

BESS WorkGroup
Internet-Draft
Intended status: Informational
Expires: May 5, 2020

S. Mohanty
M. Ghosh
A. Sajassi
Cisco Systems
S. Breeze
Claranet
J. Uttaro
ATT
November 2, 2019

BGP EVPN Flood Traffic Optimization at EVPN Gateways
draft-mohanty-bess-evpn-bum-opt-01

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. In deployments such as EVPN Gateways (EVPN GW) or Data Center Interconnect (DCI) routers, this can be quite wasteful. This is especially true if there are significantly more EVPN GW or DCI PEs all participating in the same sets of ES and vES. This draft explores the problem and provides solutions 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 May 5, 2020.

Copyright Notice

Copyright (c) 2019 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 (<https://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.	Requirements Language and Terminology	2
2.	Introduction	3
3.	Problem Description	4
4.	Solutions	5
4.1.	DF Election per-mcast-flow	5
4.2.	Suppress the advertisement of the IMET route	5
4.3.	Advertisement of the IMET route from the BDF	7
5.	Protocol Considerations	7
6.	Operational Considerations	8
7.	Security Considerations	8
8.	Acknowledgements	8
9.	Contributors	8
10.	References	8
10.1.	Normative References	8
10.2.	Informative References	9
	Authors' Addresses	9

[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](#)].

- o ES: Ethernet Segment
- o vES: Virtual Ethernet Segment
- o EVI: Ethernet virtual Instance, this is a mac-vrf.
- o IMET: Inclusive Multicast Route

- o DF: Designated Forwarder
- o BDF: Backup Designated Forwarder
- o DCI: Data Center Interconnect Router

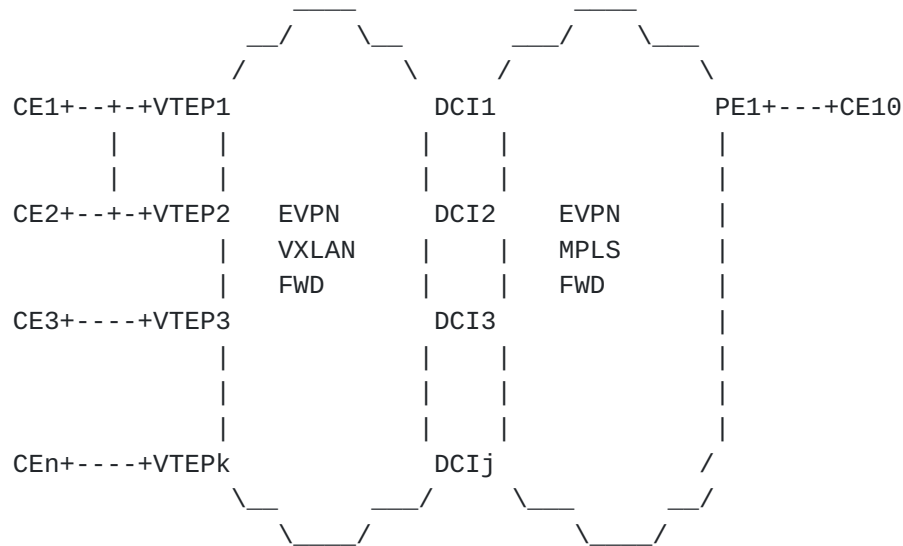
2. Introduction

EVPN [[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 drop the packet immediately.



An EVPN Datacenter network with VXLAN forwarding joined to a traditional EVPN network with MPLS forwarding. Adjoining DCI routers are said to be EVPN GW's. A DCI will have a single vES (ESI) per BD, with multiple VTEP next-hops.

Figure 1

3. Problem Description

In the Figure 1. above, DCI1, DCI2 and DCI3 are all multi-homed EVPN GW's for multiple VTEPs serving the same vES, say vES1. PE1 has a single host which is not multi-homed.

The same EVPN instance (Bridge-Domain) exists on all the PEs and DCIs. For this EVPN instance, DCI1 is the Designated Forwarder on vES1 and DCI2 is the backup DF [[RFC8584](#)]. When PE1 sends the BUM traffic, the flooded frames are received by DCI1, DCI2, DCI3 up to DCIj. DCI1 is going to forward the flood traffic on its vES towards all VTEPs participating in vES1. DCI2, DCI3 and all DCIs up to DCIj will drop the flooded frames that they receive from the core.

Here it is wasteful for DCI2, DCI3 and DCIj to receive the flooded frames. Whilst the majority of deployments usually have two DCIs as part of the redundancy group, in some cases, there may be more than two on the same vES. An example being when capacity demands of the DCI are close to the hardware limits of the DCI. In this scenario, operators may chose to protect their investments and increase their resilience by installing additional DCIs, instead of replacing them or further segmenting the datacenter network. Further, increasing

the number of DCIs results in more efficient load-balancing across VNIs.

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

1. This is an unnecessary usage of bandwidth in the EVPN Core.
2. DCIj receives traffic which it drops which is non-optimal usage of the L2 Forwarding engine.
3. PEk replicates a copy of the Ethernet Frame to DCIj which is only to be dropped. This consumes cycles at PEk.

In this draft we address the above problem and give possible solutions.

4. Solutions

4.1. DF Election per-mcast-flow

Solving the bandwidth in the EVPN core is an operators primary concern. Given the majority of traffic volume in BUM comes from large multicast flows, adopting the mechanisms described in ["I-D.draft-ietf-bess-evpn-per-mcast-flow-df-election-00"](#) not only improves the distribution of multicast traffic amongst DCI1...DCIj for a given vES, techniques such as not advertising the SMET from a non-DF DCI ensure that only DCIs who've won the election for the group, receive multicast traffic for the group.

This solution explicitly requires IGMP snooping in the BD where the vES resides.

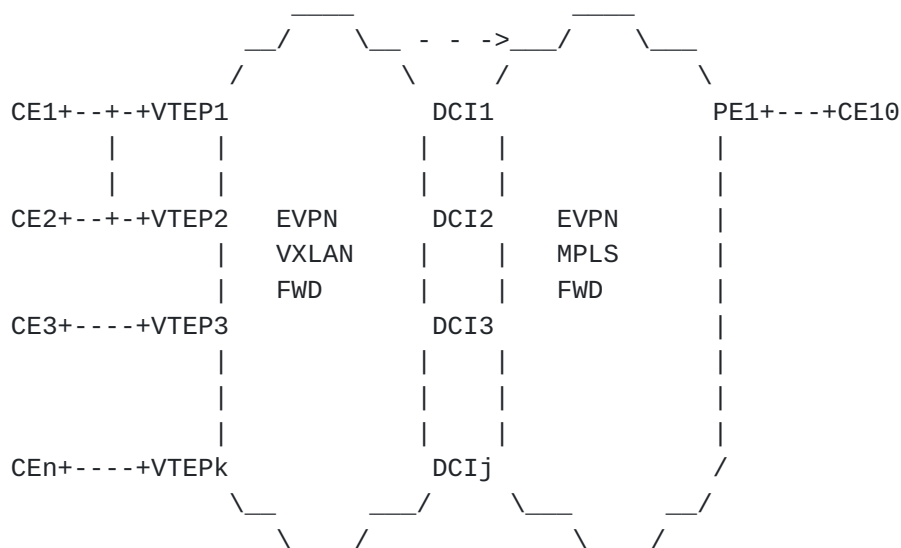
This solution does not solve the problem of unnecessary Broadcast and Unknown Unicast being replicated to nDFs, but it solves the most prominent problem of bandwidth.

4.2. Suppress the advertisement of the IMET route

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

- o DCIj 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 DCI, i.e Multihomed) and DCIj is the DF for that specific Ethernet Segment.
- o The Type-3 SHOULD also be advertised if there is a "Single-Home" Ethernet Segment on an EVI.
- o Where a DCI is the first DF for an vES 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 DCI1, DCI2, DCI3, DCIj and PE1. However, only DCI1 and PE1 advertise the IMET route. So PE1 sends the flood traffic to DCI1 only.



An EVPN GW Network

Figure 2

With this approach, on a DF DCI1 failure, BUM traffic will be dropped until the IMET from the next elected DF [DCI2 through DCIj] is received at PE1. Note however; 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.

4.3. 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 the previous solution, 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 ES or vES, for that EVI
 - * It is BDF for at least one ES or vES, for that EVI

This would mean that, in Fig. 2, in addition to the IMET routes that are being advertised from DCI1, DCI2 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 vESs, only two DCIs 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 DCIs that do not need it; but would have, nevertheless, got it, as per the existing EVPN procedures.

It is important to note that the solutions involving suppression of IMET should be limited to the following use case caveats;

1. BUM traffic for Ingress Replication (IR) cases
2. BDs with no igmp/mld/pim proxy
3. BDs with no OISM or IRBs
4. BDs with vES associated to overlay tunnels and no other ACs

With these caveats, the suppression of IMET at non DF or BDF EVPN GWs provide complete control over BUM traffic distribution per-vES (per-BD).

5. 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.draft-ietf-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.

6. Operational Considerations

None

7. Security Considerations

This document raises no new security issues for EVPN.

8. Acknowledgements

The authors would like to thank Jorge Rabadan, John Drake and Eric Rosen for discussions related to this draft.

9. Contributors

Samir Thoria
Cisco Systems
US

Email: sthoria@cisco.com

Sameer Gulrajani
Cisco Systems
US

Email: sameerg@cisco.com

10. References

10.1. Normative References

- [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>>.
- [RFC8584] Rabadan, J., Ed., Mohanty, R., Sajassi, N., Drake, A., Nagaraj, K., and S. Sathappan, "BGP MPLS-Based Ethernet VPN", [RFC 8584](#), DOI 10.17487/RFC8584, April 2019, <<https://www.rfc-editor.org/info/rfc8584>>.

10.2. 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-04](#) (work in progress), September 2019.
- [I-D.ietf-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-ietf-bess-evpn-per-mcast-flow-df-election-01](#) (work in progress), March 2019.
- [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

Ali Sajassi
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: sajassi@cisco.com

Sandy Breeze
Claranet
21 Southampton Row
London WC1B 5HA
United Kingdom

Email: sandy.breeze@eu.clara.net

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

Email: uttaro@att.com

