BESS Workgroup A. Sajassi
INTERNET-DRAFT P. Brissette
Intended Status: Standards Track Cisco
J. Uttaro
ATT
J. Drake
W. Lin
Juniper
S. Boutros
VMWare
J. Rabadan
Nokia

Expires: January 6, 2016 July 6, 2016

EVPN VPWS Flexible Cross-Connect Service
draft-sajassi-bess-evpn-vpws-fxc-00.txt

Abstract

This document describes a new EVPN VPWS VLAN-aware bundle service type referred to as flexible cross-connect service. It also describes the rational for this new service as well as a solution to deliver such service.

Table of Contents

1 Introduction

[EVPN-VPWS] describes a solution to deliver P2P services using BGP constructs defined in [RFC7432]. It delivers this P2P service between a pair of Attachment Circuits (ACs), where an AC can designate on a PE a port, a VLAN on a port, or a group of VLANs on a port. It also leverages multi-homing and fast convergence capabilities of [RFC7432] in delivering these VPWS services. Multi-homing capabilities include the support of single-active and all-active redundancy mode and fast convergence is provided using "mass withdraw" message in control-plane and fast protection switching using prefix independent convergence in data-plane upon node or link failure. Furthermore, the use of EVPN BGP constructs eliminates the need for multi-segment PW auto-discovery and signaling if the VPWS service need to span across multiple ASes.

Some service providers have very large number of ACs (in millions) that require tag manipulation (e.g., VLAN translation) to be back hauled across their MPLS/IP network. These service providers want to multiplex a large number of ACs across several physical interfaces (e.g., several Ethernet Segments) onto a single VPWS service tunnel in order to a) reduce number of EVPN service labels associated with VPWS service tunnels and thus the associated OAM monitoring, and b) reduce EVPN BGP signaling (e.g., not to signal each AC as it is the case in [EVPN-VPWS]).

These service provider want the above functionality without scarifying any of the capabilities of [EVPN-VPWS] including single-active and all-active multi-homing, and fast convergence.

This document presents a solution based on extensions to [EVPN-VPWS] to meet the above requirements.

1.1 Terminology

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].

MAC: Media Access Control

MPLS: Multi Protocol Label Switching

OAM: Operations, Administration and Maintenance

PE: Provide Edge Node

CE: Customer Edge device e.g., host or router or switch

EVPL: Ethernet Virtual Private Line

EPL: Ethernet Private Line

ES: Ethernet Segment

VPWS: Virtual private wire service

EVI: EVPN Instance

VPWS Service Tunnel: It is represented by a pair of EVPN service labels associated with a pair of endpoints. Each label is downstream assigned and advertised by the disposition PE through an Ethernet A-D per-EVI route. The downstream label identifies the endpoint on the disposition PE. A VPWS service tunnel can be associated with many VPWS service identifiers for VLAN-aware VPWS service where each identifier is a normalized VID.

Single-Active Mode: When a device or a network is multi-homed to two or more PEs and when only a single PE in such redundancy group can forward traffic to/from the multi-homed device or network for a given VLAN, then such multi-homing or redundancy is referred to as "Single-Active".

All-Active: When a device is multi-homed to two or more PEs and when all PEs in such redundancy group can forward traffic to/from the multi-homed device for a given VLAN, then such multi-homing or redundancy is referred to as "All-Active".

2 Conflicting Requirements

Two of the main motivations for service providers seeking a new solution are: 1) to reduce number of VPWS service tunnels by muxing large number of ACs across different physical interfaces instead of having one VPWS service tunnel per AC, and 2) to reduce the signaling of ACs as much as possible. Besides these two requirements, they also want multi-homing and fast convergence capabilities of [EVPN-VPWS].

In [EVPN-VPWS], a PE signals an AC indirectly by first associating that AC to a VPWS service tunnel (e.g., a VPWS service instance) and then signaling the VPWS service tunnel via a per-EVI Ethernet AD route with Ethernet Tag field set to a 24-bit VPWS service instance identifier (which is unique within the EVI) and ESI field set to a 10-octet identifier of the Ethernet Segment corresponding to that AC. Therefore, a PE device that receives such EVPN routes, can associate

the VPWS service tunnel to the remote Ethernet Segment, and when the remote ES fails and the PE receives the "mass withdraw" message associated with the failed ES per [RFC7432], it can update its BGP path list for that VPWS service tunnel quickly and achieve fast convergence for multi-homing scenarios. Even if fast convergence were not needed, there would still be a need for signaling each AC failure (via its corresponding VPWS service tunnel) associated with the failed ES, so that the BGP path list for each of them gets updated accordingly and the packets are sent to backup PE (in case of single-active multi-homing) or to other PEs in the redundancy group (in case of all-active multi-homing). In absence of updating the BGP path list, the traffic for that VPWS service tunnel will be black-holed.

When a single VPWS service tunnel multiplexes many ACs across number of Ethernet Segments (number of physical interfaces) and the ACs are not signaled via EVPN BGP to remote PE devices, then the remote PE devices neither know the association of the received Ethernet Segment to these ACs (and in turn to their local ACs) nor they know the association of the VPWS service tunnel (e.g., EVPN service label) to the far-end ACs - i.e, the remote PEs only know the association of their local ACs to the VPWS service tunnel but not the far-end ACs. Thus upon a connectivity failure to the ES, they don't know how to redirect traffic via another multi-homing PE to that ES. In other words, even if an ES failure is signaled via EVPN to the remote PE devices, they don't know what to do with such message because they don't know the association among the ES, their ACs, and the VPWS service tunnel.

In order to address this issue when multiplexing large number of ACs onto a single VPWS service tunnel, two mechanisms are devised: one to support VPWS services between two single-homed endpoints and another one to support VPWS services where one of the endpoints is multi-homed. An endpoint can be an AC, MAC-VRF, IP-VRF, global table, or etc.

For single-homed endpoints, it is OK not to signal each AC in BGP because upon connection failure to the ES, there is no alternative path to that endpoint. However, the ramification for not signaling an AC failure is that the traffic destined to the failed AC, is sent over MPLS/IP core and then gets discarded at the destination PE - i.e., it can waste network resources. However, when there is a connection failure, the application layer will eventually stop sending traffic and thus this wastage of network resources should be transient. Section 4.1 describes a solution for such single-homing VPWS service which is called VLAN-Unaware flexible cross-connect service.

For VPWS services where one of the endpoints is multi-homed, there

are two options:

1) to signal each AC via BGP so that the path list can be updated upon a failure that impacts those ACs. This solution is described in section 4.2 and it is called VLAN-Aware flexible cross-connect service.

2) to bundle several ACs on an ES together per destination ES (or PE) and associated such bundle to a single VPWS service tunnel. This is similar to VLAN-bundle service interface described in [EVPN-VPWS]. This solution is described in section 4.3.

4 Solution

This section describes a solution for providing a new VPWS service between two PE devices where a large number of ACs (e.g., VLANs) that span across many physical interfaces on each PE are multiplex onto a single P2P EVPN LSP tunnel. Since multiplexing is done across several physical interfaces, there can be overlapping VLAN IDs across these interfaces; therefore, in such scenarios, the VLAN IDs (VIDs) MUST be translated into unique VIDs to avoid collision. Furthermore, if the number of VLANs that are getting multiplex onto a single VPWS service tunnel, exceed 4K, then a single tag to double tag translation MUST be performed. This translation of VIDs into unique VIDs (either single or double) is referred to as "VID normalization". When single normalized VID is used, the lower 12-bit of Ethernet tag field in EVPN routes is set to that VID and when double normalized VID is used, the lower 12-bit of Ethernet tag field is set to inner VID and the higher 12-bit is set to the outer VID.

Since there is only a single P2P EVPN LSP tunnel associated with many normalized VIDs (either single or double), MPLS lookup at the disposition PE is no longer sufficient to forward the packet to the right egress endpoint/interface. Therefore, in addition to an EVPN label lookup corresponding to the VPWS service tunnel, a VID lookup (either single or double) is also required. On the disposition PE, one can think of the lookup of EVPN label results in identification of a VID table, and the lookup of normalized VID(s) in that table, results in identification of egress endpoint/interface. The tag manipulation (translation from normalized VID(s) to local VID) can be performed either as part of the VID table lookup or at the egress interface itself.

Since VID lookup (single or double) needs to be performed at the disposition PE, then VID normalization MUST be performed prior to the MPLS encapsulation on the ingress PE. This requires that both imposition and disposition PE devices be capable of VLAN tag

manipulation, such as re-write (single or double), addition, deletion (single or double), at their endpoints (e.g., their physical interfaces).

4.1 VLAN-Unaware Flexible Xconnect - Single-Homing

In this mode of operation, many ACs across several Ethernet Segments are multiplex into a single P2P EVPN LSP tunnel represented by a single VPWS service ID. VLAN-Unaware mode for this solution means that VLANs (normalized VIDs) are not signaled via EVPN BGP among the PEs. In this solution, there is only a single P2P EVPN LSP tunnel between a pair of PEs for all their ACs that are single-homed.

As discussed previously, since the VPWS service tunnel is used to multiplex ACs across different ES's (e.g., physical interfaces), the EVPN label alone is not sufficient for proper forwarding of the received packets (over MPLS/IP network) to egress interfaces. Therefore, normalized VID lookup is required in the disposition direction to forward packets to their proper egress end-points/ interfaces - i.e., the EVPN label lookup identifies a VID table and subsequently, the normalized VID lookup in that table, identifies the egress interface.

In this solution, on each PE, the single-homing ACs represented by their normalized VIDs are associated with a single VPWS service tunnel (in a given EVI). The EVPN route that gets generated is an EVPN Ethernet AD per EVI route with ESI=0, Ethernet Tag field set to VPWS service instance ID, MPLS label field set to dynamically generated EVPN service label representing the EVPN VPWS service. This route is sent with an RT representing the EVI. This RT can be auto-generated from the EVI per section 5.1.2.1 of [EVPN-Overlay]. Furthermore, this route is sent with the EVPN Layer-2 Extended Community defined in section 3.1 of [EVPN-VPWS] with two new flags (defined in section 5) that indicate: 1) this VPWS service tunnel is for VLAN-unaware Flexible Cross-Connect, and 2) normalized VID type (single versus double). The receiving PE uses these new flags for consistency check and MAY generate an alarm if it detects inconsistency but doesn't bring down the VPWS service because such inconsistency may be intentional - i.e., one side is configured for VLAN-aware VPWS service and another side is configured for VLAN-unaware VPWS service.

It should be noted that in this mode of operation, a single Ethernet AD route is sent upon configuration of the first AC (ie, normalized VID). Later, when additional ACs are configured and associated with this EVPN VPWS service tunnel, the PE does not advertise any additional EVPN BGP routes. The PE only associates locally these ACs

with the already created VPWS service tunnel.

4.2 VLAN-Aware Flexible Xconnect

In this mode of operation, just as the VLAN-unaware mode, many normalized VIDs (ACs) across several different ES's/interfaces are multiplexed into a single P2P EVPN LSP tunnel; however, this single tunnel is represented by many VPWS service IDs (one per normalized VID) and these normalized VIDs are signaled using EVPN BGP.

In this solution, on each PE, the multi-homing ACs represented by their normalized VIDs are configured with a single EVI. There is no need to configure VPWS service instance ID in here. A VPWS service instance ID is derived automatically from each normalized VID. For each normalized VID on each ES, the PE generates an EVPN Ethernet AD per EVI route where ESI field represents the ES ID, the Ethernet Tag field is set to the normalized VID, MPLS label field is set to dynamically generated EVPN label representing the P2P EVPN LSP tunnel. This route is sent with an RT representing the EVI. As before, this RT can be auto-generated from the EVI per section 5.1.2.1 of [EVPN-Overlay]. Furthermore, this route is sent with the EVPN Layer-2 Extended Community defined in section 3.1 of [EVPN-VPWS] with two new flags (defined in section 5) that indicate: 1) this VPWS service tunnel is for VLAN-aware Flexible Cross-Connect, and 2) normalized VID type (single versus double). The receiving PE uses these new flags for consistency check and MAY generate an alarm if it detects inconsistency but doesn't bring down the VPWS service because such inconsistency may be intentional - i.e., one side is configured for VLAN-aware VPWS service and another side is configured for VLAN-unaware VPWS service.

It should be noted that in this mode of operation, the PE sends a single Ethernet AD route for each AC that is configured - i.e., each normalized VID that is configured per ES results in generation of an EVPN Ethernet AD per EVI.

This mode of operation provides automatic cross checking of normalized VIDs used for EVPL services because these VIDs are signaled in EVPN BGP. For example, if the same normalized VID is configured on three PE devices (instead of two) for the same EVI, then when a PE receives the second EVPN Eth-AD per EVI route, it generates an error message unless the two EVPN Eth-AD per EVI routes include the same ESI. Such cross-checking is not feasible in VLAN-unaware FXC because the normalized VIDs are not signaled.

4.3 VLAN-Unaware Flexible Xconnect - Multi-Homing

In this mode of operation, a group of normalized VIDs (ACs) on a single ES that are destined to a single endpoint/interface are multiplexed into a single P2P EVPN LSP tunnel represented by a single VPWS service ID. This mode of operation is the same as VLAN-bundle service interface of [EVPN-VPWS] except for the fact that VIDs on Ethernet frames are normalized before getting sent over the LSP tunnel.

In the previous two modes of operation, only a single EVPN VPWS service tunnel is needed per pair of PEs. However, in this mode of operation, there can be lot more service tunnels per pair of PEs - i.e, there is one tunnel per group of VIDs per pair of PEs and there can be many groups between a pair of PEs, thus resulting in many EVPN service tunnels.

5. BGP Extensions

This draft uses the EVPN Layer-2 attribute extended community defined in [EVPN-VPWS] with two additional flags added to this EC as described below. This EC is to be advertised with Ethernet A-D per EVI route per section 4.

```
+-------------------------------------+
|  Type(0x06)/Sub-type(TBD)(2 octet) |
+-------------------------------------+
|  Control Flags (2 octets)          |
+-------------------------------------+
|  L2 MTU (2 octets)                 |
+-------------------------------------+
|  Reserved (2 octets)               |
+-------------------------------------+
```

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   MBZ           | V | M |C|P|B|  (MBZ = MUST Be Zero)
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following bits in the Control Flags are defined; the remaining bits MUST be set to zero when sending and MUST be ignored when receiving this community.

Name	Meaning
B,P,C	per definition in [EVPN-VPWS]
M	00 mode of operation as defined in [EVPN-VPWS] 01 VLAN-aware FXC 10 VLAN-unaware FXC
V	00 operating per [EVPN-VPWS] 01 single-VID normalization 10 double-VID normalization

The M and V fields are OPTIONAL on transmission and ignored at reception for forwarding purposes. They are used for error notifications.

6 Failure Scenarios

6.2 EVPN VPWS service Failure

The failure detection of an EVPN VPWS service can be performed via OAM mechanisms such as VCCV-BFD and upon such failure detection, the switch over procedure to the backup S-PE is the same as the one described above.

6.2 Attachment Circuit Failure

6.3 PE Port Failure

6.4 PE Node Failure

In the case of PE node failure, the operation is similar to the steps described above, albeit that EVPN route withdrawals are performed by the Route Reflector instead of the PE.

7 Security Considerations

TBD.

8 IANA Considerations

TBD

9 References

9.1 Normative References

[RFC7432] Sajassi et al., "Ethernet VPN", RFC 7432, February 2015.

[EVPN-IRB] Sajassi et al., "Integrated Routing and Bridging in EVPN", draft-ietf-bess-evpn-inter-subnet-forwarding-00, work in progress, November 2014.

[EVPN-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", draft-ietf-bess-evpn-prefix-advertisement-02, work in progress, September 2015.

[RFC6718] Muley P., et al., "Pseudowire Redundancy", RFC 6718, August 2012.

[RFC6870] Muley P., et al., "Pseudowire Preferential Forwarding Status Bit", RFC 6870, February 2013.

9.2 Informative References

[BGP-PIC] Bashandy A. et al., "BGP Prefix Independent Convergence", draft-rtgwg-bgp-pic-02.txt, work in progress, October 2013.

Authors' Addresses

A. Sajassi
Cisco
EMail: sajassi@cisco.com

P. Brissette
Cisco
EMail: pbrisset@cisco.com

J. Uttaro
ATT

EMail: ju1738@att.com

J. Drake
Juniper
EMail: jdrake@juniper.net

S. Boutros
ATT
EMail: boutros.sami@gmail.com

W. Lin
Juniper
EMail: wlin@juniper.net

J. Rabadan
jorge.rabadan@nokia.com