BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: May 19, 2020

Ali. Sajassi
Mankamana. Mishra
Samir. Thoria
Cisco Systems
Jorge. Rabadan
Nokia
John. Drake
Juniper Networks

November 16, 2019

Per multicast flow Designated Forwarder Election for EVPN
draft-ietf-bess-evpn-per-mcast-flow-df-election-02

Abstract

[RFC7432] describes mechanism to elect designated forwarder (DF) at the granularity of (ESI, EVI) which is per VLAN (or per group of VLANs in case of VLAN bundle or VLAN-aware bundle service). However, the current level of granularity of per-VLAN is not adequate for some applications.[I-D.ietf-bess-evpn-df-election-framework] improves base line DF election by introducing HRW DF election. [I-D.ietf-bess-evpn-igmp-mld-proxy] introduces applicability of EVPN to Multicast flows, routes to sync them and a default DF election. This document is an extension to HRW base draft [I-D.ietf-bess-evpn-df-election-framework] and further enhances HRW algorithm for the Multicast flows to do DF election at the granularity of (ESI, VLAN, Mcast flow).

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

EVPN based All-Active multi-homing is becoming the basic building block for providing redundancy in next generation data center deployments as well as service provider access/aggregation networks. [RFC7432] defines the role of a designated forwarder as the node in the redundancy group that is responsible to forward Broadcast, Unknown unicast, Multicast (BUM) traffic on that Ethernet Segment (CE device or network) in All-Active multi-homing.

The default DF election mechanism allows selecting a DF at the granularity of (ES, VLAN) or (ES, VLAN bundle) for BUM traffic. While [I-D.ietf-bess-evpn-df-election-framework] improve on the default DF election procedure, some service provider residential applications require a finer granularity, where whole multicast flows are delivered on a single VLAN.

```
                               (Multicast sources)
                                        |
                                        |
                                      +---+
                                      |CE4|
                                      +---+
                                        |
                                        |
                                  +-----+-----+
                     +------------|   PE-1    |------------+
                     |            |           |            |
                     |            +-----------+            |
                     |                                     |
                     |                   EVPN              |
                     |                                     |
                     |                                     |
                     | (DF)                           (NDF)|
               +-----------+                        +-----------+
               |  |EVI-1|  |                        |  |EVI-1|  |
               |   PE-2    |------------------------|   PE-3    |
               +-----------+                        +-----------+
                      AC1  \                       / AC2
                            \                     /
                             \      ESI-1        /
                              \                 /
                               \               /
                               +---------------+
                               |    CE2        |
                               +---------------+
                                      |
                                      |
                             (Multiple receivers)
```

Figure 1: Multi-homing Network of EVPN
for IPTV deployments

Consider the above topology, which shows a typical residential deployment scenario, where multiple receivers are behind an all-active multihoming segments. All of the multicast traffic is provisioned on EVI-1. Assume PE-2 get elected as DF. According to [RFC7432], PE-2 will be responsible for forwarding multicast traffic to that Ethernet segment.

o Forcing sole data plane forwarding responsibility on PE-2 is a limitation in the current DF election mechanism. The topology at Figure 1 would always have only one of the PE to be elected as DF irrespective of which current DF election mechanism is in use

defined in [RFC7432] or [I-D.ietf-bess-evpn-df-election-framework].

- The problem may also manifest itself in a different way. For example, AC1 happens to use 80% of its available bandwidth to forward unicast data. And now there is need to serve multicast receivers where it would require more than 20% of AC1 bandwidth. In this case, AC1 becomes oversubscribed and multicast traffic drop would be observed even though there is already another link (AC2) present in network which can be used more efficiently load balance the multicast traffic.

In this document, we propose an extension to the HRW base draft to allow DF election at the granularity of (ESI, VLAN, Mcast flow) which would allow multicast flows to be better distributed among redundancy group PEs to share the load.

2. 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] .

With respect to EVPN, this document follows the terminology that has been defined in [RFC7432] and [RFC4601] for multicast terminology.

3. The DF Election Extended Community

[I-D.ietf-bess-evpn-df-election-framework] defines an extended community, which would be used for PEs in redundancy group to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs in redundancy group about its willingness to support Per multicast flow base DF election capability by signaling a DF election extended community along with Ethernet-Segment Route (Type-4). The current proposal extends the existing extended community defined in [I-D.ietf-bess-evpn-df-election-framework]. This draft defines new a DF type.

- DF type (1 octet) - Encodes the DF Election algorithm values (between 0 and 255) that the advertising PE desires to use for the ES.

  * Type 0: Default DF Election algorithm, or modulus-based algorithms in [RFC7432].

  * Type 1: HRW algorithm defined in [I-D.ietf-bess-evpn-df-election-framework]

  - Type 2: Handshake defines in [I-D.ietf-bess-evpn-fast-df-recovery]

  - Type 3: Time-Synch defined in [I-D.ietf-bess-evpn-fast-df-recovery]

  - Type 4: HRW base per (S,G) multicast flow DF election (explained in this document)

  - Type 5: HRW base per (*,G) multicast flow DF election (explained in this document)

  - Type 6 - 254: Unassigned

  - Type 255: Reserved for Experimental Use.

- The [I-D.ietf-bess-evpn-df-election-framework] describes encoding of capabilities associated to the DF election algorithm using Bitmap field. When these capabilities bits are set along with the DF type-4 and type-5, they need to be interpreted in context of this new DF type-4 and type-5. For example, consider a scenario where all PEs in the same redundancy group (same ES) can support both AC-DF, DF type-4 and DF type-5 and receive such indications from the other PEs in the ES. In this scenario, if a VLAN is not active in a PE, then the DF election procedure on all PEs in the ES should factor that in and exclude that PE in the DF election per multicast flow.

- A PE SHOULD attach the DF election Extended Community to ES route and Extended Community MUST be sent if the ES is locally configured for DF type Per Multicast flow DF election. Only one DF Election Extended community can be sent along with an ES route.

- When a PE receives the ES Routes from all the other PEs for the ES, it checks if all of other PEs have advertised their desire to proceed by Per multicast flow DF election. If all peering PEs have done so, it performs DF election based on Per multicast flow procedure. But if:

  - There is at least one PE which advertised route-4 (AD per ES Route) which does not indicate its capability to perform Per multicast flow DF election. OR

  - There is at least one PE signaling single active in the AD per ES route

it MUST be considered as an indication to support of only Default DF election [RFC7432] and DF election procedure in [RFC7432] MUST be used.

4. HRW base per multicast flow EVPN DF election

This document is an extension of [I-D.ietf-bess-evpn-df-election-framework], so this draft does not repeat the description of HRW algorithm itself.

EVPN PE does the discovery of redundancy groups based on [RFC7432]. If redundancy group consists of N peering EVPN PE nodes, after the discovery all PEs build an unordered list of IP address of all the nodes in the redundancy group. The procedure defined in this draft does not require the list of PEs to be ordered. Address [i] denotes the IP address of the [i]th EVPN PE in redundancy group where (0 < i <= N).

4.1. DF election for IGMP (S,G) membership request

The DF is the PE who has maximum weight for (S, G, V, Es) where

- S - Multicast Source
- G - Multicast Group
- V - VLAN ID.
- Es - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

1. Weight

   * The weight of PE(i) to (S,G,VLAN ID, Es) is calculated by function, weight (S,G,V, Es, Address(i)), where (0 < i <= N), PE(i) is the PE at ordinal i.

   * Weight (S,G,V, Es, Address(i)) = (1103515245. ((1103515245.Address(i) + 12345) XOR D(S,G,V,ESI))+12345) (mod 2^31)

   * In case of tie, the PE whose IP address is numerically least is chosen.

2. Digest

- D(S,G,V, Es) = CRC_32(S,G,V, Es)

- Here D(S,G,V,Es) is the 31-bit digest (CRC_32 and discarding the MSB) of the Source IP, Group IP, Vlan ID and Es. The CRC MUST proceed as if the architecture is in network byte order (big-endian).

4.2. DF election for IGMP (*,G) membership request

The DF is the PE who has maximum weight for (G, V, Es) where

o G - Multicast Group

o V - VLAN ID.

o Es - Ethernet Segment Identifier

Address[i] is address of the ith PE. The PEs IP address length does not matter as only the lower-order 31 bits are modulo significant.

1. Weight

   - The weight of PE(i) to (G,VLAN ID, Es) is calculated by function, weight (G,V, Es, Address(i)), where (0 < i <= N), PE(i) is the PE at ordinal i.

   - Weight (G,V, Es, Address(i)) = (1103515245. ((1103515245.Address(i) + 12345) XOR D(G,V,ESI))+12345) (mod 2^31)

   - In case of tie, the PE whose IP address is numerically least is chosen.

2. Digest

   - D(G,V, Es) = CRC_32(G,V, Es)

   - Here D(G,V,Es) is the 31-bit digest (CRC_32 and discarding the MSB) of the Group IP, Vlan ID and Es. The CRC MUST proceed as if the architecture is in network byte order (big-endian).

4.3. Default DF election procedure

Per multicast DF election procedure would be applicable only when host behind Attachment Circuit (of the Es) start sending IGMP membership requests. Membership requests are synced using procedure defined in [I-D.ietf-bess-evpn-igmp-mld-proxy], and each of the PE in redundancy group can use per flow DF election and create DF state per

multicast flow. The HRW DF election "Type 1" procedure defined in [I-D.ietf-bess-evpn-df-election-framework] MUST be used for the Es DF election and SHOULD be performed on Es even before learning multicast membership request state. This default election procedure MUST be used at port level but will be overwritten by Per flow DF election as and when new membership request state are learnt.

5. Procedure to use per multicast flow DF election algorithm

```
                                     Multicast  Source
                                             |
                                             |
                                             |
                                             |
                                         +---------+
                          +--------------+  PE-4   +--------------+
                          |              |         |              |
                          |              +---------+              |
                          |                                       |
                          |              EVPN CORE                |
                          |                                       |
                          |                                       |
                          |                                       |
                      +---------+        +---------+         +---------+
                      |  PE-1   +--------+   PE-2  +---------+   PE-3  |
                      |  EVI-1  |        |  EVI-1  |         | EVI-1   |
                      +---------+        +---------+         +---------+
                           |__________________|___________________|
                         AC-1    ESI-1        | AC-2               AC-3
                                         +---------+
                                         |  CE-1   |
                                         |         |
                                         +---------+
                                              |
                                              |
                                              |
                                              |
                                      Multicast Receivers
```

Figure-2 : Multihomed network

Figure-2 shows multihomed network. Where EVPN PE-1, PE-2, PE-3 are multihomed to CE-1. Multiple multicast receivers are behind all active multihoming segment.

1. PEs connected to the same Ethernet segment can automatically discover each other through exchange of the Ethernet Segment

Route. This draft does not change any of this procedure, it still uses the procedure defined in [RFC7432].

2. Each of the PEs in redundancy group advertise Ethernet segment route with extended community indicating their ability to participate in per multicast flow DF election procedure. Since Per multicast flow would not be applicable unless PE learns about membership request from receiver, there is a need to have the default DF election among PEs in redundancy group for BUM traffic. Until multicast membership state are learnt, we use the the DF election procedure in Section 4.3, namely HRW per (v,Es) as defined in [I-D.ietf-bess-evpn-df-election-framework] .

3. When a receiver starts sending membership requests for (s1,g1), where s1 is multicast source address and g1 is multicast group address, CE-1 could hash membership request (IGMP join) to any of the PEs in redundancy group. Let's consider it is hashed to PE-2. [I-D.ietf-bess-evpn-igmp-mld-proxy] defines a procedure to sync IGMP join state among redundancy group of PEs. Now each of the PE would have information about membership request (s1,g1) and each of them run DF election procedure Section 4.1 to elect DF among participating PEs in redundancy group. Consider PE-2 gets elected as DF for multicast flow (s1,g1).

    1. PE-1 forwarding state would be nDF for flow (s1,g1) and DF for rest other BUM traffic.

    2. PE-2 forwarding state would be DF for flow (s1,g1) and nDF for rest other BUM traffic.

    3. PE-3 forwarding state would be nDF for flow (s1,g1) and rest other BUM traffic.

4. As and when new multicast membership request comes, same procedure as above would continue.

5. If Section 3 has DF type 4, For membership request (S,G) it MUST use Section 4.1 to elect DF among participating PEs. And membership request (*,G) MUST use Section 4.2 to elect DF among participating PEs.

6. Triggers for DF re-election

There are multiple triggers which can cause DF re-election. Some of the triggers could be

1. Local ES going down due to physical failure or configuration change triggers DF re-election at peering PE.

2. Detection of new PE through ES route.

3. AC going up / down

4. ESI change

5. Remote PE removed / Down

6. Local configuration change of DF election Type and peering PE consensus on new DF Type

This document does not provide any new mechanism to handle DF re-election procedure. It uses the existing mechanism defined in [RFC7432]. Whenever either of the triggers occur, a DF re-election would be done. and all of the flows would be redistributed among existing PEs in redundancy group for ES.

7. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

8. IANA Considerations

Allocation of DF type in DF extended community for EVPN.

9. Acknowledgement

Authors would like to acknowledge helpful comments and contributions of Luc Andre Burdet.

10. Normative References

[HRW1999] IEEE, "Using name-based mappings to increase hit rates", IEEE HRW, February 1998.

[I-D.ietf-bess-evpn-df-election-framework]
Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for EVPN Designated Forwarder Election Extensibility", draft-ietf-bess-evpn-df-election-framework-03 (work in progress), May 2018.

[I-D.ietf-bess-evpn-fast-df-recovery]
Sajassi, A., Badoni, G., Rao, D., Brissette, P., Drake, J., and J. Rabadan, "Fast Recovery for EVPN DF Election", draft-ietf-bess-evpn-fast-df-recovery-00 (work in progress), June 2018.

[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.

[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>.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <https://www.rfc-editor.org/info/rfc4601>.

[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>.

Authors' Addresses

Ali Sajassi
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sajassi@cisco.com

Mankamana Mishra
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: mankamis@cisco.com

Samir Thoria
Cisco Systems
821 Alder Drive,
MILPITAS, CALIFORNIA 95035
UNITED STATES

Email: sthoria@cisco.com

Jorge Rabadan
Nokia
777 E. Middlefield Road
Mountain View, CA 94043
UNITED STATES

Email: jorge.rabadan@nokia.com

John Drake
Juniper Networks

Email: jdrake@juniper.net