Abstract

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) services, for DC interconnect (DCI) services, and for next generation virtual private LAN services in service provider (SP) applications.

This draft describes how to support efficiently endpoints running IGMP for the above services over an EVPN network by incorporating IGMP proxy procedures on EVPN PEs.
The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright and License 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
1.1 Terminology ............................................ 4
2 IGMP Proxy ................................................. 4
2.1 Proxy Reporting ......................................... 4
  2.1.1 IGMP Membership Report Advertisement in BGP .......... 4
  2.1.1 IGMP Leave Group Advertisement in BGP ................ 6
2.2 Proxy Querier ............................................ 7
3 Operation ................................................... 7
  3.1 PE with only attached hosts/VMs for a given subnet ....... 8
  3.2 PE with mixed of attached hosts/VMs and multicast source .. 9
  3.1 PE with mixed of attached hosts/VMs, multicast source and router ................................................ 9
5 BGP Encoding ............................................... 9
  5.1 Selective Multicast Ethernet Tag Route .................. 9
  5.2 Constructing the Selective Multicast route ............... 11
6 Acknowledgement ........................................... 12
7 Security Considerations .................................... 12
8 IANA Considerations ....................................... 12
9 References .................................................. 12
  9.1 Normative References .................................... 12
  9.2 Informative References ................................... 12
Authors’ Addresses .......................................... 12
1 Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) services, for DC interconnect (DCI) services, and for next generation virtual private LAN services in service provider (SP) applications.

In DC applications, a POD can consist of a collection of servers supported by several TOR and Spine routers. This collection of servers and routers are self contained and may have their own control protocol for intra-POD communication and orchestration. However, EVPN is used as way of standard inter-POD communication for both intra-DC and inter-DC. A subnet can span across multiple PODs and DCs. EVPN provides robust multi-tenant solution with extensive multi-homing capabilities to stretch a subnet (e.g., VLAN) across multiple PODs and DCs. There can be many hosts/VMs (e.g., several hundreds) attached to a subnet that is stretched across several PODs and DCs.

These hosts/VMs express their interests in multicast groups on a given subnet/VLAN by sending IGMP membership reports (Joins) for their interested multicast group(s). Furthermore, an IGMP router (e.g., IGMPv1) periodically sends membership queries to find out if there are hosts on that subnet still interested in receiving multicast traffic for that group. The IGMP/MLD Proxy solution described in this draft has three objectives to accomplish:

1) Just like ARP/ND suppression mechanism in EVPN to reduce the flooding of ARP messages over EVPN, it is also desired to have a mechanism to reduce the flood of IGMP messages (both Queries and Reports) in EVPN.

2) If there is no physical/virtual multicast router attached to the EVPN network for a given (*,G) or (S,G), it is desired for the EVPN network to act as a distributed anycast multicast router for all the hosts attached to that subnet.

3) To forward multicast traffic efficiently over EVPN network such that it only gets forwarded to the PEs that have interest in the multicast group(s) - i.e., multicast traffic will not be forwarded to the PEs that have no receivers attached to them for that multicast group. This draft shows how both of the above objectives are achieved.

The first two objectives are achieved by using IGMP/MLD proxy on the PE and the third objective is achieved by setting up a multicast tunnel (ingress replication or P2MP) only among the PEs that have interest in that multicast group(s) based on the trigger from
IGMP proxy processing. The proposed solutions for each of these objectives are discussed in the following sections.

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 [KEYWORDS].

2 IGMP Proxy

IGMP Proxy mechanism is used to reduce the flooding of IGMP messages over EVPN network similar to ARP proxy used in reducing the flooding of ARP messages over EVPN. It also provided triggering mechanism for the PEs to setup their underlay multicast tunnels. IGMP Proxy mechanism consist of two components: a) Proxy for IGMP Reports and b) Proxy for IGMP Queries.

2.1 Proxy Reporting

When IGMP protocol is used between host/VMs and its first hop EVPN router (EVPN PE), Proxy-reporting is used by the EVPN PE to summarize (when possible) reports received from downstream hosts and propagate it in BGP to other PEs that are interested in the info. This is done by terminating IGMP Reports in the first hop PE, translating and exchanging the relevant information between EVPN BGP speakers. The information is again translated back to IGMP message at the recipient EVPN speaker. Thus it helps create an IGMP overlay subnet using BGP. In order to facilitate such an overlay, this document also defines a new EVPN route type NLRI (EVPN Selective Multicast Ethernet Tag route) along with its procedures to help exchange and register IGMP multicast groups [section 5].

2.1.1 IGMP Membership Report Advertisement in BGP

When a PE wants to advertise an IGMP membership report (Join) using the BGP EVPN route, it follows the following rules:

1) When the first hop PE receives several IGMP membership reports (Joins), belonging to the same IGMP version, from different attached hosts/VMs for the same (*,G) or (S,G), it only sends a single BGP message corresponding to the very first IGMP Join. This is because BGP is a statefull protocol and no further transmission of the same report is needed. If the IGMP Join is for (*,G), then multicast group address along with the corresponding version flag (v1, v2, or v3) are set. In case of IGMPv3, exclude flag also needs to be set to indicate...
that no source IP address to be excluded (e.g., include all sources "*"). If the IGMP Join is for \((S,G)\), then besides setting multicast group address along with the version flag v3, the source IP address and the include/exclude flag must be set. It should be noted that when advertising the EVPN route for \((S,G)\), the only valid version flag is v3 (i.e., v1 and v2 flags must be set to zero).

2) When the first hop PE receives an IGMPv3 Join for \((S,G)\), then the PE checks to see if the source \((S)\) is attached to self. If so, it does not send the corresponding BGP EVPN route advertisement.

3) When the first hop PE receives an IGMP version-X Join first for \((*,G)\) and then later it receives an IGMP version-Y Join for the same \((*,G)\), then it will readvertise the same EVPN Selective Multicast route with flag for version-Y set in addition to any previously-set version flag(s). In other words, the first hop PE does not withdraw the EVPN route before sending the new route because the flag field is not part of BGP route key processing.

4) When the first hop PE receives an IGMP version-X Join first for \((*,G)\) and then later it receives an IGMPv3 Join for the same multicast group address but for a specific source address S, then the PE will readvertise a new EVPN Selective Multicast route with v3 flag set (and v1 and v2 reset). Include/exclude flag also need to be set accordingly. Since source IP address is used as part of BGP route key processing, it is considered as a new BGP route advertisement.

5) When a PE receives an EVPN Selective Multicast route with more than one version flag set, it will generate the corresponding IGMP report for \((*,G)\) for each version specified in the flag field. With multiple version flags set, there should be no source IP address in the receive EVPN route. If there is, then an error should be logged. If v3 flag is set (in addition to v1 or v2), then the include/exclude flag needs to indicate "exclude". If not, then an error should be logged. The PE MUST generate an IGMP membership report (Join) for that \((*,G)\) and each IGMP version in the version flag.

6) When a PE receives a list of EVPN Selective Multicast NLRIs in its BGP update message, each with a different source IP address and the multicast group address, and the version flag is set to v3, then the PE generates an IGMPv3 membership report with a record corresponding to the list of source IP addresses and the group address along with the proper indication of inclusion/exclusion.

7) Upon receiving EVPN Selective Multicast route(s) and before
generating the corresponding IGMP Join(s), the PE checks to see whether it has any multicast router’s AC(s) (Attachment Circuits connected to multicast routers). If it has router’s ACs, then the generated IGMP Join(s) are sent to those ACs. If it doesn’t have any router’s AC, then no IGMP Join(s) needs to be generated because sending IGMP Joins to other hosts can result in unintentionally preventing a host from joining a specific multicast group for IGMPv1 and IGMPv2 – i.e., if the PE does not receive a join from the host it will not forward multicast data to it. Per [RFC4541], when an IGMPv1 or IGMPv2 host receives a membership report for a group address that it intends to join, the host will suppress its own membership report for the same group. This message suppression is a requirement for IGMPv1 and IGMPv2 hosts. This is not a problem for hosts running IGMPv3 because there is no suppression of IGMP Membership reports.

2.1.1 IGMP Leave Group Advertisement in BGP

When a PE wants to withdraw an EVPN Selective Multicast route corresponding to an IGMPv2 Leave Group (Leave) or IGMPv3 "Leave" equivalent message, it follows the following rules:

1) For IGMPv1, there is no explicit membership leave; therefore, the PE needs to periodically send out an IGMP membership query to determine whether there is any host left who is interested in receiving traffic directed to this multicast group. This proxy query function will be described in more details in section 2.2.

2) When a PE receives an IGMPv2 Leave Group or its "Leave" equivalent message for IGMPv3 from its attached host, it checks to see if this host is the last host who is interested in this multicast group by sending a query for the multicast group. If the host was indeed the last one, then the PE re-advertises EVPN Selective Multicast route with the corresponding version flag reset. If this is the last version flag to be reset, then instead of readvertising the EVPN route with all version flags reset, the PE withdraws the EVPN route for that (*,G).

3) When a PE receives an EVPN Selective Multicast route for a given (*,G), it compares the received version flags from the route with its per-PER stored version flags. If the PE finds that a version flag associated with the (*,G) for the remote PE is reset, then the PE generates IGMP Leave for that (*,G) toward its local interface (if any) attached to the multicast router for that multicast group. It also removes the remote PE from the OIF list associated with that multicast group. It should be noted that the received EVPN route should at least have one version flag set. If all version flags are reset, it is an error because the PE should have received an EVPN
route withdraw for the last version flag.

4) If the reset version flag is for version-1 or if the EVPN route withdraw is for version-1, the PE removes the remote PE from its OIF list for that multicast group. If there are no more OIF entries for that multicast group (either locally or remotely), then the PE MUST stop responding to queries from the locally attached router (if any). If there is a source for that multicast group, the PE stops sending multicast traffic for that source.

2.2 Proxy Querier

As mentioned in the previous sections, each PE need to have proxy querier functionality for the following reasons:

1) To enable the collection of EVPN PEs providing L2VPN service to act as distributed multicast router with Anycast IP address for all attached hosts/VMs in that subnet.

2) To enable suppression of IGMP membership reports and queries over MPLS/IP core.

3) To enable generation of query messages locally to their attached host. In case of IGMPv1, the PE needs to send out an IGMP membership query to verify that at least one host on the subnet is still interested in receiving traffic directed to that group. When there is no reply to three consecutive IGMP membership queries, the PE times out the group, stops forwarding multicast traffic to the attached hosts for that (*,G), and sends a EVPN Selective Multicast route associated with that (*,G) with the version-1 flag reset or withdraws that route.

3 Operation

Consider the EVPN network of figure-1, where there is an EVPN instance configured across the PEs shown in this figure (namely PE1, PE2, and PE3). Lets consider that this EVPN instance consist of a single bridge domain (single subnet) with all the hosts, sources and the multicast router shown in this figure connected to this subnet. PE1 only has hosts connected to it. PE2 has a mix of hosts and multicast source. PE3 has a mix of hosts, multicast source, and multicast router. Further more, lets consider that for (S1,G1), R1 is used as the multicast router but for (S2, G2), distributed multicast router with Anycast IP address is used. The following subsections describe the IGMP proxy operation in different PEs with regard to whether the locally attached devices for that subnet are:
- only hosts/VMs
- mix of hosts/VMs and multicast source
- mix of hosts/VMs, multicast source, and multicast router

![Figure 1:](image_url)

### 3.1 PE with only attached hosts/VMs for a given subnet

When PE1 receives an IGMPv1 Join Report from H1, it does not forward this join to any of its other ports (for this subnet) because all these local ports are associated with the hosts/VMs. PE1 sends an EVPN Multicast Group route corresponding to this join for \((*,G1)\) and setting \(v1\) flag. This EVPN route is received by PE2 and PE3 that are the member of the same EVI. PE3 reconstructs IGMPv1 Join Report from this EVPN BGP route and only sends it to the port(s) with multicast routers attached to it (for that subnet). In this example, PE3 sends the reconstructed IGMPv1 Join Report for \((*,G1)\) to only R1. Furthermore, PE2 although receives the EVPN BGP route, it does not send it to any of its port for that subnet – namely ports associated with H6 and H7.

When PE1 receives the second IGMPv1 Join from H2 for the same multicast group \((*,G1)\), it only adds that port to its OIF list but it doesn’t send any EVPN BGP route because there is no change in information. However, when it receives the IGMPv2 Join from H3 for
the same (*,G1), besides adding the corresponding port to its OIF list, it re-advertises the previously sent EVPN Selective Multicast route with the version-2 flag set.

Finally when PE1 receives the IMGMPv3 Join from H4 for (S2,G2), it advertises a new EVPN Selective Multicast route corresponding to it.

3.2 PE with mixed of attached hosts/VMs and multicast source

The main difference in here is that when PE2 receives IGMPv3 Join from H7 for (S2,G2), it does not advertises it in BGP because PE2 knows that S2 is attached to its local AC. PE2 adds the port associated with H7 to its OIF list for (S2,G2). The processing for IGMPv2 received from H6 is the same as the v2 Join described in previous section.

3.1 PE with mixed of attached hosts/VMs, multicast source and router

The main difference in here relative to the previous two sections is that Join messages received locally needs to be sent to the port associated with router R1. Furthermore, the Joins received via BGP need to be passed to the R1 port but filtered for all other ports.

5 BGP Encoding

This document defines a new BGP EVPN route to carry IGMP membership reports. This route type is known as:

+ 6 - Selective Multicast Ethernet Tag Route

The detailed encoding and procedures for this route type is described in subsequent section.

5.1 Selective Multicast Ethernet Tag Route

An Selective Multicast Ethernet Tag route type specific EVPN NLRI consists of the following:
For the purpose of BGP route key processing, all the fields are considered to be part of the prefix in the NLRI except for the one-octet optional flag field (if included). The Flags fields are defined as follows:

```
0 1 2 3 4 5 6 7
+---+---+---+---+---+---+---+
|  reserved  |IE|v3|v2|v1|
+---+---+---+---+---+---+
```

The least significant bit, bit 7 indicates support for IGMP version 1.

The second least significant bit, bit 6 indicates support for IGMP version 2.

The third least significant bit, bit 5 indicates support for IGMP version 3.

The forth least significant bit, bit 4 indicates whether the (S, G) information carried within the route-type is of Include Group type (bit value 0) or an Exclude Group type (bit value 1). The Exclude Group type bit MUST be ignored if bit 5 is not set.

This EVPN route type is used to carry tenant IGMP multicast group information. The flag field assists in distributing IGMP membership interest of a given host/VM for a given multicast route. The version
bits help associate IGMP version of receivers participating within the EVPN domain.

The include/exclude bit helps in creating filters for a given multicast route.

5.2 Constructing the Selective Multicast route

This section describes the procedures used to construct the Selective Multicast route. Support for this route type is optional.

The Route Distinguisher (RD) SHOULD be a Type 1 RD [RFC4364]. The value field comprises an IP address of the PE (typically, the loopback address) followed by a number unique to the PE.

The Ethernet Tag ID MUST to zero for VLAN-based service and to a valid normalized VID for VLAN-aware bundle service.

The Multicast Source length MUST be set to length of multicast source address in bits. In case of a (*, G) Join, the Multicast Source Length is set to 0.

The Multicast Source is the Source IP address of the IGMP membership report. In case of a (*, G) Join, this field does not exist.

The Multicast Group length MUST be set to length of multicast group address in bits.

The Multicast Group is the Group address of the IGMP membership report.

The Originator Router Length is the length of the Originator Router address in bits.

The Originator Router Address is the IP address of Router Originating the prefix.

The Flags field indicates the version of IGMP protocol from which the membership report was received. It also indicates whether the multicast group had INCLUDE or EXCLUDE bit set.

IGMP protocol is used to receive group membership information from hosts/VMs by TORs. Upon receiving the hosts/VMs expression of interest of a particular group membership, this information is then forwarded to BGP EVPN using Ethernet Multicast Source Group Route NLRI. The NLRI also keeps track of receiver’s IGMP protocol version and any “source filtering” for a given group membership. All Ethernet Multicast Source Group Routes are announced with ES-Import Route...
Target extended communities.

6 Acknowledgement

7 Security Considerations

Same security considerations as [RFC7432].

8 IANA Considerations

Allocation of Extended Community Type and Sub-Type for EVPN.

9 References

9.1 Normative References


9.2 Informative References


Authors’ Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com