MAC Flush Loop Detection in VPLS
draft-ietf-l2vpn-vpls-macflush-ld-03

Abstract

MAC Address Withdrawal is a mechanism described in [RFC4762] to remove or unlearn MAC addresses that have been dynamically learned in a VPLS instance, for faster convergence on topology change. Failure of mechanisms that control loop free connectivity among VPLS PE-rs nodes may cause MAC Address Withdrawal messages looping among those nodes, leading to Denial of Service (DoS) or complete failure of control plane in the PE-rs nodes. This document describes a mechanism to detect and prevent loops of MAC Address Withdrawal messages in a VPLS PE-rs node on such failures.

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 RFC 2119 [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 7, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.
This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction .................................................. 3
2. Terminology .................................................. 4
3. Problem Statement ............................................. 4
   3.1. MTU-s Initiated MAC Flush ............................... 6
       3.1.1. MAC Flush Loop Due to Misconfiguration .......... 7
   3.2. PE-rs Initiated MAC Flush ............................... 10
4. Loop Detection in MAC Flush ................................... 10
   4.1. Loop Detection Procedure in MAC Flush Message ...... 11
5. Applicability .................................................. 12
6. IANA Considerations .......................................... 12
7. Security Considerations ...................................... 12
8. Major Contributing Authors ................................... 13
9. Acknowledgements ............................................. 13
10. References .................................................. 13
   10.1. Normative References ................................. 13
   10.2. Informative References ............................... 14
Authors’ Addresses ............................................... 14
1. Introduction

A method of Virtual Private LAN Service (VPLS) is described in [RFC4762]. A full mesh of Pseudowire (PW) s is established between PE-rs (Provider Edge - Routing and Switching Capable) nodes to construct the VPLS core. The mesh topology provides a LAN segment or broadcast domain that is fully capable of learning and forwarding based on Ethernet MAC addresses at the PE-rs nodes. Since every PE-rs node is now directly connected to every other PE-rs nodes in the core via a PW, the topology forms a loop in data plane. A simple loop breaking rule is used - the "spilt horizon" rule described in section 4.4 in [RFC4762], whereby a PE-rs node does not forward traffic from one PW to another in the same VPLS mesh. Various mechanisms used to enforce split-horizon rule in the VPLS full mesh are out of scope of this document.

For scalability, this VPLS full mesh core configuration can be augmented with additional non-meshed spoke or MTU-s (Multi-Tenant Unit - Switching Capable) [RFC4762] nodes to provide a Hierarchical VPLS (H-VPLS) service [RFC4762]. "Spoke" PW connecting the MTU-s node and "hosting" PE-rs node can forward traffic to and receive traffic from a "Mesh" PW in the VPLS core. To protect from connection failure of the spoke PW or the hosting PE-rs node, the MTU-s node may be dual-homed into two PE-rs nodes in the same VPLS instance (see Figure 1). The dual homed connectivity also forms a loop that requires a loop breaking mechanism. The various loop breaking mechanisms used in H-VPLS dual homing are out of scope of this document.

MAC Address Withdrawal [RFC4762] is required to remove or unlearn MAC addresses for faster convergence on topology change in H-VPLS (e.g., failure of the primary spoke PW for a dual-homed MTU-s). In this document the term "MAC Flush Message" means the LDP Address Withdraw Message MAC List TLV used for MAC Address Withdrawal in VPLS.

Lack of of mechanisms or failure of mechanisms that control data plane loop free connectivity among VPLS PE-rs nodes may also cause loops of MAC Flush Messages in the LDP control plane among the participating PE-rs nodes. Such looping of MAC Flush Messages may be similar to denial of service (DoS) attacks which may lead to complete failure of the control plane in the node(s). This document describes the looping problem with MAC Flush Messages and defines a mechanism to detect and prevent such loops in the LDP control plane.

Note misconfiguration in VPLS networks may cause loops and broadcast storms in traffic forwarding as well. Data path loop detection and prevention is outside the scope of this document. Implementations may deploy methodologies available in native service forwarding to
detect and prevent such loops in data path. The control plane may not necessarily benefit from these capabilities since the control messages are handled differently at each hop.

2. Terminology

This document uses the terminology defined in [RFC5036] 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 devices offering the VPLS [RFC4762].

The terms spoke node and MTU-s in H-VPLS are used interchangeably.

"Spoke PW" means the PW 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 with MAC List TLV with 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 on 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.

3. Problem Statement

This section describes MAC Flush Loop in the LDP control plane.
An example of usage of the MAC Flush mechanism is the dual-homed H-VPLS where an edge node termed as MTU-s is connected to two hosting PE-rs nodes via primary spoke PW and backup spoke PW respectively (see Figure 1). This model is described in [RFC4762].

Such redundancy is designed to protect against the failure of primary spoke PW or hosting primary PE-rs node. 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".

When the MTU-s switches over to the backup PW (activates the backup PW), it is required to flush the MAC addresses learned in the corresponding VSI (Virtual Switch Instance) in all PE-rs devices participating in full mesh, to avoid black holing of frames to those

Figure 1: Dual homed MTU-s in two tier hierarchy H-VPLS
addresses.

In Figure 1, 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 Z, 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. 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 ageing timer expires (default is 5 minutes) or a packet flows from X to Z through PE2-rs.

To avoid traffic blackholing the MAC addresses that have been learned in the upstream VPLS full-mesh through PE1-rs, those addresses must be relearned or removed from the MAC FIBs in the VSIs at PE2-rs, PE3-rs and PE4-rs. This is accomplished by sending a MAC Flush Message with the list of MAC addresses to be removed to all PE-rs devices in the VPLS. 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 MAC Flush Message with an empty MAC List.

3.1. MTU-s Initiated MAC Flush

When dual homing in H-VPLS is achieved by manual configuration in MTU-S node, the hosting PE-rs nodes are dual homing "agnostic". It is the MTU-s node that controls the primary and backup status of spoke PWs connected to PE1-rs and PE2-rs respectively. In figure 1, PE2-rs can send and receive traffic over the backup spoke PW any time and only MTU-s blocks the traffic on the spoke PW at its end. For example Broadcast, Unicast Unlearned and Multicast (BUM) traffic received from the VPLS core is continually forwarded by PE3-rs node over the backup spoke PW to get it dropped by MTU-s node.

Therefore PE2-rs node (PE-rs node with now-active PW) cannot initiate MAC Flush message on activation of backup PW by MTU-s. When the backup PW is made active by the MTU-s node, it initiates a MAC Flush Message to PE2-rs node. The Message is sent over in the context of the newly activated spoke PW. On receiving the MAC Flush Message from MTU-s node, PE2-rs node would flush all the MAC addresses it has learned except the ones learned over the newly activated spoke PW. (Here it is assumed that MTU-s initiated a MAC Flush Message with
empty MAC List). PE2-rs node further "propagates" the MAC Flush Message to all other PE-rs nodes in the VPLS core to incur the same MAC flushing action in peer PE-rs nodes.

The MAC Flush forwarding rules in LDP control plane strictly follow the "split-horizon" forwarding rules in H-VPLS data plane (Refer to section 4.4 in [RFC4762]). So a MAC Flush is forwarded in the context of mesh PW(s) only 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 forwarded in the control plane further - means not forwarded to any mesh or spoke PWs.

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. MTU-s initiated MAC Flushing may be also applicable when RSTP may be used to ensure loop free connectivity and failure protection between MTU-s node and host PE-rs node (Refer to section 11.2 in [RFC4762]).

This method of MTU-s initiated MAC Flush is not explicitly described in [RFC4762], although it mentions about dual-homing with manual configuration or with xSTP (Any of the Spanning Tree Protocol Family) . Note that forced switchover to backup spoke PW can be also performed at MTU-s node administratively due to maintenance activities on the formerly primary spoke PW.

3.1.1. MAC Flush Loop Due to Misconfiguration

In the figure below, PE1-rs, PE2-rs, PE3-rs and PE4-rs comprises of the VPLS full mesh core. The loop free connectivity between the PE-rs devices is supposed to be ensured by split-horizon forwarding rules between mesh PWs that connects the VPLS core[RFC4762].
Figure 2. above describes a set of misconfigurations in the H-VPLS topology, which are as follows:

1. The PW that connects between PE2-rs and PE3-rs has been misconfigured at PE3-rs as spoke PW instead of mesh PW.

2. The PW that connects between PE3-rs and PE1-rs has been misconfigured at PE1-rs as spoke PW instead of mesh PW.

3. The PW that connects between PE1-rs and PE2-rs has been misconfigured at PE2-rs as spoke PW instead of mesh PW.

Such misconfiguration can occur due to configuration error in manual provisioning of a VPLS instance. Further such misconfiguration may also happen due to misconfiguration of the protocol(s) used to automate provisioning of VPLS instances.

Note the misconfiguration now has created a loop in data plane along the path PE1-rs->PE2-rs->PE3-rs->PE1-rs. An example is as follows:
1. BUM traffic received by PE1-rs from MTU-s over the primary spoke PW would be flooded to PE2-rs and PE3-rs (would be also forwarded to MTU-s over the backup spoke PW if PE2-rs is dual-homing agnostic).

2. PE2-rs on receiving the packets from PE1-rs on a spoke PW, PE2-rs would forward to PE3-rs over the mesh PW.

3. PE3-rs on receiving the packets from PE2-rs on a spoke PW, PE3-rs would forward to PE1-rs and PE4-rs over the respective mesh PWs.

4. PE1-rs on receiving the packets from PE3-rs on a spoke PW, PE1-rs would forward to PE2-rs and PE4-rs over the respective mesh PWs.

Implementations may deploy several data plane loop breaking methods that are available in native service forwarding to break such loops in data path. Data plane loop breaking methods are outside the scope of this document.

In this misconfigured topology, MTU-s activates the backup spoke PW to PE2-rs and initiates a MAC Flush Message towards PE2-rs. Since MAC flush forwarding rules are derived from the VPLS split-horizon rules data plane, a MAC flush loop would occur in the LDP control plane along PE2-rs->PE3-rs->PE1-rs->PE2-rs as follows:

1. PE2-rs on receiving the MAC Flush Message from MTU-s in the context of a spoke PW, PE2-rs would perform required flushing action and would forward the MAC Flush Message towards PE3-rs and PE4-rs in the context of respective mesh PWs. Note that PE2-rs would also forward the message to PE1-rs as well since the PW connecting to PE1-rs has been misconfigured as spoke.

2. PE3-rs on receiving the MAC Flush Message from PE2-rs in the context of a spoke PW, PE3-rs would perform required flushing action and would forward the MAC Flush Message to PE1-rs and PE4-rs in the context of respective mesh PWs.

3. PE1-rs on receiving the MAC Flush Message from PE3-rs in the context of a spoke PW, PE1-rs would perform required flushing action and would forward the MAC Flush Message to PE2-rs and PE4-rs in the context of respective mesh PWs.

Each such loop of a MAC Flush Message causes replication to \((N-1)\) messages, where \(N\) is number of PE-rs devices a replicating PE-rs is connected to. Such looping of MAC Flush Messages may lead to a denial of service (DoS) attacks or complete failure of the control plane in the node(s) and bring down services in the entire network. The LDP control plane in PE-rs or MTU-s nodes requires a fault tolerant mechanism to detect such loops and to protect against such
failures.

3.2. PE-rs Initiated MAC Flush

PE-rs initiated MAC Flush [RFC4762] specifies that on failure of the primary PW, it is the PE2-rs node (Figure 2) that initiates MAC flush towards the core. This is specified in section 10.2.2 in [RFC4762]. However note that PE2-rs node can initiate MAC Flush only when PE2-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).

When PE2-rs detects that the backup spoke PW is now active then PE2-rs initiates a MAC Flush Message towards all peer PE-rs nodes in the VPLS full-mesh core. The MAC Flush Processing and Forwarding Rules are same as explained in section 3.1. When a VPLS topology is misconfigured as described in Figure 2, a MAC Flush loop would occur in the same way as explained in section 3.1.1.

4. Loop Detection in MAC Flush

This section describes the method for Detection and Protection against MAC Flush loops in the control plane.

MAC Flush Loop Detection is a configurable option in a VPLS capable PE-rs or MTU-s node that provides a mechanism for finding and preventing MAC Flush messages from looping across the VPLS network. The mechanism makes use of LDP Path Vector TLVs defined in [RFC5036]. For loop detection in MAC Flush, Path Vector TLVs are carried in the MAC Flush Messages.

The following shows the encoding of Path Vector TLV when used for MAC Flush Loop Detection. For backward compatibility with [RFC4762] the U bit and F bit MUST be set as 1.

```
+------------------+-+------------------+-+------------------+-+------------------+-+
| 3                | 2                | 1                | 0                |
|------------------+------------------+------------------+------------------|
|0 1 2 3 4 5 6 7 8 | 0 1 2 3 4 5 6 7 8 |
|9 0 1 2 3 4 5 6 7 | 9 0 1 2 3 4 5 6 7 |
|8 9 0 1 2 3 4 5 6 | 8 9 0 1 2 3 4 5 6 |
|7 8 9 0 1 2 3 4 5 | 7 8 9 0 1 2 3 4 5 |
|6 7 8 9 0 1 2 3 4 | 6 7 8 9 0 1 2 3 4 |
|5 6 7 8 9 0 1 2 3 | 5 6 7 8 9 0 1 2 3 |
|4 5 6 7 8 9 0 1 2 | 4 5 6 7 8 9 0 1 2 |
|3 4 5 6 7 8 9 0 1 | 3 4 5 6 7 8 9 0 1 |
```

Kwok & Dutta            Expires October 2, 2013                [Page 10]
The method builds on the following basic property of the TLV:

- A Path Vector TLV contains a list of the PE-rs devices that the containing MAC Flush Message has traversed. A PE-rs is identified in a Path Vector list by its unique LDP LSR Identifier (LSR-ID), which is the first four octets of its LDP Identifier ([RFC5036]). When a PE-rs propagates a MAC flush message containing a Path Vector, it appends its LSR-ID to the Path Vector list. An LSR that receives a message with a Path Vector that contains its LSR-ID, detects that the message has traversed a loop.

The following paragraphs describe the MAC Flush Loop Detection Procedures. For these paragraphs, and only these paragraphs, "MUST" is redefined to mean "MUST if configured for Loop Detection". Further the term "PE-rs device" is used specify a VPLS capable PE-rs or a MTU-s device.

4.1. Loop Detection Procedure in MAC Flush Message

The rules that govern use of the Path Vector TLV are LDP MAC Flush messages by a PE-rs when MAC Flush Loop Detection is enabled are the following:

- If PE-rs device is originating the MAC Flush Message, it MUST include a Path Vector TLV of length 1 containing its own LSR-ID along with the MAC Flush Message originated to downstream PE-rs.

- If PE-rs device is propagating the MAC flush as a result of having received a MAC Flush from an upstream PE-rs, then:

  * If the MAC Flush message contains no Path Vector TLV, then PE-rs MUST include a Path Vector TLV of length 1 containing its own LSR-ID along with the MAC Flush Message propagated to downstream PE-rs.

  * If the MAC Flush contains a Path Vector TLV then:
If the Path Vector TLV contains its LSR-ID then the PE-rs device detects the MAC flush message has traveled in a loop and it MUST drop the MAC Flush Message without processing, else

If the Path Vector TLV exceeds the maximum allowable length, then the PE-rs device detects that the MAC flush message has traveled in a loop and it MUST drop the MAC Flush Message without processing, else,

the TLV PE-rs device MUST add its own LSR ID to the Path Vector, and MUST pass the resulting Path Vector to its downstream PE-rs along with the propagated MAC flush message.

By the definition of Path Vector TLV in [RFC5036], it supports the notion of a maximum allowable Path Vector Length; a PE-rs node that detects a Path Vector has reached the maximum length behaves as if containing message has traversed a loop. Such limit MAY be a locally configurable option in an implementation based on the scope of H-VPLS topology. The limit MAY also be driven by payload size limitations of MAC flush messages.

5. Applicability

If MAC Flush Loop Detection is desired in VPLS Network, then it should be enabled on all PE-rs nodes within that VPLS Network, otherwise Loop Detection will not operate properly and may result in undetected loops or in falsely detected loops.

PE-rs nodes that are configured for MAC Flush Loop Detection are not expected to store the Path Vectors as part of the VPLS service.

There could be other H-VPLS topologies where the problem and the solution described in this document may be applicable.

6. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

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. Major Contributing Authors

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

Don Fedyk Alcatel-Lucent, EMail: donald.fedyk@alcatel-lucent.com

9. Acknowledgements

The authors would like acknowledge the comments and suggestions from Wim Henderickx, Vach Kompella, Florin Balus, Mustapha Aissaoui, Mathew Bocci, Miroslav Vrana, Nabil Bitar, Giles Heron and Ali Sajassi.

10. References

10.1. Normative References


10.2. Informative References

[I-D.ietf-l2vpn-vpls-multihoming]


Authors’ Addresses

Paul Kwok
Alcatel-Lucent
701 E Middlefield Road
Mountain View, CA  94043
USA
Email: paul.kwok@alcatel-lucent.com

Pranjal Kumar Dutta
Alcatel-Lucent
701 E Middlefield Road
Mountain View, CA  94043
USA
Email: pranjal.dutta@alcatel-lucent.com