.

"MAC Flush Message" means LDP Address Withdraw Message without MAC List TLV.

MAC Flush Message in the "context of a PW" means the Message that has been received over the LDP session that is used to set up the PW used to provide connectivity in VPLS. The MAC Flush Message carries the context of the PW in terms of FEC TLV associated with the PW [RFC4762][RFC4447].

In general, "MAC Flush" means the method of initiating and processing of MAC Flush Messages across a VPLS instance.

2. Introduction

A method of Virtual Private LAN Service (VPLS), also known as Transparent LAN Service (TLS) is described in [RFC4762]. A VPLS is created using a collection of one or more point-to-point pseudowires (PWs) [RFC4664] configured in a flat, full-mesh topology. The mesh topology provides a LAN segment or broadcast domain that is fully capable of learning and forwarding on Ethernet MAC addresses at the PE devices.

This VPLS full mesh core configuration can be augmented with additional non-meshed spoke nodes to provide a Hierarchical VPLS (H-VPLS) service [RFC4762]. Throughout this document this configuration is referred to as "regular" H-VPLS.

[RFC7041] describes how Provider Backbone Bridging (PBB) can be integrated with VPLS to allow for useful PBB capabilities while

continuing to avoid the use of MSTP in the backbone. The combined solution referred to as PBB-VPLS results in better scalability in terms of number of service instances, PWs and C-MAC (Customer MAC) Addresses that need to be handled in the VPLS PEs depending on the location of the I-component in the PBB-VPLS topology.

A MAC Address Withdrawal mechanism for VPLS is described in [RFC4762] to remove or unlearn MAC addresses for faster convergence on topology change in resilient H-VPLS topologies. Note that the H-VPLS topology in [RFC4762] describes two tier hierarchy to VPLS as the basic building block of H-VPLS, but it is possible to have multi-tier hierarchy in an H-VPLS.

The figure 1. described below is taken from [RFC4762] that describes dual-homing in H-VPLS.

```
                                                            PE2-rs
                                                          +--------+
                                                          |        |
                                                          |   --   |
                                                          |  /  \  |
      CE-1                                                |  \S /  |
        \                                                 |   --   |
         \                                                +--------+
          \  MTU-s                          PE1-rs        /   |
          +--------+                      +--------+     /    |
          |        |                      |        |    /     |
          |   --   |   Primary PW         |   --   |---/      |
          |  /  \  |- - - - - - - - - - - |  /  \  |          |
          |  \S /  |                      |  \S /  |          |
          |   --   |                      |   --   |---\      |
          +--------+                      +--------+    \     |
            /      \                                     \    |
           /        \                                  +--------+
          /          \                                 |        |
         CE-2         \                                |  --    |
                       \     Secondary PW              | /  \   |
                        - - - - - - - - - - - - - - - -| \S /   |
                                                       |  --    |
                                                       +--------+
                                                         PE3-rs
```

Figure 1: An example of a dual-homed MTU-s

An example of usage of the MAC Flush mechanism is the dual-homed H-VPLS where an edge device termed as MTU-s is connected to two PE devices via primary spoke PW and backup spoke PW respectively. Such redundancy is designed to protect against the failure of primary spoke PW or primary PE device. There could be multiple methods of

dual homing in H-VPLS that are not described in [RFC4762]. For example, note the following statement from section 10.2.1 in [RFC4762].

"How a spoke is designated primary or secondary is outside the scope of this document. For example, a spanning tree instance running between only the MTU-s and the two PE-rs nodes is one possible method. Another method could be configuration".

This document intends to clarify several H-VPLS dual-homing models that are deployed in practice and various use cases of LDP based MAC flush in these models.

When the MTU-s switches over to the backup PW, the requirement is to flush the MAC addresses learned in the corresponding VSI in peer PE devices participating in the full mesh, to avoid black holing of frames to those addresses. This is accomplished by sending an LDP Address Withdraw Message from the PE that is no longer connected to the MTU-s with the primary PW, with the list of MAC addresses to be removed to all other PEs over the corresponding LDP sessions [RFC4762].

In order to minimize the impact on LDP convergence time and scalability when a MAC List TLV contains a large number of MAC addresses, many implementations use a LDP Address Withdraw Message with an empty MAC List. Throughout this document the term "MAC Flush Message" is used to specify LDP Address Withdraw Message with an empty MAC List described in [RFC4762] unless specified otherwise. The solutions described in this document are applicable only to LDP Address Withdraw Message with empty MAC List.

In a VPLS topology, the core PWs remain active and learning happens on the PE-rs nodes. However when the VPLS topology changes, the PE-rs must relearn using MAC Addresses withdrawal or flush. As per the MAC Address Withdrawal processing rules in [RFC4762] a PE device on receiving a MAC Flush Message removes all MAC addresses associated with the specified VPLS instance (as indicated in the FEC TLV) except the MAC addresses learned over the PW associated with this signaling session over which the message was received. Throughout this document we use the terminology "Positive" MAC Flush or "Flush-all-but-mine" for this type of MAC Flush Message and its actions.

2.1. MAC Flush on activation of backup spoke PW

This section describes scenarios where MAC Flush withdrawal is initiated on activation of backup PW in H-VPLS.

2.1.1. PE-rs initiated MAC Flush

[RFC4762] specifies that on failure of the primary PW, it is the PE3-rs (Figure 1) that initiates MAC flush towards the core. However note that PE3-rs can initiate MAC Flush only when PE3-rs is dual homing "aware" - that is, there is some redundancy management protocol running between MTU-s and its host PE-rs devices. The scope of this document is not specific to any dual homing protocols. One example could be BGP based multi-homing in LDP based VPLS that uses the procedures defined in [I-D.ietf-l2vpn-vpls-multihoming]. In this method of dual-homing, PE3-rs would neither forward any traffic to MTU-s neither would receive any traffic from MTU-s while PE1-rs is acting a primary (or designated forwarder).

2.1.2. MTU-s initiatied MAC flush

When dual homing is achieved by manual configuration in MTU-s, the hosting PE-rs devices are dual homing "agnostic" and PE3-rs can not initiate MAC Flush message. PE3-rs can send or receive traffic over the backup PW since the dual-homing control is with MTU-s only. When the backup PW is made active by the MTU-s, the MTU-s triggers MAC Flush Message. The message is sent over the LDP session associated with the newly activated PW. On receiving the MAC Flush Message from MTU-s, PE3-rs (PE-rs device with now-active PW) would flush all the MAC addresses it has learned except the ones learned over the newly activated spoke PW. PE3-rs further initiates a MAC Flush Message to all other PE devices in the core. Note that forced switchover to backup PW can be also performed at MTU-s administratively due to maintenance activities on the "erstwhile" primary spoke PW.

MTU-s initiated method of MAC flushing is modeled after Topology Change Notification (TCN) in Rapid Spanning Tree Protocol (RSTP) [IEEE.802.1Q-2011]. When a bridge switches from a failed link to the backup link, the bridge sends out a TCN message over the newly activated link. The upstream bridge upon receiving this message flushes its entire MAC addresses except the ones received over this link and sends the TCN message out of its other ports in that spanning tree instance. The message is further relayed along the spanning tree by the other bridges.

The MAC Flush information is propagated in the control plane. The control plane message propagation is associated with the data path and hence follows similar rules for propagation as the forwarding in the LDP data plane. For example PE-rs nodes follow the data plane "split-horizon" forwarding rules in H-VPLS (Refer to section 4.4 in [RFC4762]). Therefore a MAC Flush is propagated in the context of mesh PW(s) when it is received in the context of a spoke PW. When a PE-rs node receives a MAC Flush in the context of a mesh PW then it

is not propagated to other mesh PWs.

Irrespective of whether a MAC Flush is initiated by a PE-rs or MTU-s, when a PE-rs device in the full-mesh of H-VPLS receives a MAC flush message it also flushes MAC addresses which are not affected due to topology change, thus leading to unnecessary flooding and relearning. This document describes an optional mechanism to optimize the MAC flush procedure in [RFC4762] so that it flushes only the set of MAC addresses that require relearning when topology changes in H-VPLS.

2.2. MAC Flush on failure

MAC Flush on failure is introduced in this document. In this model, the MAC Flush is initiated by PE1-rs (Figure 1) on detection of failure of the primary spoke PW and is sent to all participating PE-rs devices in the VPLS full-mesh. PE1-rs SHOULD initiate MAC flush only if PE1-rs is dual homing aware. (If PE1-rs is dual homing agnostic, the policy is do not initiate a MAC flush on failure, since that could cause unnecessary flushing in the case of single homed MTU-s.) The dual-homing protocols for this scenario are outside the scope of this document. For example, the case of PE1-rs initiated MAC flush on failure may arise when the dual-homing segment is native ethernet as opposed to spoke PWs. In this case the PE-rs devices that receives the MAC flush from PE1-rs are required to flush all the MAC addresses learned over the PW connected to PE1-rs. This cannot be achieved with the MAC Address Withdraw Message defined in [RFC4762]. This document describes extensions to MAC Flush procedures defined in [RFC4762] in order to implement MAC Flush on Failure. We use the term "negative" MAC flush or "Flush-all-from-me" for this kind of flushing action as opposed to "positive" MAC Flush action in [RFC4762]. The negative MAC flush typically results is a smaller set of MACs to be flushed.

Note that in the case of negative flush the list SHOULD be only the MACs for the affected MTU-s. If the list is empty then the negative flush will result in flushing and relearning all attached MTU-s for the originating PE-rs.

2.3. MAC Flush in PBB-VPLS

[RFC7041] describes how PBB can be integrated with VPLS to allow for useful PBB capabilities while continuing to avoid the use of MSTP in the backbone. The combined solution referred to as "PBB-VPLS" results in better scalability in terms of number of service instances, PWs and C-MACs that need to be handled in the VPLS PE-rs devices. This document describes extensions to LDP MAC Flush procedures described in [RFC4762] required to build desirable capabilities to PBB-VPLS solution.

The solution proposed in this document is generic and is applicable when MS-PWs are used in interconnecting PE devices in H-VPLS. There could be other H-VPLS models not defined in this document where the solution may be applicable.

3. Problem Description

This document describes the problems in detail with respective to various MAC flush actions described in section 2.

3.1. MAC Flush Optimization in VPLS Resiliency

This section describes the optimizations required in MAC flush procedures when H-VPLS resiliency is provided by primary and backup spoke PWs.

3.1.1. MAC Flush Optimization for regular H-VPLS

Figure 2. describes a dual-homed H-VPLS scenario for a VPLS instance where the problem with the existing MAC flush method (section 2) is explained. [RFC4762]

```
                                 PE1-rs                       PE3-rs
                               +--------+                  +--------+
                               |        |                  |        |
                               |   --   |                  |   --   |
   Customer Site 1             |  /  \  |------------------|  /  \  |->Z
   X->CE-1               /-----|  \s /  |                  |  \s /  |
       \     primary spoke PW  |   --   |           /------|   --   |
        \             /        +--------+          /       +--------+
         \    (MTU-s)/              |    \        /             |
          +--------+/               |     \      /              |
          |        |                |      \    /               |
          |   --   |                |       \  /                |
          |  /  \  |                |      H-VPLS Full Mesh Core|
          |  \s /  |                |       / \                 |
          |   --   |                |      /   \                |
         /+--------+\               |     /     \               |
        /     backup spoke PW       |    /       \              |
       /              \        +--------+         \--------+--------+
   Y->CE-2             \       |        |                  |        |
   Customer Site 2      \------|  --    |                  |  --    |
                               | /  \   |------------------| /  \   |->
                               | \s /   |                  | \s /   |
                               |  --    |                  |  --    |
                               +--------+                  +--------+
                                 PE2-rs                      PE4-rs
```

Figure 2: Dual homed MTU-s in two tier hierarchy H-VPLS

In Figure 2, the MTU-s is dual-homed to PE1-rs and PE2-rs. Only the primary spoke PW is active at MTU-s, thus PE1-rs is acting as the active device (designated forwarder) to reach the full mesh in the VPLS instance. The MAC addresses of nodes located at access sites (behind CE1 and CE2) are learned at PE1-rs over the primary spoke PW. Let's say X represents a set of such MAC addresses located behind CE-1. As packets flow from X to other MACs in the VPLS network, PE2-rs, PE3-rs and PE4-rs learn about X on their respective mesh PWs terminating at PE1-rs. When MTU-s switches to the backup spoke PW and activates it, PE2-rs becomes the active device (designated forwarder) to reach the full mesh core for MTU-s. Traffic entering the H-VPLS from CE-1 and CE-2 is diverted by the MTU-s to the spoke PW to PE2-rs. Traffic destined from PE2-rs, PE3-rs and PE4-rs to X will be blackholed till MAC address aging timer expires (default is 5 minutes) or a packet flows from X to other addresses through PE2-rs.

For example, if after the backup spoke PW is active, if a packet flows from MAC Z to MAC X, packets from MAC Z travel from PE3-rs to PE-1rs and are dropped. However, if a packet with MAC X as source and MAC Z as destination arrives at PE2-rs, PE2-rs will now learn MAC

X is on the backup spoke PW and will forward to MAC Z. At this point traffic from PE3-rs to MAC X will go to PE2-rs, since PE-3rs has also learned about MAC X. Therefore a mechanism is required to make this learning more timely in cases where traffic is not bidirectional.

To avoid traffic blackholing the MAC addresses that have been learned in the upstream VPLS full-mesh through PE1-rs, must be relearned or removed from the MAC FIBs in the VSIs at PE2-rs, PE3-rs and PE4-rs. If PE1-rs and PE2-rs are dual-homing agnostic then on activation of the standby PW from MTU-s, a MAC flush message will be sent by MTU-s to PE2-rs that will flush all the MAC addresses learned in the VPLS instance at PE2-rs from all the other PWs but the PW connected to MTU-s.

PE2-rs further relays MAC flush messages to all other PE-rs devices in the full mesh. The same processing rule applies at all those PE-rs devices: all the MAC addresses are flushed but the ones learned on the PW connected to PE2-rs. For example, at PE3-rs all of the MAC addresses learned from the PWs connected to PE1-rs and PE4-rs are flushed and relearned subsequently. Before the relearning happens flooding of unknown destination MAC addresses takes place throughout the network. As the number of PE-rs devices in the full-mesh increases, the number of unaffected MAC addresses flushed in a VPLS instance also increases, thus leading to unnecessary flooding and relearning. With large number of VPLS instances provisioned in the H-VPLS network topology the amount of unnecessary flooding and relearning increases. An optimization, described below, is required that will flush only the MAC addresses learned from the respective PWs between PE1-rs and other PE devices in the full-mesh minimizing the relearning and flooding in the network. In the example above, only the MAC addresses in set X and Y need to be flushed across the core.

The same case is applicable when PE1-rs and PE2-rs are dual homing aware and participate in a designated forwarder election. When PE2-rs becomes the active device for MTU-s then PE2-rs MAY initiate MAC flush towards the core. The receiving action of the MAC Flush in other PE-rs devices is the same as in MTU-s initiated MAC Flush. This is the [RFC4762] specified behavior.

3.1.2. MAC Flush Optimization for native Ethernet access

The analysis in section 3.1.1 applies also to the native Ethernet access into a VPLS. In such a scenario one active and one or more standby endpoints terminate into two or more VPLS or H-VPLS PE-rs devices. Examples of these dual homed access are ITU-T [ITU.G8032] access rings or any proprietary multi-chassis LAG emulations. Upon failure of the active native Ethernet endpoint on PE1-rs, an

optimized MAC flush is required to be initiated by PE1-rs to ensure that on PE2-rs, PE3-rs and PE4-rs only the MAC addresses learned from the respective PWs connected to PE1-rs are being flushed.

3.2. Black holing issue in PBB-VPLS

In a PBB-VPLS deployment a B-component VPLS (B-VPLS) may be used as infrastructure to support one or more I-component instances. The B-VPLS control plane (LDP Signaling) and learning of "Backbone" MACs (BMACs) replaces I-component control plane and learning of customer MACs (CMACs) throughout the MPLS core. This raises an additional challenge related to black hole avoidance in the I-component domain as described in this section. Figure 3 describes the case of a CE device (node A) dual-homed to two I-component instances located on two PBB-VPLS PEs (PE1-rs and PE2-rs).

```
IP/MPLS Core
                       +--------------+
                       |PE2-rs        |
                      +----+          |
                      |PBB |   +-+    |
                      |VPLS|---|P|    |
                    S/+----+  /+-+\   |PE3-rs
                    / +----+ /     \+----+
              +---+/  |PBB |/  +-+  |PBB |   +---+
     CMAC X--|CE |---|VPLS|---|P|--|VPLS|---|CE |--CMAC Y
              +---+ A +----+   +-+  +----+   +---+
                A      |PE1-rs        |        B
                       |              |
                       +--------------+
```

Figure 3: PBB Black holing Issue - CE Dual-Homing use case

The link between PE1-rs and CE-A is active (marked with A) while the link between CE-A and PE2-rs is in Standby/Blocked status. In the network diagram CMAC X is one of the MAC addresses located behind CE-A in the customer domain, CMAC Y is behind CE-B and the B-VPLS instances on PE1-rs are associated with BMAC B1 and PE2-rs with BMAC B2.

As the packets flow from CMAC X to CMAC Y through PE1-rs with BMAC B1, the remote PE-rs devices participating in the B-VPLS with the same I-SID (for example, PE3-rs) will learn the CMAC X associated with BMAC B1 on PE1-rs. Under a failure condition of the link between CE-A and PE1-rs and on activation of the link to PE2-rs, the remote PE-rs devices (for example, PE3-rs) will black-hole the traffic destined for customer MAC X to BMAC B1 until the aging timer expires or a packet flows from X to Y through the PE B2. This may take a long time (default aging timer is 5 minutes) and may affect a

large number of flows across multiple I-components.

A possible solution to this issue is to use the existing LDP MAC Flush as specified in [RFC4762] to flush the BMAC associated with the PE-rs in the B-VPLS domain where the failure occurred. This will automatically flush the CMAC to BMAC association in the remote PE-rs devices. This solution has the disadvantage of producing a lot of unnecessary MAC flush in the B-VPLS domain as there was no failure or topology change affecting the Backbone domain.

A better solution which propagates the I-component events through the backbone infrastructure (B-VPLS) is required in order to flush only the CMAC to BMAC associations in the remote PBB-VPLS capable PE-rs devices. Since there are no I-component control plane exchanges across the PBB backbone, extensions to B-VPLS control plane are required to propagate the I-component MAC Flush events across the B-VPLS.

4. Solution Description

This section describes the solution for the requirements described in section 3.

4.1. MAC Flush Optimization for VPLS Resiliency

The basic principle of the optimized MAC flush mechanism is explained with reference to Figure 2. The optimization is achieved by initiating MAC Flush on failure as described in section 2.2.

PE1-rs would initiate MAC Flush towards the core on detection of failure of primary spoke PW between MTU-s and PE1-rs (or status change from active to standby [RFC6718]). This method is referred as "MAC Flush on Failure" throughout this document. The MAC Flush message would indicate to receiving PE-rs devices to flush all MACs learned over the PW in the context of the VPLS for which the MAC flush message is received. Each PE-rs device in the full mesh that receives the message identifies the VPLS instance and its respective PW that terminates in PE1-rs from the FEC TLV received in the message and/or LDP session. Thus the PE-rs device flushes only the MAC addresses learned from that PW connected to PE1-rs, minimizing the required relearning and the flooding throughout the VPLS domain.

This section defines a generic MAC Flush Parameters TLV for LDP [RFC5036]. Through out this document the MAC Flush Parameters TLV is referred as MAC Flush TLV. A MAC Flush TLV carries information on the desired action at the PE-rs device receiving the message and is used for optimized MAC flushing in VPLS. The MAC Flush TLV can also

be used for [RFC4762] style of MAC Flush as explained in section 2.

4.1.1. MAC Flush Parameters TLV

The MAC Flush Parameters TLV is described as below:

```
0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1|1| MAC Flush Params TLV(TBD) |           Length              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Flags     | Sub-TLV Type  |         Sub-TLV Length        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Sub-TLV Variable Length Value                  |
|                             "                                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The U and F bits are set to forward if unknown so that potential intermediate VPLS PE-rs devices unaware of the new TLV can just propagate it transparently. (In the case of an B-VPLS network that has PBB-VPLS in the core with no I-components attached this message can still be useful to edge B-VPLS that do have the I-components with the ISIDs and understand the message.) The MAC Flush Parameters TLV type is to be assigned by IANA. The encoding of the TLV follows the standard LDP TLV encoding in [RFC5036]

The TLV value field contains a one byte Flag field used as described below. Further the TLV value MAY carry one or more sub-TLVs. Any sub-TLV definition to the above TLV MUST address the actions in combination with other existing sub-TLVs.

The detailed format for the Flags bit vector is described below:

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|C|N|    MBZ    | (MBZ = MUST Be Zero)
+-+-+-+-+-+-+-+-+
```

1 Byte Flag field is mandatory. The following flags are defined:

C flag, used to indicate the context of the PBB-VPLS component in which MAC flush is required. For PBB-VPLS there are two contexts of MAC flushing - The Backbone VPLS (B-component VPLS) and Customer VPLS (I-component VPLS). C flag MUST be ZERO (C=0) when a MAC Flush for the B-VPLS is required. C flag MUST be set (C=1) when the MAC Flush for I-component is required. In the regular H-VPLS case the C flag MUST be ZERO (C=0) to indicate the flush applies to the current VPLS context.

N flag, used to indicate whether a positive (N=0, Flush-all-but-mine) or negative (N=1 Flush-all-from-me) MAC Flush is required. The source (mine/me) is defined either as the PW associated with the LDP session on which the LDP MAC Withdraw was received or with the BMAC(s) listed in the BMAC Sub-TLV. For the optimized MAC Flush procedure described in this section the flag MUST be set (N=1).

Detailed usage in the context of PBB-VPLS is explained in section 4.2.

MBZ flags, the rest of the flags SHOULD be set to zero on transmission and ignored on reception.

The MAC Flush TLV SHOULD be placed after the existing TLVs in MAC Flush message in [RFC4762].

4.1.2. Application of MAC Flush TLV in Optimized MAC Flush

For optimized MAC flush, the MAC Flush TLV MAY be sent as in existing LDP Address Withdraw Message with empty MAC List but from the core PE-rs on detection of failure of its local/primary spoke PW. The N bit in TLV MUST be set to 1 to indicate Flush-all-from-me. If the optimized MAC Flush procedure is used in a Backbone VPLS or regular VPLS/H-VPLS context the C bit MUST be ZERO (C=0). If it is used in an I-component context the C bit MUST be set (C= 1). See section 4.2 for details of its usage in PBB-VPLS context.

Note that the assumption is the MAC flush TLV is understood by all devices before it is turned on in any network. See Operational Considerations section 5.

The MAC withdraw procedures defined in [RFC4762], MTU-s or PE2-rs SHOULD be sent in cases where the network is being upgraded and devices are not capable of understanding the optimized MAC flush. This would result in the same flushing action as [RFC4762] at the receiving PE-rs devices.

For the case of B-VPLS devices optimized MAC flush message SHOULD be supported.

4.1.3. MAC Flush TLV Processing Rules for Regular VPLS

This section describes the processing rules of MAC Flush TLV that SHOULD be followed in the context of MAC flush procedures in VPLS.

For optimized MAC Flush a multi-homing PE-rs initiates MAC flush message towards the other related VPLS PE-rs devices when it detects a transition (failure or to standby) in its active spoke PW. In such

case the MAC Flush TLV MUST be sent with N= 1. A PE-rs device receiving the MAC Flush TLV SHOULD follow the same processing rules as described in this section.

Note that if MS-PW is used in VPLS then a MAC flush message is processed only at the T-PE nodes since S-PE(s) traversed by the MS-PW propagate MAC flush messages without any action. In this section, a PE-rs device signifies only T-PE in MS-PW case unless specified otherwise.

When a PE-rs device receives a MAC Flush TLV with N = 1, it SHOULD flush all the MAC addresses learned from the PW in the VPLS in the context on which the MAC Flush message is received.

If a MAC Flush TLV is received with N = 0 in the MAC flush message then the receiving PE-rs SHOULD flush the MAC addresses learned from all PWs in the VPLS instance except the ones learned over the PW on which the message is received.

If a PE-rs device receives a MAC flush with the MAC Flush TLV option and a valid MAC address list, it SHOULD ignore the option and deal with MAC addresses explicitly as per [RFC4762]. It is assumed when these procedures are used all nodes support the MAC Flush Message. See section 5 Operational Considerations for details.

4.1.4. Optimized MAC Flush Procedures

This section explains the optimized MAC flush procedure in the scenario in Figure 2. When the primary spoke PW transition (failure or standby transition) is detected by PE1-rs, it MAY send MAC flush messages to PE2-rs, PE3-rs and PE4-rs with MAC Flush TLV and N = 1. Upon receipt of the MAC flush message, PE2-rs identifies the VPLS instance that requires MAC flush from the FEC element in the FEC TLV. On receiving N=1, PE-2 removes all MAC addresses learned from that PW over which the message is received. The same action is followed by PE3-rs and PE4-rs.

Figure 4 shows another redundant H-VPLS topology to protect against failure of MTU-s device. Provider RSTP [IEEE.802.1Q-2011] may be used as selection algorithm for active and backup PWs in order to maintain the connectivity between MTU-s devices and PE-rs devices at the edge. It is assumed that PE-rs devices can detect failure on PWs in either direction through OAM mechanisms such as VCCV procedures for instance.

```
                  MTU-1================PE-1===============PE-3
                    ||                  || \             /||
                    ||  Redundancy      ||  \           / ||
                    ||  Provider RSTP   ||   Full-Mesh .  ||
                    ||                  ||  /           \ ||
                    ||                  || /             \||
                  MTU-2----------------PE-2===============PE-4
                         Backup PW
```

Figure 4: Redundancy with Provider RSTP

MTU-1, MTU-2, PE1-rs and PE2-rs participate in provider RSTP. By configuration in RSTP it is ensured that the PW between MTU-1 and PE1-rs is active and the PW between MTU-2 and PE2-rs is blocked (made backup) at MTU-2 end. When the active PW failure is detected by RSTP, it activates the PW between MTU-2 and PE2-rs. When PE1-rs detects the failing PW to MTU-1, it MAY trigger MAC flush into the full mesh with MAC Flush TLV that carries N=1. Other PE-rs devices in the full mesh that receive the MAC flush message identify their respective PWs terminating on PE1-rs and flush all the MAC addresses learned from it.

[RFC4762] describes multi-domain VPLS service where fully meshed VPLS networks (domains) are connected together by a single spoke PW per VPLS service between the VPLS "border" PE-rs devices. To provide redundancy against failure of the inter-domain spoke, full mesh of inter-domain spokes can be setup between border PE-rs devices and provider RSTP may be used for selection of the active inter-domain spoke. In case of inter-domain spoke PW failure, PE-rs initiated MAC withdrawal MAY be used for optimized MAC flushing within individual domains.

Further, the procedures are applicable with any native Ethernet access topologies multi-homed to two or more VPLS PE-rs devices. The text in this section applies for the native Ethernet case where active/standby PWs are replaced with the active/standby Ethernet endpoints. An optimized MAC Flush message can be generated by the VPLS PE-rs that detects the failure in the primary Ethernet access.

4.2. LDP MAC Flush Extensions for PBB-VPLS

The use of Address Withdraw message with MAC List TLV is proposed in [RFC4762] as a way to expedite removal of MAC addresses as the result of a topology change (e.g. failure of a primary link of a VPLS PE-rs device and implicitly the activation of an alternate link in a dual-homing use case). These existing procedures apply individually to B-VPLS and I-component domains.

When it comes to reflecting topology changes in access networks connected to I-component across the B-VPLS domain certain additions should be considered as described below.

MAC Switching in PBB is based on the mapping of Customer MACs (CMACs) to Backbone MAC(s) (BMACs). A topology change in the access (I-domain) should just invoke the flushing of CMAC entries in PBB PEs' FIB(s) associated with the I-component(s) impacted by the failure. There is a need to indicate the PBB PE (BMAC source) that originated the MAC Flush message to selectively flush only the MACs that are affected.

These goals can be achieved by including the MAC Flush Parameters TLV in the LDP Address Withdraw message to indicate the particular domain(s) requiring MAC flush. On the other end, the receiving PEs SHOULD use the information from the new TLV to flush only the related FIB entry/entries in the I-component instance(s).

At least one of the following sub-TLVs MUST be included in the MAC Flush Parameters TLV if the C-flag is set to 1:

o PBB BMAC List Sub-TLV:

Type: 0x01 IANA TBA

Length: value length in octets. At least one BMAC address MUST be present in the list.

Value: one or a list of 48 bits BMAC addresses. These are the source BMAC addresses associated with the B-VPLS instance that originated the MAC Withdraw message. It will be used to identify the CMAC(s) mapped to the BMAC(s) listed in the sub-TLV.

o PBB ISID List Sub-TLV:

Type: 0x02, IANA TBA

Length: value length in octets. Zero indicates an empty ISID list. An empty ISID list means that the flush applies to all the ISIDs mapped to the B-VPLS indicated by the FEC TLV.

Value: one or a list of 24 bits ISIDs that represent the I-component FIB(s) where the MAC Flush needs to take place.

4.2.1. MAC Flush TLV Processing Rules for PBB-VPLS

The following steps describe the details of the processing rules for MAC Flush TLV in the context of PBB-VPLS:

The MAC Flush can be for the B-VPLS B-component (which applies to the BMACs and the corresponding CMACs) or the B-VPLS I-component (which applies to the CMACs) which is described in more detail here.

- The MAC Flush Message, including the MAC Flush Parameters TLV is initiated by the PBB PE(s) experiencing a Topology Change event in one or multiple customer I-component(s).

- The flags are set accordingly to indicate the type of MAC Flush required for this event: For example for an B-VPLS I-Component N=0 (Flush-all-but-mine), C=1 (Flush only CMAC FIBs).

- The PBB Sub-TLVs (BMAC and ISID Lists) are included according to the context of topology change.

- On reception of the MAC Flush message, the B-VPLS instances corresponding to the FEC TLV in the message must interpret the content of MAC Flush Parameters TLV. If the C-bit is set to 1 then Backbone Core Bridges (BCB) in the PBB-VPLS SHOULD NOT flush their BMAC FIBs. The B-VPLS control plane SHOULD propagate the MAC Flush following the data-plane split-horizon rules to the established B-VPLS topology.

- The usage and processing rules of MAC Flush Parameters TLV in the context of Backbone Edge Bridges (BEB) is as follows:

- The PBB ISID List is used to determine the particular ISID FIBs (I-component) that need to be considered for flushing action. If the PBB ISID List sub-tlv is not included in a received message then all the ISID FIBs associated with the receiving B-VPLS SHOULD be considered for flushing action.

- The PBB BMAC List is used to identify from the ISID FIBs in the previous step to selectively flush BMAC to CMAC associations depending on the N flag specified below. If PBB BMAC List Sub-TLV is not included in a received message then all BMAC to CMAC association in all ISID FIBs (I-component) as specified by the ISID List are considered for required flushing action, again depending on the N flag specified below.

- Next, depending on the N flag value the following actions apply:

- N=0, all the CMACs in the selected ISID FIBs SHOULD be flushed with the exception of the resulted CMAC list from the BMAC List mentioned in the message. ("Flush all but the CMACs associated with the BMAC(s) in the BMAC List Sub-TLV from the FIBs associated with the ISID list").

- N=1, all the resulted CMACs SHOULD be flushed ("Flush all the CMACs associated with the BMAC(s) in the BMAC List Sub-TLV from the FIBs associated with the ISID list").

4.2.2. Applicability of MAC Flush Parameters TLV

If MAC Flush Parameters TLV is received by a BEB in a PBB-VPLS that does not understand the TLV then it may result in undesirable MAC flushing action. It is RECOMMENDED that all PE-rs devices participating in PBB-VPLS support MAC Flush Parameters TLV. If this is not possible the MAC Flush Parameters TLV SHOULD be disabled as mentioned in section 5 Operational Considerations.

The MAC Flush Parameters TLV is also applicable to regular VPLS context as well as explained in section 3.1.1. To achieve negative MAC Flush (flush-all-from-me) in regular VPLS context, the MAC Flush Parameters TLV SHOULD be encoded with C=0 and N = 1 without inclusion of any Sub-TLVs. Negative MAC flush is highly desirable in scenarios when VPLS access redundancy is provided by Ethernet Ring Protection as specified in ITU-T [ITU.G8032]specification etc.

5. Operational Considerations

As mentioned before, if MAC Flush Parameters TLV is not understood by a receiver then it would result in undesired flushing action. To avoid this one solution is to develop an LDP based capability negotiation mechanism to negotiate support of various MAC Flushing capability between PE-rs devices in a VPLS instance. A negotiation mechanism is outside the scope of this document but is not required to deploy this optimized MAC flush as described below.

VPLS may be used with or without the optimization. For the case of PBB-VPLS this operation is the only method supported for ISIDs. If an operator wants the optimizations for VPLS it is the operators responsibility to make sure the VPLS that are capable of supporting the optimization are properly configured. From operational standpoint, it is RECOMMENDED that implementations of the solution provide administrative control to select the desired MAC Flushing action towards a PE-rs device in the VPLS. Thus in the topology figure 2. it is possible that PE1-rs would initiate optimized MAC Flush towards the PE-rs devices that supports the solution , whereas PE2-rs would initiate [RFC4762] style of MAC Flush towards the PE-rs devices that does not support the optimized solution. The PE-rs that supports the MAC Flush Parameters TLV MUST support the RFC4762 MAC flush procedure for completeness.

6. IANA Considerations

This document requests code point for following LDP TLV:

- o MAC Flush Parameters TLV.

Also this document requests two Sub-TLV values for

- o PBB BMAC List Sub-TLV 0x01 IANA TBA

- o PBB ISID List Sub-TLV 0x02 IANA TBA

7. Security Considerations

Control plane aspects:

- LDP security (authentication) methods as described in [RFC5036] is applicable here. Further this document implements security considerations as in [RFC4447] and [RFC4762].

Data plane aspects:

- This specification does not have any impact on the VPLS forwarding plane.

8. Contributing Authors

The authors would like to thank Marc Lasserre and Don Fedyk who made a major contribution to the development of this document.

Marc Lasserre

Alcatel-Lucent

Email: marc.lasserre@alcatel-lucent.com

Don Fedyk

Hewlett-Packard Company

Email: don.fedyk@hp.com

9. Acknowledgements

The authors would like to thank the following people who have

provided valuable comments and feedback on the topics discussed in this document: Dimitri Papadimitriou, Jorge Rabadan, Prashanth Ishwar, Vipin Jain, John Rigby, Ali Sajassi, Wim Henderickx, Paul Kwok, Maarten Vissers, Daniel Cohn, Nabil Bitar and Giles Heron.

10. References

10.1. Normative References

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

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

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

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

10.2. Informative References

[RFC7041] Balus, F., Sajassi, A., and N. Bitar, "Extensions to the Virtual Private LAN Service (VPLS) Provider Edge (PE) Model for Provider Backbone Bridging",RFC 7041, November 2013.

[I-D.ietf-l2vpn-vpls-multihoming]
Kothari, B., Kompella, K., Henderickx, W., Balus, F., Palislamovic, S., Uttaro, J., and W. Lin, "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming-06 (work in progress), October 2012.

[IEEE.802.1Q-2011]
IEEE, "IEEE Standard for Local and metropolitan area networks -- Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q, 2011.

[ITU.G8032]
International Telecommunications Union, "Ethernet ring protection switching", ITU-T Recommendation G.8032, March 2010.

[RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.

[RFC6718] Muley, P., Aissaoui, M., and M. Bocci, "Pseudowire Redundancy", RFC 6718, August 2012.

Authors' Addresses

Pranjal Kumar Dutta
Alcatel-Lucent
701 E Middlefield Road
Mountain View, California 94043
USA

Email: pranjal.dutta@alcatel-lucent.com

Florin Balus
Alcatel-Lucent
701 E Middlefield Road
Mountain View, California 94043
USA

Email: florin.balus@alcatel-lucent.com

Olen Stokes
Extreme Networks
PO Box 14129, RTP
Raleigh, North Carolina 27709
USA

Email: ostokes@extremenetworks.com

Geraldine Calvinac
France Telecom
2, avenue Pierre-Marzin
Lannion Cedex,   22307
France

Email: geraldine.calvignac@orange-ftgroup.com