BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: October 26, 2019

T. Yu

April 24, 2019

EVPN Enhanced Mass Withdraw
draft-yu-bess-evpn-mass-withdraw-01

Abstract

This document aims to define an enhanced mass withdraw process in case of failure of multiple ESs or vESs. This document also improves the withdraw efficiency of failure of single-homed ES or vES.

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 https://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 October 26, 2019.

Table of Contents

1. Introduction

EVPN [RFC7432] defines a mass withdraw mechanism to efficiently and quickly signal to remote PE nodes in case of a connection to ES fails. But there are particular scenarios that cannot be covered by [RFC7432]:

Multi-homed scenario:

- Failure of a line-card leads to failure of multiple ESs/vESs.

- EVC scenario (described in section 1.1 of [I-D.ietf-bess-evpn-virtual-eth-segment]):

  * Failure of physical port leads to failure of multiple multi-homed vESs aggregating EVC.

  * Failure of LAG leads to failure of multiple multi-homed vESs aggregating EVC.

- PW scenario (described in section 1.2 of [I-D.ietf-bess-evpn-virtual-eth-segment]):

  * Failure of PW leads to failure of multiple multi-homed vESs in EVPN. One of the example is: PW is using RAW mode (section 4.4.1 of [RFC4448]), with multiple VLAN services inside, and EVPN is using vlan-based interface and the services. This scenario is called "PW 1:N" in the following context of this document.

  * Failure of PW leads to failure of particular VLAN(s) in EVPN. One of the example is: a couple of PWs terminated by a EVPN using vlan-aware-bundle interface. This scenario is called "PW N:1" in the following context of this document.

The mass withdraw mechanism MUST handle both single-active and active-active multi-homed vES in scenarios described above.

Single-homed scenario:

- A failure of single-homed ES or vES interface requires a per-MAC based flush, which brings burden to the control plane.
- A failure of line-card leads to failure of multiple single-homed ESs/vESs.
- EVC scenario:
  - Failure of physical port leads to failure of multiple single-homed vESs aggregating EVC.
  - Failure of LAG leads to failure of multiple single-homed vESs aggregating EVC.
- PW scenario:
  - Failure of PW leads to failure of multiple single-homed vESs in EVPN.
  - Failure of PW leads to failure of particular VLAN(s) in EVPN.

The mass withdraw mechanism SHOULD handle a huge number of vES. Convergence mechanism independent of number of (v)ES and MAC/IP routes is preferred when possible.

2. Specification of Requirements

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

3. Terminology

EVPN: BGP MPLS-Based Ethernet VPN defined in [RFC7432]

EVI: EVPN Instance

EVPN VPWS: Refers to [RFC8214]

vES: Virtual Ethernet Segment [I-D.ietf-bess-evpn-virtual-eth-segment]

EVC: Ethernet Virtual Circuit

PW: Pseudowire

4. Solution Description

To achieve a fast convergence time in case of multiple vES fails, a concept of Administrative Group (AG) is introduced into EVPN. (v)ESs belonging to the same failure domain will be set with the same Administrative Group. A (v)ES MAY have more than one Administrative Groups.

A new EVPN BGP Extended Community called EVPN Administrative Group Community is defined as below. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of TBD.

This community MUST be ignored if not supported on the the receiving PE.

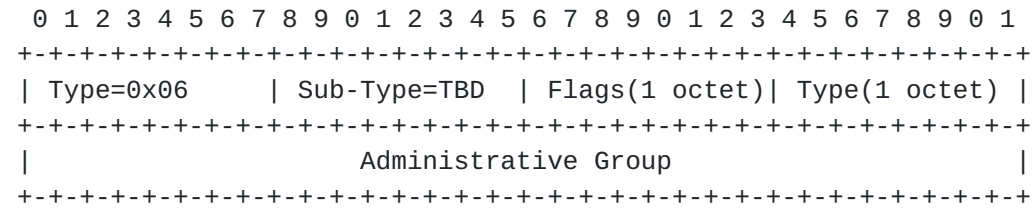

Figure 1: EVPN Administrative Group Extended Community

EVPN Administrative Group Extended community is included along with EAD route when applicable. But it needs to be included with MAC/IP Advertisement Route instead of EAD route in single-homed (v)ES and PW N:1 scenario.

When a remote PE (PE2) receives EAD containing EVPN Administrative Group Community from PE1, it first check if the corresponding (v)ES exists in the same EVI, If the (v)ES does not exist on the remote PE (PE2), the remote PE maintains a relationship between the Administrative Group and the MAC/IP routes from the corresponding (v)ES. This procedure colors the MAC/IP routes with Administrative Group. If the (v)ES exists on PE2 within the same EVI, PE2 MUST not maintain the color relationship between AG and following MAC/IP routes from PE1, as PE1 and PE2 belongs to the same multi-homed (v)ES and a failure of (v)ES in PE1 does not requires withdraw of PE2. The MAC/IP route is applicable to usage defined in [RFC7432] and also ARP/ND proxy usage defined in [I-D.ietf-bess-evpn-proxy-arp-nd]

For the scenarios mentioned that AG being included in MAC/IP route, if ESI is not all 0 (multi-homed), after checking the existence of (v)ES in the same EVI, the color relationship is directly retrieved

from the MAC/IP route. If ESI is all 0 (single-homed), then existence validation of the (v)ES is not required.

The 1 octet type field is defined to distinguish different types of Administrative Group to avoid overlap of the values across each other:

- Type 0 (0x00): This type indicates the Administrative Group is managed and configured by the operator.

- Type 1 (0x01): The Administrative Group is retrieved via ifindex of the interface (applicable to both physical and LAG interface). Refer to [RFC2863] for the usage of ifindex.

- Type 2 (0x02): The Administrative Group is retrieved via ID of the PW the EVPN is terminating.

- Type 3 (0x03): The Administrative Group is retrieved via service instance identifier of EVPN VPWS the EVPN is terminating. There is also a scenario that EVPN ELAN is terminating EVPN VPWS instead of FEC128-based [RFC4762] or FEC129-based [RFC6074] PW, in such case the Administrative Group is retrieved via service instance identifier which is defined in [RFC8214].

- Self-defined (0xF0~0xFF): These values are used for proprietary implementations to retrieve system parameters to generate self-defined value of the Administrative Group. An example is to use a type in this range to color the MAC address with ID of the line-card, which is a single-point-of-failure of a series of (v)ESs. In case of failure of the line-card, a withdraw message with the self-defined type plus ID of the line-card can be sent to remote PE to withdraw all impacted (v)ESs. The remote PE is not required to understand the meaning of self-defined type. There is no difference on the coloring and flushing procedure when using self-defined type.

Examples are given below to demonstrate the usage of Administrative Group.

- Example 1: vES1~vES1000 are under the same LAG interface, and are used to terminate EVC. In such case, these vESs belong to the same AG, the identifier of the AG is the ifindex of the LAG.

- Example 2: vES1001~vES2000 are terminating the same RAW PW. In such case, these vESs belong to the same AG, the identifier of the AG is set to the PW-ID.

- Example 3: vES2001 is a vlan-aware-bundle service interface in an EVPN, and terminating VLAN 3000~3100 (PW N:1 scenario). Each VLAN is accessed via a corresponding PW. In case of failure of a PW, only a VLAN under the vES is impacted. So the vES requires an AG for each PW with index of PW-ID (type 2). In this case, the AG community is included in MAC/IP routes instead of EAD route.

The 1 octet flags field is defined as below:

- Value 1 (0x01): Means "flush-all-from-me". When a remote PE receives a withdraw message with flags=0x01, a MAC flush procedure for MAC colored with the corresponding AG in the withdraw message is executed on both control and forwarding plane. For a single-homed (v)ES, this procedure withdraws the MAP/IP routes from remote PEs. For a multi-homed (v)ES, after the withdraw of the failed (v)ES and flush procedure of remote PEs, the MAC address will be learned again from other active (v)ES and advertised to the remote PEs.

- Value 2 (0x02): Means "frr-all-from-me". Only when the MPLS label assigned in the MAC/IP Address route of the source PE is not mapped to more than one Administrative Group, the flag is allowed be set to 0x02. For example, a PE is using label assignment per <ESI, Ethernet tag>, and the Administrative Group is retrieved via Type 1 (ifindex). In such case, the remote PE can identify the impacted ES and set the corresponding MPLS label as invalid without impact on traffic of other ESs under other interfaces within the same EVI. Another example is per <MAC-VRF> based, with this assignment method, the other (v)ESs without failure is impacted if remote PE set the label invalid. For detailed information on the assignment of label in MAC/IP Address route, refer to section 9.2.1 of [RFC7432]. When a remote PE receives a withdraw message with flags=0x02, it requires a validation of existence of aliasing labels. If the aliasing label (section 8.4 of [RFC7432]) does not exist, the procedure downgrades to "flush-all-from-me". If the aliasing label exists, the PE should process a Fast-Re-Route procedure, directly set the MPLS label of impacted (v)ES to invalid on the control and forwarding plane. This will speed up the convergence time and independent of amount of MAP/IP routes. At the same time, the remote PE needs to start a timer (T0) on control plane, to mark the corresponding MAP/IP routes to "polluted" status. During T0, if new MAC/IP routes are learned via other multi-homed PEs, update the routing table and clear the "polluted" flag of corresponding MAC/IP routes. After T0 expires, MAC/IP routes with "polluted" flag SHOULD be cleared on both control plane and forwarding plane. The length of T0 SHOULD be configurable and RECOMMENDED to be equal to MAC aging time.

To construct a flushing message, ESI of the EAD route filled with MAX-ESI, Ethernet Tag and MPLS field with all 0 and the Administrative Group Community together with a list of Route Targets corresponding to the impacted service instances. If the number of Route Targets is more than they can fit into a single attribute, then can split the RTs into multiple messages with same Administrative Group Community attached.

5. Acknowledgments

TBD

6. Security Considerations

TBD

7. IANA Considerations

IANA is requested to allocate a new "EVPN Extended Community Sub-Types" registry defined in [RFC7153] as follow:

SUB-TYPE	NAME	Reference
TBD	EVPN Administrative Group Community	This document

This document creates registry below.

Administrative Group Type:

Value	Meaning	Reference
0x00	Manually managed Administrative Group	This document
0x01	AG is retrieved via ifindex	This document
0x02	AG is retrieved via ID of PW	This document
0x03	AG is retrieved via EVPN service instance id	This document
0xF0~0xFF	Reserved for proprietary implementation	This document

Administrative Group Flags:

Value	Meaning	Reference
0x01	Flush-all-from-me	This document
0x02	FRR-all-from-me	This document

8. References

8.1. Normative References

[I-D.ietf-bess-evpn-proxy-arp-nd]
Rabadan, J., Sathappan, S., Nagaraj, K., Henderickx, W., Hankins, G., King, T., Melzer, D., and E. Nordmark, "Operational Aspects of Proxy-ARP/ND in EVPN Networks", draft-ietf-bess-evpn-proxy-arp-nd-05 (work in progress), November 2018.

[I-D.ietf-bess-evpn-virtual-eth-segment]
Sajassi, A., Brissette, P., Schell, R., Drake, J., and J. Rabadan, "EVPN Virtual Ethernet Segment", draft-ietf-bess-evpn-virtual-eth-segment-04 (work in progress), January 2019.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

8.2. Informative References

[RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, DOI 10.17487/RFC2863, June 2000, <https://www.rfc-editor.org/info/rfc2863>.

[RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, DOI 10.17487/RFC4448, April 2006, <https://www.rfc-editor.org/info/rfc4448>.

[RFC4762] Lasserre, M., Ed. and V. Kompella, Ed., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, DOI 10.17487/RFC4762, January 2007, <https://www.rfc-editor.org/info/rfc4762>.

[RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, DOI 10.17487/RFC6074, January 2011, <https://www.rfc-editor.org/info/rfc6074>.

Author's Address

Tianpeng Yu

EMail: yutianpeng.ietf@gmail.com