Workgroup: BESS WG
Internet-Draft:
draft-saum-bess-dampening-backoff-04
Published: 1 June 2022
Intended Status: Informational
Expires: 3 December 2022
Authors: S. Dikshit V. Joshi S. Shankar
Aruba, HPE Aruba, HPE Aruba, HPE

Defreezing Optimization post EVPN Mac Dampening

Abstract

MAC move handling in EVPN deployments is discussed in detail in [RFC7432]. There are few optimizations which can be done in existing way of handling the mac duplication. This document describes few of the potential techniques to do so. This document is of informational type based on comments in the ietf meeting.

Table of Contents

1. Important Terms

MDAS: Mac Dampening Attribute Set:

MDT: Mac Dampening Timer

MDC: Mac Dampening Count

MFT: MAC Freezing Timer

Mac Dampening: Process of stalling the mobility of MAC as define in [RFC7432].

VTEP: Virtual Tunnel End Point or Vxlan Tunnel End Point

DT: Dampened Time: Actual time taken to dampen the contentious MAC

2. Introduction

The host mobility solution described in [RFC7432] elaborates on few use-cases related to dual mac discovery which leads to dampening logic coming into play. The host move handling logic addresses the problem of frequent mac-moves and culminates by freezing the MACs against further moves. If there is no mellowing down of the issue, then it leads to unending cycle of mac dampening and freezing. Hence

this problem needs organic measures for arriving at MAC freezing point, sooner than later.

The events that can lead to never ending duplication are as follows:

(a) Misconfiguration of hosts with identical configuration, in the same bridge-domain, across ESIs and across NVEs.

(b) Looping of traffic due to layer 2 loops created in the bridge domain in the tenant network behind the NVEs.

2.1. Misconfiguration of Hosts

Consider the following figure wherein two hosts, Host-1 and Host-2, are misconfigured with same mac-address MAC-1. These hosts are placed behind two different Ethernet segments, ES12 and ES3 respectively and hooked to the same bridge-domain (BD-1). PE1, PE2 and PE3 will get into a never ending loop of learning the MAC-1 locally and also from the remote Vtep. Thus entering into a control-plane BGP-EVPN cycle of bumping up the sequence number in the MACMobility Extended Community till the maximum MAC move count is hit with the stipulated time. The MAC published to other Vteps like PE4 also changes accordingly based on the latest update with highest sequence number.

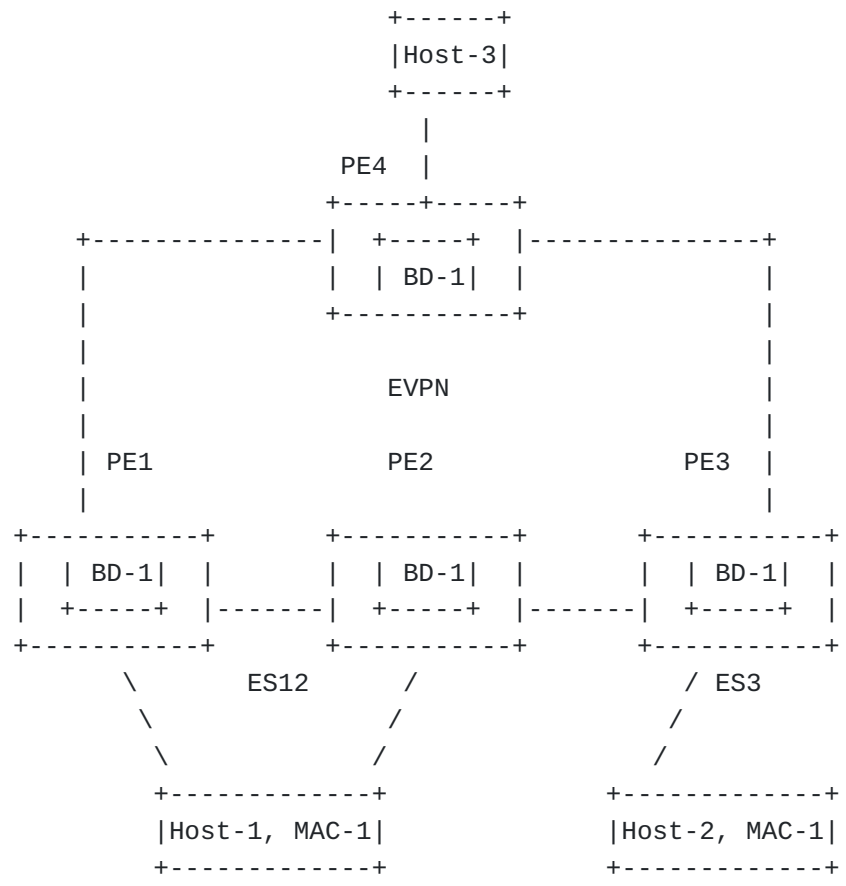

Figure 1: Figure 1: Misconfiguration of Hosts

LEGEND:
PE1, PE2, PE3: Vxlan/overlay gateways
HOST-1, HOST-2: Hosts behind PEs
MAC_1 : MAC address which is duplicated across hosts HOST-1 a
ES12: Ethernet segment between PE1 and PE2 for BD-1
ES3: Ethernet segment attached to PE3 for BD-1
BD-1: Bridge Domain 1

2.2. Loopy Traffic in Tenant Network

Consider the following case of a loopy tenant network, leading to MAC duplicity in the network. Lets say, Host-1 generates a BUM traffic like GARP (Gratuitous ARP) and sends it over the VLAN which is part of BD-1 and mapped to a configured EVI on the PEs. PE1 sprays the BUM over the EVPN fabric tying it with the mapped EVI. The BUM packet arrives at PE1 (assuming it's the elected DF) over the EVPN fabric. PE1 sprays the traffic towards the directly attached tenant network attached, tagging it with Vlan that maps to to the bridge domain, BD-1, which inturn maps to the MAC-VRF pointed to by the EVI. If the layer-2 network on tenant side is loopy due to STP network not converging or STP not configured at all, or for some

other unknown reasoni (not under the purview of this document); then the BUM traffic may loop back to PE1, thus creating a duplicate MAC learning for MAC-1. Till the tenant network is curtailed or put to order via admin intervention or otherwize, continuous MAC moves for MAC-1 can be observed between PEs attached to ethernet segment ES12 (PE1) and ES3 (PE2).

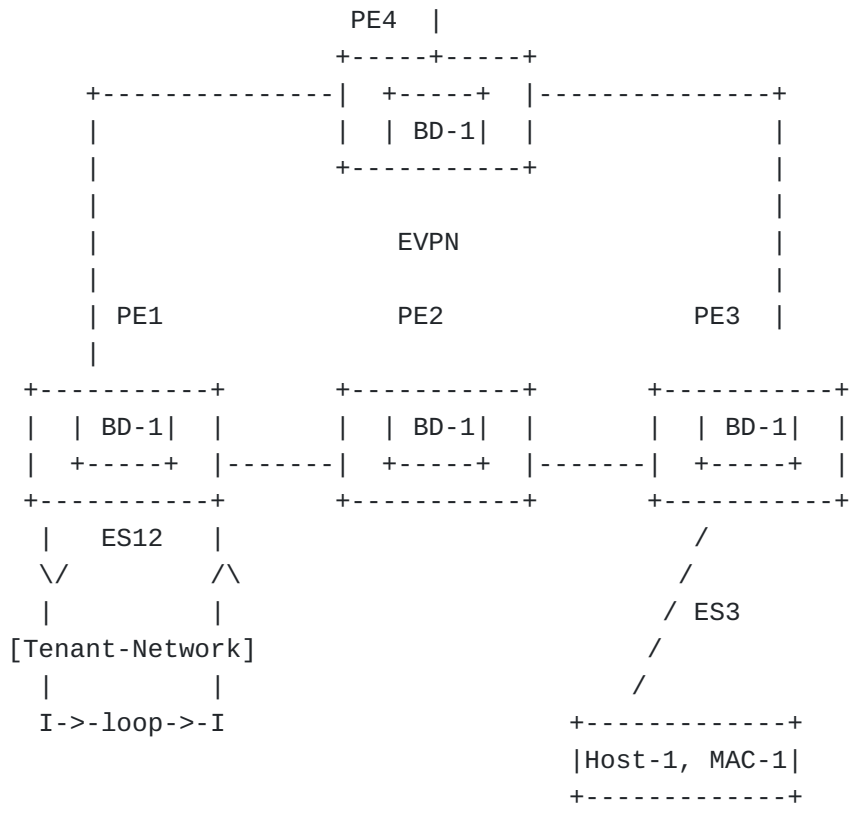

Figure 2: Figure 2: Loopy traffic in Tenant Network

LEGEND:
PE1, PE2, PE3: Vxlan/overlay gateways
HOST-1: Hosts behind PE3
MAC_1 : MAC address of Host-1
ES12: Ethernet segment between PE1 and PE2 for BD-1
ES3: Ethernet segment attached to PE3 for BD-1
BD-1: Bridge Domain 1

3. Requirements

3.1. 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].

When used in lowercase, these words convey their typical use in common language, and they are not to be interpreted as described in [RFC2119].

4. Problem Description

The mac dampening procedure mentioned in [RFC7432], suggests that a Overlay Tunnel Endpoint that detects the mac mobility event upon local learning, should start a 'M' seconds timer and track the MAC for 'N' moves before the timer expires. Hence forth, concluding that it is a MAC Duplication issue and freezing the MAC while also raising the alarm, for the admin to take corrective action. It is observed in few vendor implementations, that involves defreezing the MAC in deterministic time (configurable or derived) after freezing it, with a positive assumption that admin shall take corrective action meanwhile. Else, the subsequent unfreeze shall end up in the same cycle of MAC Duplication detection and freezing of the MAC. In case of lazy, none or inaccurate intervention by the admin, this can potentially result in ia prolong state of network disarray:

(1) Unnecessary and periodic control-plane protocol churn

(2) Exchange of control plane states which are transient and inaccurate

(3) Reachability to end device remains in the realms of ambiguity for prolonged duration

(4) Traffic destined to the Duplicate MAC case, panning across fabrics, sites or across geographies, ends up hogging the precious WAN bandwidth.

Potential solutions are discussed subsequent sections.

5. Solution(s)

The potential solutions are as follows:

5.1. Mac Freeze

The eventual solution is to FREEZE the MAC forever till admin does the clearing of the MAC. The unfreeze and clearing actions are not organic in nature and can be accompanied by unwarranted impact like clearing of other MACs in the bridge-domain. The way out may be resetting the layer-2 port and thus impacting all tenant bridge-domains hosted on the port. This solution, hence, does not always solves or mitigate the situation, or, it may create a situation from which the eventual bail-out is expensive and not restricted to the impacted Host.

5.2. Backing Off MAC Mobility Timer and Count

The best-bet to organically mellow down the never ending MAC-mobility (indicating Duplicate MAC), is to freeze the MAC temporarily, for lets say, the same time as MAC Dampening Timer(MDT). Lets term this timer as MAC Freeze Timer(MFT). MFT is the time span for which the contentious MAC is frozen, i.e., no further control plane and data flow is allowed for this MAC. The duplicity/un-ending-mobility is expected to be addressed by the admin. In case the problem is not addressed within MAC Freeze Timer, the MAC duplicity is again identified based on the MAC mobility count within the MAC Dampening timer. The best way forward MAY be:

(1) to get to the duplicity conclusion faster than the earlier iteration

(2) and freeze the MAC for a longer duration than earlier iteration

(3) , With the assumption that the problem shall be resolved in that time frame.

The MAC Dampening Attribute Set (MDAS), comprises of following three parameters:

(1) MAC Dampening Timer (MDT): Defined in [RFC7432]

(2) MAC Dampening Count (MDC): Defined in [RFC7432]

(3) MAC Freeze Timer (MFT): Time for which the MAC is frozen after MAC duplicity is detected

For example, let the first iteration of MDAS_iter_1 {MDT=180 seconds, MDC=5, MFT=180 seconds}. The default values of MDT and MDC are picket from [RFC7432], while lets define the default value of MFT same as MDT. In case admin fails to intervene, the MAC is unfrozen after MFT expires.

For second iteration of the MDAS for the problem-MAC, i.e. MDAS_iter_2 = function (MDAS_iter_1). The MDT and MDC values in second iteration are derived by backing off the MDT and MCD values by a pre-defined delta, i.e.

(1) MDAS_iter_2 (MDT) = MDAS_iter_1 (MDT) decrement_timer_delta

(2) MDAS_iter_2 (MDC) = MDAS_iter_1 (MDC) decrement_count_delta

Thus reducing the time and moves to conclude on duplicity of the MAC. The values of decrement_timer_delta and decrement_count_delta can be configured or derived on a case to case basis. [TBD:

Elaborate on the case]. The next step is to freeze the MAC for some more time as compared to the previous iteration set of MDAS, thus increasing the probability of the admin, correcting the issue:

(1) MDAS_iter_2 (MFT) = MDAS_iter_1 (MFT) + increment_timer_delta

(2) The value of increment_timer_delta is also configurable in nature.

5.2.1. MDAS Derivation

The following formulae generalizes the derivation of MDAS attributes in the Nth iteration of Duplicate MAC detection on a PE:

(1) MDAS_iter_(N) (MDT) = (MDAS_iter_(N-1) (MDT)) - decrement_mdt_delta

(2) MDAS_iter_(N) (MDC) = (MDAS_iter_(N-1) (MDC)) decrement_mdc_delta

(3) MDAS_iter_(N) (MFT) = MDAS_iter_(N-1) (MFT) + increment_mft_delta

Where in, the following values for 1st iteration can be define as follows:

MDAS_iter_1 (MDT) = 180 seconds

MDAS_iter_1 (MDC) = 5

MDAS_iter_1 (MFT) = 180 seconds. Many implementations keep the MDT and MFT values as same.

The derivation of MDAS perimeters can be exponential in nature. The delta values can be exponentially increased or decreased after certain iterations, thus triggering a exponential backing off the delta values.

5.2.1.1. MDAS Boundry Values

The new MDT value SHOULD not be less than the time taken to Dampen the MAC movement in last set of MDAS iteration. On the same lines, the new MDC count SHOULD not go below '2', as count below 2, the MAC Dampening procedure does not gets triggered.

5.2.2. Delta Values Calculation

Following bullets give a overview of potential ways the delta values, i.e. decrement_mdt_delta, increment_mdc_delta and decrement_mft_delta:

(a)

Delta values should be such that they SHOULD not infringe the time or count taken to reach Dampening state in the last set

(b) Delta values are static all through the sets

(c) Delta variable gets incremented/decremented based on the reduction in time (proportionally) to achieve the 'Dampened state' in the last 'MDAS set' as compared to the time to reach the 'Dampened state' in the MDAS set previous to the last one. For the same, the time taken to reach the Dampened State should be cached so that comparisons can be made in subsequent sets. In case, it is the first 'MDAS Set', the delta values MAY be either default or configured ones. For the second 'MDAS set', the value MAY be cross-checked against the Dampened time for the first set.

(d) Delta values are always inherited from admin configuration.

As mentioned in the Section 5.2.1 , the derivation of new delta values can done by exponentially backing them off in subsequent MDAS set(s).

5.3. Backing Off Example

This section describes the example of MDAS calculation with respect to the use-case defined in Section 2.1. Though it's equally applicable to the case described in Section 2.2. This example explains the logic in perspective of PE1. Let's say PE1 learns the MAC-1 locally and publishes it over EVPN control plane before PE2 does the same. PE1 publishes it over control plane before PE2 learns it locally (ignoring the case where both learn in tandem and publish it over control plane). Subsequently, PE2 learns it and publishes it over control plane with sequence number 1. PE1 starts the dampening logic by incrementing the local count by 1 and starting the dampening timer. If this jiggle goes on for 5 counts at PE1, MAC Dampening logic described in [RFC7432]. shall freeze the MAC. PE1 SHOULD cache the time it took to dampen the MAC. Let's say it's 30 seconds.

Assuming admin does not takes any action, before MAC freeze timer expires and PE1 defreezes the MAC, it will start moving again. PE1 shall reduce the MDT value by decrement_mdt_delta = 30 seconds to 150 seconds. The MDC counts are reduced by decrement_mdc_count = 1 to 4 and the MFT is incremented by increment_mft_delta = 20 seconds to 170 seconds. Thus PE1 shall wait for 150 seconds for concluding the dampening logic and tracks the MAC for 4 moves. Once dampening is hit, MAC is rendered as frozen for 170 seconds for admin to take action thus giving some more time for admin to take action.

The whole intention is to gradually move towards a permanent freeze of the MAC if no admin does not do the needful in the stipulated time frame.

6. Backward Compatibility

The backward comptability is a no-op for MDAS derivation and recalculation, as MAC Dampening logic is very local to the Vtep. Even if the remote Vtep does not conforms to the logic presented in this literature, it will still work towards the dampening the frequent mac-mobility with the same parameters of MDT and MDS. The instant freezing or temporary freezing of the dampened MAC is implementation dependent and should not impact or get impacted by the MDAS derivations presented in this document.

7. Security Considerations

This document inherits all the security considerations discussed in [RFC7432].

8. IANA Considerations

This document inherits all the IANA considerations discussed in [RFC7432].

9. Acknowledgements

The authors of this draft would like to thank Jorge Rabadan, Sergey Fomin and Luc Andre Burdet for their valuable comments.

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, <http://www.rfc-editor.org/rfc/rfc2119.txt>.

10.2. Informative References

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3

Networks", RFC 7348, August 2014, <http://www.rfc-editor.org/rfc/rfc7348.txt>.

[RFC7432] Sajassi, A., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015, <http://www.rfc-editor.org/rfc/rfc7432.txt>.

[RFC9014] Rabadan, J., Sathappan, S., Henderickx, W., Sajassi, A., and W. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, May 2021, <http://www.rfc-editor.org/rfc/rfc9014.txt>.

Authors' Addresses

Saumya Dikshit
Aruba Networks, HPE
Mahadevpura
Bangalore 560 048
Karnataka
India

Email: saumya.dikshit@hpe.com

Vinayak Joshi
Aruba Networks, HPE

Email: vinayak.joshi@hpe.com

Swathi Shankar
Aruba Networks, HPE

Email: swathi.shankar@hpe.com