BESS Working Group
Internet Draft
Category: Standard Track

Ali Sajassi
Gaurav Badoni
Priyanka Warade
Suresh Pasupula
Cisco Systems

Expires: January 2, 2017
July 2, 2017

L3 Aliasing and Mass Withdrawal Support for EVPN
draft-sajassi-bess-evpn-ip-aliasing-00.txt

Abstract

This draft proposes an extension to [RFC7432] to do Aliasing for Layer 3 routes that is needed for symmetric IRB to build a complete IP ECMP.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

```
                                     +---------+
                  +-------------+    |         |
                  |             |    |         |
                / |    PE1      |----|         |   +-------------+
               /  |             |    |  MPLS/  |   |             |
              /   +-------------+    |  VxLAN/ |   |     PE3     |---H3
         H1---                       |  NVGRE  |   |             |
              \   +-------------+    |         |---|             |
               \  |             |    |         |   +-------------+
                \ |     PE2     |----|         |
                  |             |    |         |
                  +-------------+    |         |
                                     |         |
                                     |         |
                                     +---------+
```

Figure 1: Inter-subnet traffic between Multihoming PEs and Remote PE

Consider a pair of multi-homing TORs PE1 and PE2. Let there be a host H1 attached to them. Consider another TOR PE3 and a host H3 attached to it.

With Asymmetric IRB, if H3 sends inter-subnet traffic to H1, routing will happen at PE3. PE3 will have the destination SVI and will trigger ARP if it does not have an ARP adjacency to H1. Finally routing lookup will resolve destination MAC to H1's MAC address. Furthermore, H1's MAC will point to a VxLAN ECMP to T1 and T2, either due to host route advertisement or MAC Aliasing as detailed in [RFC 7432].

With Symmetric IRB, if H3 sends inter-subnet traffic to H1, routing lookup will happen at PE3. PE3 will do a routing lookup in the L3VNI-VRF context and is not expected to have the destination SVI. Therefore at PE3, we need an IP ECMP list (PE1/PE2) to be built for H1's IP address for proper load balancing. If H1 is locally learnt only at one of the PEs, PE1 or PE2 due to port-channel hashing, we will not be able to build IP ECMP at PE3 as we do not do Aliasing for Layer 3 addresses.

This draft proposes an extension to do Aliasing for Layer 3 routes that is needed for symmetric IRB to build a complete IP ECMP.

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].

IRB: Integrated Routing and Bridging

IRB Interface: A virtual interface that connects the bridging module and the routing module on an NVE.

Broadcast Domain: In a bridged network, the broadcast domain corresponds to a Virtual LAN (VLAN), where a VLAN is typically represented by a single VLAN ID (VID) but can be represented by several VIDs where Shared VLAN Learning (SVL) is used per [802.1Q].

Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.

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

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

LACP: Link Aggregation Control Protocol.

PE: Provider Edge device.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

2 IP Aliasing and Backup Path

Host IP and MAC routes are learnt by PEs on the access side via a control plane protocol like ARP. In case where a CE is multihomed to multiple PE nodes using a LAG and is running in All-Active Redundancy Mode, the Host IP will be learnt and advertised in the MAC/IP Advertisement only by the PE that receives the ARP packet. As a result, the remote PE sees only one next-hop for the Host IP and forwards traffic to that advertising PE. Hence, the remote PE is not be able to effectively load balance the traffic towards the multihomed Ethernet Segment.

To address this issue, concept of Aliasing that was introduced in RFC 7432 [RFC7432], can be extended for Layer 3 routes as well. The PE SHOULD advertise reachability to an L3 VRF instance on a given ES for IP addresses using the existing EAD/EVI route. In this case, the EVPN instance is the VRF table to which the host IP address belongs. This will henceforth be referred to as the IP-EAD/EVI route.

A remote PE that receives an IP route with a non reserved ESI SHOULD consider it reachable by all PEs that have advertised the IP-EAD/EVI advertisement route and the EAD/ES advertisement route containing the VRF Route-Targets for that ES. The EAD/ES route must have the Single-Active bit in the flags of the ESI Label extended community set to 0 for Aliasing to take effect.

The IP-EAD/EVI route cannot be used for route forwarding until the associated Ethernet A-D per ES route is received.

In case of Single-Active redundancy mode, the remote PE SHOULD use the IP-EAD/EVI route EVPN Layer 2 attribute extended community as mentioned in draft-ietf-bess-evpn-vpws-07 in combination with the EAD/ES route to determine the Backup Path for the IP addresses for the given IP VRF context. This alternate path SHOULD be installed as a backup path for the IP address.

2.1 Constructing Ethernet A-D per EVPN Instance Route

This draft proposes the advertisement of per EVI Ethernet A-D route for IP VRFs to enable Aliasing for IP addresses. The usage/construction of this route remains similar to that described in RFC 7432 with a few notable exceptions as below.

* The Route-Distinguisher should be set to the corresponding L3VPN context.

* The Ethernet Tag should be set to 0.

* The L3 EAD/EVI SHOULD carry one or more IP VRF Route-Target (RT)

attributes.

* The L3 EAD/EVI SHOULD carry the RMAC Extended Community attribute.

* The MPLS Label usage should be as described in RFC 7432.

It is important to note that the prefix for a IP-EAD/EVI and L2-EAD/EVI may be identical. However, since the RD of the IP-EAD/EVI is set to the corresponding L3VPN context and the RD of the L2-EAD/EVI is set to the corresponding MAC-VRF context, the import will happen in the respective IP-VRFs and MAC-VRFs and hence, the prefix will not be overwritten.

3 Fast Convergence for Routed Traffic

In EVPN, Host IP reachability is learned via the BGP control plane over the MPLS network. All the hosts that are dually connected behind an ES are advertised by the PEs belonging to the redundancy group. A remote TOR receiving these host routes can loose reachability from any of the PEs either due to box reload or core failure or access failure for that PE.

BGP PIC functionality is the existing mechanism for fast convergence as described in https://tools.ietf.org/html/draft-rtgwg-bgp-pic-02. PIC feature doesn't solve the convergence issue for the access failure cases as the PEs are still reachable from the remote TOR.

To alleviate this, EVPN defines a mechanism to efficiently and quickly signal, to remote PE nodes, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment. This is done by having each PE advertise a set of one or more Ethernet A-D per ES routes for each locally attached Ethernet segment (refer to Section 3.1 below for details on how these routes are constructed). A PE may need to advertise more than one Ethernet A-D per ES route for a given ES because the ES may be in a multiplicity of EVIs and the RTs for all of these EVIs may not fit into a single route. Advertising a set of Ethernet A-D per ES routes for the ES allows each route to contain a subset of the complete set of RTs. Each Ethernet A-D per ES route is differentiated from the other routes in the set by a different Route Distinguisher (RD).

Upon failure in connectivity to the attached ES, the PE withdraws the corresponding set of Ethernet A-D per ES routes. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all IP addresses across IP VRFs associated with the Ethernet segment in question. If no other PE has advertised an Ethernet A-D route for the same segment, then the PE that received the withdrawal simply invalidates the IP entries for that segment. Otherwise, the

PE updates its next-hop adjacencies accordingly.

These routes should be processed with higher priority than other MAC or MAC-IP withdrawals upon failure. Similar priority processing is needed even on the intermittent RRs.

This draft is addressing the mass withdrawal behavior for routed traffic. For Layer-2, please refer to Section 8.2 of RFC 7432.

3.1 Constructing Ethernet A-D per Ethernet Segment Route

This section describes the procedures used to construct the Ethernet A-D per ES route, which is used for fast convergence (as discussed above). The usage/construction of this route remains similar to that described in section 8.2.1. of RFC 7432 with a few notable exceptions as explained in following sections.

3.1.1 Ethernet A-D Route Targets

Each Ethernet A-D per ES route MUST carry one or more Route Target (RT attributes). The set of Ethernet A-D routes per ES MUST carry the entire set of IP VRF RTs for all the IP VRFs in addition to MAC VRF RTS for all the EVPN instance to which the Ethernet segment belongs.

3.2 Avoiding convergence issues by syncing IP prefixes

Consider a pair of multi-homing TORs PE1 and PE2. Let there be a host H1 attached to them. Consider another TOR PE3 and a host H3 attached to it.

If the host H1 is learnt on both the PEs, ECMP path list is formed on PE3 pointing to (PE1/PE2). Traffic from H3 to H1 is not impacted even if one of the TORs becomes unreachable as the path list gets corrected upon receiving the mass withdrawal route (Ethernet A-D segment).

Let us consider a case where H1 is locally learnt only on PE1 due to port-channel hashing. At PE3, H1 has ECMP path list (PE1/PE2) using Aliasing as described in section 2 of this draft. Traffic from H3 can reach either of the TORs PE1 or PE2.

On PE2, all the remote MAC-IP routes belonging to the same Ethernet Segment that are advertised by it's respective peers (PE1 in our example) should be synced and installed locally on PE2 but not advertised as local routes by BGP. When the traffic from H3 reaches PE2, it will be able forward the traffic to H1 without any convergence delay caused by triggering ARP/ND. In a scaled setup, the convergence can be significant as the ARP and ND resolution can take

a lot of time. So syncing the IPv4/6 prefixes that belong to same Ethernet Segment helps in solving convergence issues.

3.3 Handling Silent Host

In continuation with the discussion above, if the reachability of PE1 is lost, PE3 will update the ECMP list for H1 to PE2, upon receiving mass withdrawal from PE1. If host H1 is also withdrawn from PE1, then the same route is withdrawn from PE2 and PE3. Hence traffic from H3 to H1 is black-holed till H1 is re-learnt on PE2.

This black-holing can be much worse if the H1 behaves like a silent host. IP address of H1 will not be re-learnt on PE2 till H1 re-ARPs or some traffic triggers ARP for H1.

PE2 can detect the failure of PE1's reachability in following ways

a) When core failure or box reload happens on PE1, next hop reachability to PE1 can be detected by the underlay routing protocols.

b) Upon access failure, PE1 sends withdraws the EAD/ES Route and PE2 can use this as a trigger to detect failure.

Thus to avoid the black-holing, when PE2 detects loss of reachability to PE1, it should trigger ARP/ND for all remote IP prefixes received from it's ES peers (i.e. PE1) belonging to same Ethernet Segment across IP-VRF contexts. This will force host H1 to reply to the solicited ARP/ND from PE2 and refresh both MAC and IP for the corresponding host in its tables.

Even in core failure scenario on PE1, PE1 must withdraw all its local L2 connectivity, as L2 traffic should not be received by PE1. So when ARP/ND is triggered from PE2 the replies from host H1 can only be received by PE2. Thus H1 will be learnt as local route and also advertised from PE2.

It is recommended to have a staggered or delayed deletion of the IP routes from PE1, so that ARP/ND refresh can happen on PE2 before the deletion.

3.4 MAC Aging

PE1 would do ARP/ND refresh for H1 before it ages out. During this process, H1 on can age out genuinely or due to the ARP/ND reply landing on PE2. PE1 must withdraw the local entry from BGP when H1 entry ages out. PE1 deletes the entry from the local forwarding only

when there are no remote synced entries.

4 Determining Reach-ability to Unicast IP Addresses

4.1 Local Learning

The procedures for local learning do not change from [RFC7432].

4.2 Remote Learning

The procedures for remote learning do not change from [RFC7432].

4.2.1 Constructing MAC/IP Address Advertisement

The procedures for constructing MAC/IP Address Advertisement do not change from RFC 7432

4.2.2 Route Resolution

If the ESI field is set to reserved values of 0 or MAX-ESI, the the IP route resolution MUST be based on the MAC-IP route alone.

If the ESI field is set to a non-reserved ESI, the IP route resolution MUST happen only when both the MAC-IP route and the associated set of Ethernet AD per ES routes have been received. To illustrate this with an example, consider a pair of multi-homed TORs PE1 and PE2 connected to an Ethernet Segment. ES1 in an all-active redundancy mode. A given host with IP address H1 is leant by PE1 but not by PE2. When the MAC-IP advertisement route from PE1 and a set of EAD/ES and Layer 3 EAD/EVI routes from PE1 and PE2 are received, PE3 can forward traffic destined to H1 to both PE1 and PE2.

If after (1) PE1 withdraws EAD/ES, then PE3 will forward the said traffic to PE2 only.

If after (1) PE2 withdraws EAD/ES, then PE3 will forward the said traffic to PE1 only.

If after (1) PE1 withdraws the MAC-IP route, then PE3 will do delayed deletion of H1, as described in section 3.3.

If after (1) PE2 advertised the MAC-IP route, but PE1 withdraws it, PE3 will continue forwarding to both PE1 and PE2 as long as it has the EAD/ES and the Layer 3 EAD/EVI route from both.

5 Forwarding Unicast Packets

Please refer to Section 5 in the draft-ietf-bess-evpn-inter-subnet-forwarding-01

6 Load Balancing of Unicast Packets

The procedures for load balancing of Unicast Packets do not change from [RFC7432]

7 Security Considerations

The mechanisms in this document use EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable.

This document uses MPLS and IP-based tunnel technologies to support data plane transport. Security considerations described in [R7432] and in [ietf-evpn-overlay] are equally applicable.

8 IANA Considerations

9 References

9.1 Normative References

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

[RFC1776] Crocker, S., "The Address is the Message", RFC 1776, April 1 1995.

[TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, April 1 1996.

9.2 Informative References

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

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Suresh Pasupula
Cisco
Email: spasupula@cisco.com

Gaurav Badoni
Cisco
Email: gbadoni@cisco.com

Priyanka Warade
Cisco
Email: pwarade@cisco.com