BESS WorkGroup
Internet-Draft
Intended status: Standards Track
Expires: September 4, 2018

Ali. Sajassi
Mankamana. Mishra
Samir. Thoria
Cisco Systems
Jorge. Rabadan
Nokia
John. Drake
Juniper Networks
March 3, 2018

Per multicast flow Designated Forwarder Election for EVPN
draft-sajassi-bess-evpn-per-mcast-flow-df-election-00

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 of applications. [I-D.ietf-bess-evpn-ac-df] and [I-D.ietf-bess-evpn-df-election] improves base line DF election. This document is an extension to HRW base drafts ([I-D.ietf-bess-evpn-ac-df] and [I-D.ietf-bess-evpn-df-election]) and further enhances HRW algorithm 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 network. [RFC7432] defines 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 an All-Active multi-homing.

This DF election mechanism allows selecting a DF at the granularity of (ES, VLAN) or (ES, VLAN bundle) for Broadcast, Unknown Unicast, or Multicast (BUM) traffic. Though [I-D.ietf-bess-evpn-ac-df] and [I-D.ietf-bess-evpn-df-election] improves the default DF election procedure , still it does not fit well for some of service provider

residential application, where whole multicast traffic is delivered on 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 residential deployment scenario, where multiple receivers are behind all active multihoming segment. 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.

- Forcing sole data plane forwarding responsibility on the PE-2 proves a limitation in the current DF election mechanism. In 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-ac-df] and [I-D.ietf-bess-evpn-df-election]).

- In the above deployment we have to consider one more factor, Network bandwidth is shared between multicast and unicast flow. At any given point of time if AC1 already has unicast traffic flow which is taking good amount of network bandwidth. we would have very limited bandwidth available for multicast flows. Even though PE-3 to CE2 (AC2) has not been used much, still we would end up having limitation about how much multicast can flow though AC1.

In this document, we propose an extension to HRW base drafts to allow DF election at the granularity of (ESI, VLAN, Mcast flow) which would allow multicast flows to be distributed among redundancy group PE's 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-ac-df] and [I-D.ietf-bess-evpn-df-election] defines extended community, which would be used for PE's in redundancy group to come to an agreement about which DF election procedures is supported. A PE can notify other participating PE's 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). current proposal extends the existing extended community defined in [I-D.ietf-bess-evpn-ac-df] and [I-D.ietf-bess-evpn-df-election]. 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-ac-df] and [I-D.ietf-bess-evpn-df-election]

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

  - Type 5 - 254: Unassigned

  - Type 255: Reserved for Experimental Use.

- The [I-D.ietf-bess-evpn-ac-df] and [I-D.ietf-bess-evpn-df-election] 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, then these capabilities need to be interpreted in context of this new DF type-4. For example consider a scenario where all PEs in the same redundancy group (same ES) can support both AC-DF and DF type-4 and thus they 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 PE's for the ES, it check if all of other PE's have advertised their capability about Per multicast flow DF election procedure. If all of them have advertised capability, 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 indicates its capability to perform Per multicast flow DF election. OR

  - There is at least one PE signals 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-ac-df] and [I-D.ietf-bess-evpn-df-election], so this draft does not repeat description of HRW algorithm itself.

EVPN PE does the discovery of redundancy group based on [RFC7432]. If redundancy group consists of N EVPN PE nodes. Then after the discovery all PEs build an unordered list of IP address of all the nodes in redundancy group. Procedure defined in this draft does not require PE's to be ordered list.Address [i] denotes the IP address of 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 affinity for (S, G, V, ESI) where

- S - Multicast Source
- G - Multicast Group
- V - Vlan ID for Ethernet Tag V.
- ESI - Ethernet Segment Identifier

In case of tie choose the PE whose IP address is numerically least.

The affinity of PE(i) to (S,G,VLAN ID, ESI) is calculated by function, affinity (S,G,V, ESI, Address(i)), where (0 < i <= N), PE(i) is the PE at ordinal i, address(i) is the IP address of PE at ordinal i

- affinity (S,G,V, ESI, Address(i)) = (1103515245. ((1103515245.Address(i) + 12345) XOR D(S,G,V,ESI))+12345) (mod 2^31)
- D(S,G,V, ESI) = CRC_32(S,G,V, ESI).

Here D(S,G,V,ESI) is the 32-bit digest (CRC_32) of the Source IP, Group IP, Vlan ID for Ethernet Tag V. Source and Group IP address length does not matter as only the lower order 31 bits are modulo significant.

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

In case of IGMP membership request where source is not known. The DF is the PE which has maximum affinity for (G,V, ESI) where

- G - Multicast Group

- V - Vlan ID for Ethernet Tag V.

- ESI - Ethernet Segment Identifier

In case of tie choose the PE whose IP address is numerically least.

The affinity of PE(i) to (G,V, ESI) is calculated by function, affinity (G,V, ESI, Address(i)), where (0 < i <= N), PE(i) is the PE at ordinal i, address(i) is the IP address of PE at ordinal i

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

- D(G,V, ESI) = CRC_32(G,V, ESI).

Here D(G,V,ESI) is the 32-bit digest (CRC_32) of the Group IP, Vlan ID for Ethernet Tag V. Source and Group IP address length does not matter as only the lower order 31 bits are modulo significant.

4.3. Default DF election procedure

Even if all of the PE's indicate their availability to participate in per multicast flow DF election procedure, there is need to have default DF election algorithm. Since Per multicast flow DF election is applicable for only those multicast flows for which PE has received membership request. For other BUM traffic, forwarding plane need default DF election procedure. And we use HRW based DF election procedure as default one in these cases which is defined in [I-D.ietf-bess-evpn-ac-df] and [I-D.ietf-bess-evpn-df-election].

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. PE's 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 procedure defined in [RFC7432].

2. Each of the PE's 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 need to have default DF election among PE's in redundancy group for BUM traffic. In initial phase we use Section 4.3 DF election procedure.

3. When receiver starts sending membership request 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 PE's in redundancy group. Lets consider it is hashed to PE-2. [I-D.ietf-bess-evpn-igmp-mld-proxy] defines procedure to sync IGMP join state among redundancy group of PE's. 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 PE's 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.

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

2. Detection of new PE through ES route.

3. AC going up / down

This document does not provide any new mechanism to handle DF re-election procedure. it does uses existing mechanism defined in [RFC7432]. When ever either of trigger occur, DF re-election would be done. and all of the flows would be redistributed among existing PE's in redundancy group for ES.

7. Protocol Considerations

More details to be added in next version.

8. Security Considerations

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

9. IANA Considerations

There are no new IANA considerations in this document.

10. Acknowledgement

11. Normative References

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

[I-D.ietf-bess-evpn-ac-df]
Rabadan, J., Nagaraj, K., Sathappan, S., Prabhu, V., Liu, A., and W. Lin, "AC-Influenced Designated Forwarder Election for EVPN", draft-ietf-bess-evpn-ac-df-03 (work in progress), January 2018.

[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.

[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