BESS Working Group
Internet Draft
Intended status: Standards Track
Expires: August 21, 2021

Yisong Liu
China Mobile
M. McBride
Futurewei
Z. Zhang
ZTE
J. Xie
Huawei

Feb 21, 2021

Multicast DF Election for EVPN based on bandwidth or quantity
draft-liu-bess-evpn-mcast-bw-quantity-df-election-03

Status of this Memo

Copyright Notice

Abstract

Ethernet Virtual Private Network (EVPN, RFC7432) is becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. When multi-homing from a CE to multiple PEs, including links in an EVPN instance on a given Ethernet Segment, in an all-active redundancy mode, [RFC7432] describes a basic mechanism to elect a Designated Forwarder (DF), and [RFC8584] improves basic DF election by a HRW algorithm. [I-D.ietf-bess-evpn-per-mcast-flow-df-election] enhances the HRW algorithm for the multicast flows to perform DF election at the granularity of (ESI, VLAN, Mcast flow). This document specifies a new algorithm, based on multicast bandwidth utilization and multicast state quantity, in order for the multicast flows to elect a DF.

Table of Contents

1. Introduction

Ethernet Virtual Private Network (EVPN [RFC7432]) solutions are becoming prevalent in Data Centers, Data Center Interconnect (DCI) and Service Provider VPN applications. When multi-homing from a CE to multiple PEs, with links in an EVPN instance on a given Ethernet Segment (ES), in an all-active redundancy mode, [RFC7432] defines the role of Designated Forwarder (DF) as the node that is responsible to forward multicast flows.

Per [RFC7432], the basic method of DF election is specified. The same ES is sorted in ascending order according to the IP address of the EVPN peer. The PE set is generated, and then the number of PEs is modulo according to the VLAN. The modulo value is equal to the position of the PE in the PE set. The election is the primary DF of the corresponding VLAN, and the other PEs are elected as standby.

[RFC8584] defines extended community attributes for DF elections, which can be extended to use different DF election algorithms and would be used for PEs in a redundancy group to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs in a redundancy group about its DF election algorithm by signaling a DF election extended community along with the ES route. The document also improves the basic DF election by a HRW algorithm.

[I-D.ietf-bess-evpn-per-mcast-flow-df-election] proposes a method for DF election by enhancing the HRW algorithm, adding the source and group address of the multicast flow as hash factors, and extending the types 4 and 5 of the extended community of the DF election for (S, G) and (*, G) types for different multicast flows. The source and group address is introduced as new elements to HRW algorithm, and the PE with the largest weight is selected as the DF of the multicast flow.

However, the relationship between the bandwidth of the multicast flows and the link capacity of different PEs, to the same CE device, is not considered in any of the current DF election algorithms. This may result in severe bandwidth utilization of different links due to different bandwidth usage of multicast flows. This document specifies a new algorithm for multicast flow DF election based on multicast bandwidth or multicast state quantity and extends the existing extended community defined in [I-D.ietf-bess-evpn-df-election-framework].

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.2. Terminology

CE: Customer Edge equipment

PE: Provider Edge device

EVPN: Ethernet Virtual Private Network

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

IGMP: Internet Group Management Protocol

MLD: Multicast Listener Discovery

PIM: Protocol Independent Multicast

2. Solution

In the DF election calculation, the bandwidth weight of each multi-homed link of the PE is added, and the bandwidth occupation of the multicast flows is calculated and divided into two scenarios:

* The specific bandwidth value of the multicast flow exists, and the ratio of the current multicast flow bandwidth value to the link bandwidth weight is calculated according to the bandwidth weight of each multi-homed link, and the link with the smallest ratio is elected as the new multicast flow DF.

* The specific bandwidth value of the multicast flow does not exist, and the ratio of the current multicast flow state quantity to the link bandwidth weight is calculated according to the bandwidth weight of each multi-homed link, and the link with the smallest ratio is elected as the new multicast flow DF.

In particular, if there are multiple PEs with the same calculated ratio, the DF is elected according to the method of maximum bandwidth weight of the link or maximum IP address of the EVPN peer.

Since [I-D.ietf-idr-link-bandwidth] defines the link bandwidth extended community, it can be reused to transfer the link bandwidth value of the local ES to other multi-homed PEs, so that each PE can calculate the bandwidth weight ratio of each link of the ES in advance.

2.1. DF Election Based on Bandwidth

Each PE obtains the link bandwidth values of the other multi-homed PEs in the same EVPN instance on a given ES according to the extended community of the Link bandwidth, and calculates the link bandwidth weight ratio, for example W1:W2:...:Wn for N multi-homed PEs.

When the CE sends an IGMP or PIM join to one of the PEs, like PE1, PE1 advertises the PE2, PE3, ... and PEn by the EVPN IGMP/PIM Join Synch route defined in [I-D.ietf-bess-evpn-igmp-mld-proxy] and [I-D.skr-bess-evpn-pim-proxy]. If PE2, PE3, ... or PEn receives an IGMP or PIM join, the procedure will be the same.

Each PE calculates the ratio of the current multicast flows bandwidth to the link bandwidth weight. The one PE in PE1, PE2, ... and PEn, which has the smallest ratio, is elected as the DF of the new multicast flow. When the smallest ratios of more than one PE are the same, the PE with the maximum bandwidth weight of the link or the maximum EVPN peer IP address is elected as the DF.

2.2. DF Election Based on State Qunatity

The procedure is almost the same as described in section 2.1. The only difference is that each PE calculates the ratio of the current number of multicast states instead of the bandwidth to the link bandwidth weight because of lacking specific bandwidth value of the multicast flows.

2.3. Inconsistent Timing between Multi-homed PEs

As a result of the same multicast join, only one of the multi-homed PEs can receive the multicast join message and advertise the EVPN Join Synch route (Type 7). The other PEs need to install the new multicast join state according to the received Synch route.

The inconsistent processing timing of the same multicast group joining process between PEs may cause electing different DFs. For example:

* Multicast group G1, G2, and G3 join packets are sent from the CE to PE1, PE2 and PE3.

* PE1 calculates the DF of G1, while PE2 calculates the DF of G2, and PE calculates the DF of G3, and at this moment each PE has not received the EVPN Join Synch route.

* PE1, PE2 and PE3 select the link on the same ES to the CE using the algorithm as described in section 2.1 or 2.2, and the same DF may be elected for G1, G2, and G3.

* After receiving the EVPN Join Synch route sent by PE2, PE1 may calculate the DF of G2 as PE3, which is inconsistent with the calculation result of PE2.

The DF calculation results of the PEs are inconsistent, which may result in multiple flows or traffic interruptions of the same multicast flow state. Therefore, EVPN Join Synch routes need to carry elected DF information in the route advertisement as the extended community called Multicast DF Extended Community, which can make the DF information for a given multicast flow state between PEs consistent. The actual effect is that the PE that receives the multicast join packet completes the calculation of the DF election and notifies other PEs on the same ES.

2.4. Increase or Decrease of Multi-homed PEs

2.4.1. Decrease of Multi-homed PEs

When one of the multi-homed PEs on the same ES fails or is shut down for maintenance reasons, because the other PEs have received the synch routes of all the multicast flows, the multicast flows destined to the failed PE need to be in a specific order (for example, the group and source address ascending order) to reassign the DF. The DF election calculation based on the multicast flows bandwidth, or the number of multicast states, is completed by one of the specified multi-homing PEs, and the specified calculated PE can be selected according to the link bandwidth weight value or the IP address of the EVPN peer. The specified PE needs to advertise each DF election result of the multicast flow that belongs to the original faulty PE to the other multi-homed PEs that belong to the same ES by the EVPN Join Synch route carrying the Multicast DF Extended Community.

If a new multicast join is received in the above calculation process, the DF election calculation of the new multicast flow is still completed by the PE receiving the multicast join packet. Similarly, the PE needs to advertise the DF information to other multi-homed PEs belonging to the same ES by the EVPN Join Synch route carrying the Multicast DF Extended Community.

2.4.2. Increase of Multi-homed PEs

One multi-homing PE of the same ES is added, and no active adjustment can be performed. The DF of the subsequent new multicast flow is elected according to the algorithm of this document. The new multicast flow must be preferentially assigned to the new PE, and finally the multicast flows on the PEs of the same ES are approximately equalized.

If active adjustment is required, consider calculating the ratio using the algorithm as described in section 2.1 and 2.2. Each time the multicast entries in the PE, whose ratio of the existing multi-homed PE is the largest, are migrated to the new PE. The multicast entries are migrated in descending order of multicast flow bandwidth or in ascending order of the group and source address until the ratio of the new PE is greater than the existing smallest ratio of other multi-homed PEs.

The calculation of the active adjustment is still performed by one specific PE among the multi-homed PEs. The specified calculated PE can be selected according to the link bandwidth weight value or the IP address of the EVPN peer.

After the new PE is started, in the synchronization process of all the multicast entries of other multi-homed PEs, the existing multicast join packet may be received on the new PE. To avoid having the existing multicast join appear as a new multicast join, and recalculating the DF and notifying the other PEs belonging to the same ES, it is necessary to start a timer to suppress the synchronization process from the new PE to other existing PE's. The timer range should also be configured.

3. BGP Encoding

3.1. DF Election Extended Community

[RFC8584] defines an extended community, which would be used for multi-homed PEs to reach a consensus as to which DF election procedure is desired. A PE can notify other participating PEs its DF election capability by signaling a DF election extended community along with Ethernet-Segment Route (Type-4). The current document extends the existing extended community defined in [RFC8584]. This document defines a new DF type.

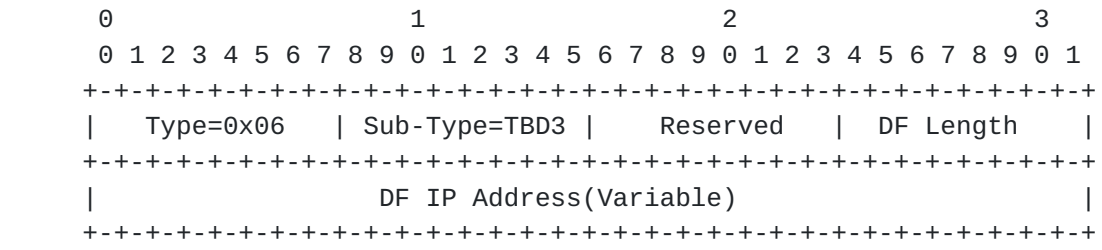
o DF type (1 octet) - Encodes the DF Election algorithm values (between 0 and 255) that the advertising PE desires to use for the ES.

* Type TBD1: Based on bandwidth of multicast flow DF election(detailed in this document)

* Type TBD2: Based on quantity of multicast flow state DF election(detailed in this document)

3.2. Multicast DF Extended Community

This document defines a new extended community in EVPN Type 7 route to notify other multi-homed PEs the elected DF of a given multicast flow. The new extended community is called Multicast DF Extended Community and it belongs to the transitive extended community. The type is to be assigned. It is used to carry DF information of a given (S,G) or (*,G) multicast flow selection. The role of this extended community has been described in sections 2.3 and 2.4.



o Type is 0x06 as registered with IANA for EVPN Extended Communities

o Sub-Type: TBD3

o DF Length: the length of the DF IP Address field, 4 octets for IPv4 address, 16 octets for IPv6 address

o DF IP Address: the elected DF IP address of the given (S,G) or (*,G) route in the EVPN type 7 route

4. Security Considerations

For general EVPN Security Considerations, see [RFC7432].

TBD

5. IANA Considerations

TBD

6. References

6.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC7432] A. Sajassi, Ed., R. Aggarwal, N. Bitar, A. Isaac, J. Uttaro, J. Drake, and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017

[RFC8584] J. Rabadan Ed., S. Mohanty, Ed., A. Sajassi, J. Drake, K. Nagaraj and S. Sathappan, " Framework for Ethernet VPN Designated Forwarder Election Extensibility ", RFC8584, April 2019.

[I-D.ietf-bess-evpn-per-mcast-flow-df-election] Ali Sajassi, Mankamana Mishra, Samir Thoria, Jorge Rabadan and John Drake, " Per multicast flow Designated Forwarder Election for EVPN ", August 2020, work-in-progress, draft-ietf-bess-evpn-per-mcast-flow-df-election-04.

[I-D.ietf-idr-link-bandwidth] P. Mohapatra and R. Fernando, " BGP Link Bandwidth Extended Community ", March 2018, expired, draft-ietf-idr-link-bandwidth-07.

[I-D.ietf-bess-evpn-igmp-mld-proxy] Ali Sajassi, Samir Thoria, Keyur Patel, John Drake and Wen Lin, "IGMP and MLD Proxy for EVPN", January 2021, work-in-progress, draft-ietf-bess-evpn-igmp-mld-proxy-06.

[I-D.skr-bess-evpn-pim-proxy] J. Rabadan, Ed., J. Kotalwar, S. Sathappan, Z. Zhang and A. Sajassi, "PIM Proxy in EVPN Networks", October 2017, expired, draft-skr-evpn-bess-pim-proxy-01.

6.2. Informative References

TBD

7. Acknowledgments

The authors would like to thank the following for their valuable contributions of this document:

TBD

Authors' Addresses

Yisong Liu
China Mobile

Email: liuyisong@chinamobile.com

Mike McBride
Futurewei Inc.

Email: michael.mcbride@futurewei.com

Zheng(Sandy) Zhang
ZTE Corporation

Email: zhang.zheng@zte.com.cn

Jingrong Xie
Huawei Technologies

Email: xiejingrong@huawei.com