BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: September 6, 2018

S. Mohanty
M. Ghosh
Cisco Systems
S. Breeze
Claranet
J. Uttaro
ATT

March 5, 2018

BGP EVPN Flood Traffic Optimization
draft-mohanty-bess-evpn-bum-opt-00

Abstract

In EVPN, the Broadcast, Unknown Unicast and Multicast (BUM) traffic is sent to all the routers participating in the EVPN instance. In a multi-homing scenario, when more than one PEs share the same Ethernet Segment, i.e. there are more than one PEs in a redundancy group, only the PE that is the Designated-Forwarder (DF) for the ES will forward that packet on the access interface whereas all non-DF PEs will drop the packet. From the perspective of the network, this is quite wasteful. This is especially true if there are significantly more PEs on the Ethernet Segment. This draft explores this problem and provides a solution for the same.

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 September 6, 2018.

Copyright Notice

Table of Contents

1. Requirements Language and Terminology 2
2. Introduction 2
3. Problem Description 4
4. Solution 1. Suppress the advertisement of the IMET route . . 5
5. Solution 2. Advertisement of the IMET route from the BDF . . 6
6. Protocol Considerations 6
7. Operational Considerations 7
8. Security Considerations 7
9. Acknowledgements 7
10. Contributors 7
11. References 7
11.1. Normative References 7
11.2. Informative References 8
Authors' Addresses 8

1. Requirements Language and 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 [RFC2119].

- ES: Ethernet Segment
- EVI: Ethernet virtual Instance, this is a mac-vrf.
- IMET: Inclusive Multicast Route
- DF: Designated Forwarder
- BDF: Backup Designated Forwarder

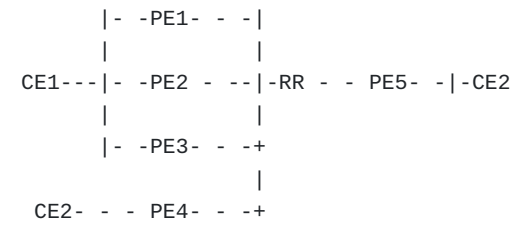
2. Introduction

BGP [RFC7432] describes a solution for disseminating mac addresses over an mpls core via the Border Gateway Protocol. In EVPN, data plane learning is confined to the access, and the control plane

learning happens via BGP in the core. This prevents unnecessary flooding in the data plane as the traffic is directed to where the destination is learnt from. However, in case of Broadcast, Unknown Unicast and Multicast (BUM) traffic, the PE needs to do a flooding to all the other PEs in the domain.

PEs elect a Designated Forwarder (DF) amongst themselves, for a given ES, by exchanging type-4 routes via BGP. The role of a DF is to forward BUM traffic received from the core, towards its access facing interface. A PE in a non-DF role will drop flood traffic received on its core-facing interface. Note that the DF election process is only confined to the set of PEs who host the same Ethernet Segment. Remote PEs are not interested in type-4 routes for Ethernet Segments that they do not host. Hence remote PEs are ignorant of the DFs for segments which is not local to them. Consequently, when the remote PE needs to do a BUM flooding using ingress replication, it will flood the frames to all participating PEs, irrespective of whether DFs or not. The key to creating a list of PEs with which to flood to, is the Inclusive multicast ethernet tag route which is described below.

The IMET route (type-3) in EVPN advertises the BUM label for the EVI to all the other PEs who are interested in the same EVI. For ingress replication the label is encapsulated in the PMSI attribute. The label is used to encapsulate the BUM traffic at the ingress entity. This label is inserted just above the split-horizon label in the BUM frame. When the BUM packet is received by a PE that is multi-homed to the same Ethernet segment as the PE that originated the BUM packet, and, is the DF for that (EVI, ES) pair, after popping the transport label, the receiving PE is going to check if the split-horizon label is its own. If so, it will drop the packet if no other ES is configured. Otherwise it will forward the frame on all other Segments that are part of the same EVI. if the PE is not the DF, it will straightaway drop the packet immediately.



An EVPN Network

Figure 1

3. Problem Description

In the Figure 1. above, PE1, P2 and PE3 are all multi-homed to CE1 on the same Ethernet Segment, say ES1. PE4 has a single host which is not multi-homed. The same EVPN instance (Bridge-Domain) exists on all the PEs. For this EVPN instance, PE1 is the Designated Forwarder on ES1. Also, PE3 is the backup DF [I-D.ietf-bess-evpn-df-election]. When PE5 sends the BUM traffic, the flooded frames are received by PE1, PE2, PE3 and PE4. PE1 is going to forward the flood traffic on its access link towards CE1. PE2 and PE3 will drop the flooded frames that they receive from the core. PE4 will forward it as it has a single-homed host on the same EVPN instance.

Here it is wasteful for PE2 and PE3 to receive the flooded frames. Whilst the majority of deployments usually have two PEs as part of the redundancy group, in some cases, there may be more than two PEs on the same ES. An example being when capacity demands of the PE are close to the hardware limits of the PE. In this scenario, operators may chose to protect their investments and increase their resilience by installing additional PEs, instead of replacing them or further segmenting the access network. Further,increasing the number of PEs results in efficient load-balancing across vlans.

We can now formally describe the issue. In general, consider an EVPN instance, EVIi, that exists in a PE, say PEk. As per existing EVPN behavior, even If PEk is not the DF for any of its Ethernet Segments (that are multi-homed to other PEs) and also there are no other single-homed Ethernet Segments that are part of EVIi in PEk , PEk will still receive BUM traffic meant for EVIi from a remote PE, PEj. This traffic is simply dropped as PEk is not a DF for any of these Ethernet Segments.

1. This is an unnecessary usage of bandwidth in the EVPN Core.

2. PEk receives traffic which it drops which is non-optimal usage of the L2 Forwarding engine.

3. PEj replicates a copy of the Ethernet Frame to PEk which is anyway to be dropped. This consumes cycles at PEj.

In this draft we address the above problem and give two simple solutions for the same. These solutions do not mandate any protocol changes and are backwards compatible.

4. Solution 1. Suppress the advertisement of the IMET route

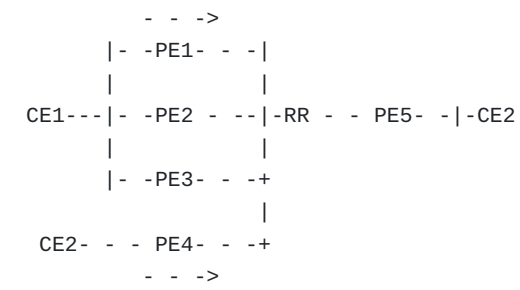
The first solution is for a PE not to advertise the IMET route if the outcome is to drop the flooded traffic

- PEk only needs to advertise "Inclusive Multicast Ethernet Tag route" (Type-3 route) for an EVPN Instance, EVIi if and only if EVIi is configured on at least one Ethernet Segment (which also has a presence in another PEj, i.e Multihomed) and PEk is the DF for that specific Ethernet Segment.

- The Type-3 SHOULD also be advertised if there is a "Single-Home" Ethernet Segment on an EVI.

- Where a PE is the first DF for an ES on an EVPN Instance, the IMET should be advertised, whereas on the Last DF to Non-DF transition, it should be withdrawn.

In the Figure 2 the same EVPN instance exists in PE1, PE2, PE3, PE4 and pE5. But only PE1 and PE4 advertise the IMET route. So PE5 sends the flood traffic to PE1 and PE4 only.



An EVPN Network

Figure 2

With this approach, on a DF PE (PE1) failure, BUM traffic will be dropped until the IMET from the next elected DF [PE2 or PE3], is received at PE5. Note, present behaviour is that BUM is also dropped based on route type 4 withdraw in the peering PEs. In comparison of this proposal with the existing methods, convergence delay will be MAX[Type 4, Type 3 Propagation delays] after the New DF is elected. This leads to our next solution extension, where convergence cannot be traded off over bandwidth optimization.

5. Solution 2. Advertisement of the IMET route from the BDF

1. Multihomed PEs can easily compute the Backup DF, based on the DF election mode in operation.

2. Extending Solution 1, we are proposing that a PE should only advertise Type-3 for an EVI if and only if one of the conditions hold:

    * It has an Single Home Ethernet Segment, in the EVI

    * It is DF for at least one Ethernet-Segment, for that EVI

    * It is BDF for at least one Ethernet-Segment, for that EVI

This would mean that, in Fig. 2, in addition to the IMET routes that are being advertised from PE1 and PE4, PE3 also advertises the IMET route since it is the BDF. It can be seen from the above example that with increasing number of multi-homed PEs sharing the same Ethernet-Segment and Vlans, only two PEs will advertise IMET on behalf of an EVI. Of course, if there are some single-homed hosts, there may be some additional IMET advertisements. But the real benefits are in the data plane since this results in no BUM traffic for PEs that do not need it; but would have, nevertheless, got it, as per the existing EVPN procedures.

6. Protocol Considerations

This idea conforms to existing EVPN drafts that deal with BUM handling [RFC7432], and [I-D.ietf-bess-evpn-igmp-mld-proxy]. Additionally, to take DF Type 4 as explained in [I-D.sajassi-bess-evpn-per-mcast-flow-df-election] into consideration, along the other conditions specified in Sections 4 and 5, the PE should advertise IMET if and only if there is at least one (S,G) for which it is DF. For all other DF Types, no additional considerations are required.

7. Operational Considerations

None

8. Security Considerations

This document raises no new security issues for EVPN.

9. Acknowledgements

The authors would like to thank Ali Sajassi for his feedback and insight into the deployments that can benefit from this proposal.

10. Contributors

Samir Thoria
Cisco Systems
US

Email: sthoria@cisco.com

Sameer Gulrajani
Cisco Systems
US

Email: sameerg@cisco.com

11. References

11.1. Normative References

[I-D.ietf-bess-evpn-df-election]
satyamoh@cisco.com, s., Patel, K., Sajassi, A., Drake, J., and T. Przygienda, "A new Designated Forwarder Election for the EVPN", draft-ietf-bess-evpn-df-election-03 (work in progress), October 2017.

[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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[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>.

11.2. Informative References

[I-D.ietf-bess-evpn-igmp-mld-proxy]
Sajassi, A., Thoria, S., Patel, K., Yeung, D., Drake, J., and W. Lin, "IGMP and MLD Proxy for EVPN", draft-ietf-bess-evpn-igmp-mld-proxy-00 (work in progress), March 2017.

[I-D.sajassi-bess-evpn-per-mcast-flow-df-election]
Sajassi, A., mishra, m., Thoria, S., Rabadan, J., and J. Drake, "Per multicast flow Designated Forwarder Election for EVPN", draft-sajassi-bess-evpn-per-mcast-flow-df-election-00 (work in progress), March 2018.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Mrinmoy Ghosh
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: mrghosh@cisco.com

Sandy Breeze
Claranet
University of Warwick
United Kingdom

Email: sandy.breeze@eu.clara.net

Jim Uttaro
ATT
200 S. Laurel Avenue
Middletown, CA 07748
USA

Email: uttaro@att.com