BESS Working Group
INTERNET-DRAFT

W. Hao
L. Wang
Y. Li
S. Zhuang
Huawei

Intended Status: Standard Track
Expires: August 15, 2016

February 15, 2016

Centralized EVPN DF Election
draft-hao-bess-evpn-centralized-df-00.txt

Abstract

This document proposes a centralized DF Designated Forwarder election mechanism to be used between the SDN(Software defined network) controller and each PE(Provider Edge) in EVPN network. Such a mechanism overcomes the issues of current standalone DF election defined in RFC7432. A new BGP capability and an additional DF Election Result Route Type are proposed to support the centralized DF mechanism.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Table of Contents

1. Introduction

RFC7432 defines the Designated Forwarder (DF) election mechanism in EVPN networks to appoint one PE as DF from a candidate list of PEs connecting to a multi-homed CE device or access network. The DF PE is responsible for sending broadcast, multicast and unknown unicast traffic (BUM) to the multi-homed CE device or network and non-DF PEs should drop such traffic. DF based mechanism is used to prevent the duplicated packet injected into the multi-homed access network via multiple PEs.

DF is selected according to the VLAN modulus 'service-carving' algorithm in order to perform load balancing for multi-destination traffic destined to a given segment [RFC7432]. The algorithm can ensure each participating PE independently and unambiguously selects one of the participating PEs as the DF; while it has some drawbacks as follows [EVPN-HRW-DF].

1. Low loadbalancing uniformity in some VLAN configuration case when the ethernet tag follows a non-uniform distribution, for instance when the Ethernet tags are all even or all odd.

2. Unnecessary service disruption upon member PE joining and leaving in a redundancy group. In Figure 1, say v1, v2 and v3 are vlans configured on ES2 with associated ethernet tags of value 3, 4 and 5 respectively. So PE1, PE2 and PE3 are also the DFs for v1, v2 and v3 respectively. Now when PE3 goes down, PE2 will become the DF for v1 and v3 while PE1 will become the DF for v2, needless churn of v1 and v2 occurs and it will cause unnecessary service disruption in v1 and v2.

3. Non-deterministic DF election result which lacks user control.In some cases, the user may want to flexibly control the loadbalancing based on VLAN number, bandwidth consumption, and other factors. The user should be allowed to use some specific DF re-election algorithm to avoid service disruption. The user also should be allowed to specify revertive and non-revertive mode for on-demand DF switchover in order to carry out some maintenance tasks.

This document proposes a centralized DF election method to overcome the issues aforementioned. A physically distributed but logically centralized controller is suggested to be deployed to perform the DF election calculation for all multi-homed PEs. Each individual multi-homed PE should turn off its own DF election process and listen to the DF election result from the SDN controller only. RFC7432 DF election procedures defined in RFC7432 are extended for the interaction between SDN Controller and each PE.

2. Conventions used in this document

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 RFC 2119 [RFC2119].

The terms and acronyms in [RFC6325] are used with the following additions:

CE: Customer Edge device, e.g., a host, router, or switch.

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

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'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

3. Solution overview

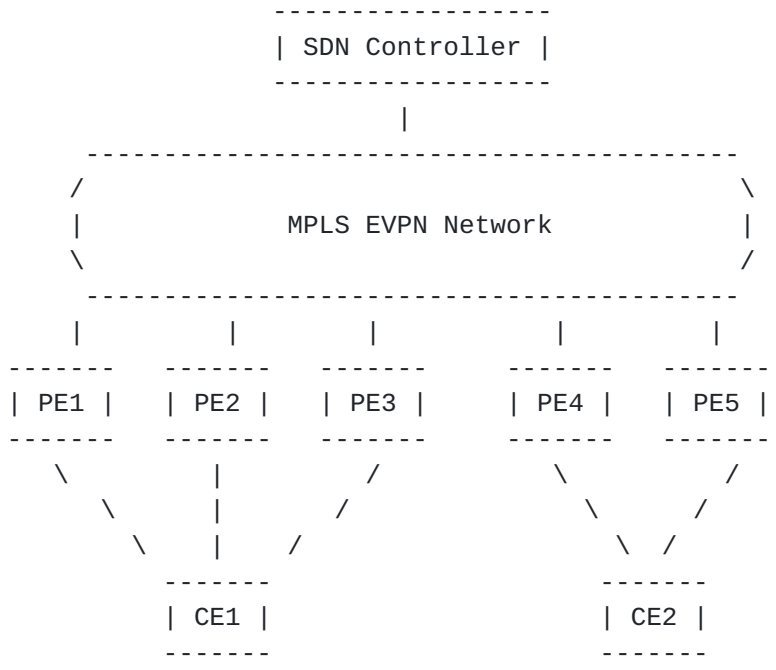


Figure 1: Centralized DF Election scenario

In figure 1, CE1 is multi-homed to PE1,PE2 and PE3, the ESI is 1. CE2 is multi-homed to PE4 and PE5, the ESI is 2. The controller will be pre-provisioned entire network's ESI related configuration, which includes EVI, the Ethernet Tags on each ESI, redundancy mode of active-active or active-standby for each ESI, <ESI, Ethernet Tag> and EVI correspondence.

Before each PE and SDN controller exchange BGP route information for DF election, it's required for the SDN controller and each PE to negotiate a new BGP centralized DF election capability and role firstly when OPEN messages are exchanged, each multi-homed PE is the client for DF election while the SDN controller is the server. For the DF election Client, regular DF election process as per RFC7432 will be turned off, and it will only listen to the DF/Non-DF result from the SDN controller at the granularity of <ES,VLAN> or <ES, VLAN bundle>. For the DF election server, after it receives Ethernet Segment route from each PE, it will perform DF election calculation based on local algorithm and will notify each EVPN PE the election

result through new EVPN route type.

3.1. Centralized DF Election Capability

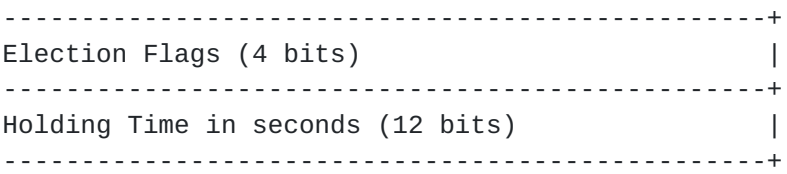
The centralized DF election capability is a new BGP capability [BGP-CAP] that can be used by a BGP speaker to indicate its ability to support for the new DF election process.

This capability is defined as follows:

Capability code: TBD

Capability length: variable

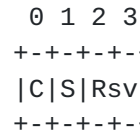
Capability value: Consists of the "Election Flags" field, "Waiting Time" field as follows:



The use and meaning of the fields are as follows:

Election Flags:

This field contains bit flags related to restart.



The most significant bit is defined as the election client(C) bit to indicate the BGP speaker is the Client which will wait the DF election result from Controller(Server). When set (value 1), this bit indicates that the BGP speaker is the server(Controller) which has the DF election calculation capability for all multi-homed PEs in entire EVPN network.

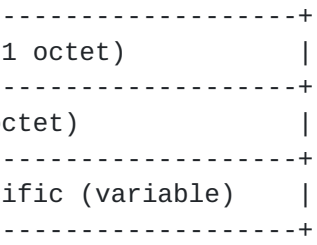
Holding Time:

This is the estimated time (in seconds) it will take for the client to get DF election result from the controller after BGP session establishes. When no result about DF election is received after the holding time goes to zero,), PEs will revert to traditional EVPN DF election process as per RFC7432.

The remaining bits are reserved and MUST be set to zero by the sender and ignored by the receiver.

4. DF Election Result Route Type

The current BGP EVPN NLRI as defined in [RFC7432] is shown below:



This document defines an additional Route Type used for the server(Controller) to notify DF election result to each client(PE). The Route Type is ''DF Election Result Route Type''.

The detailed encoding of this route and associated procedures are described in the following sections.

4.1. DF Election Result Route encoding

A DF Election Result Route NLRI consists of the following fields:

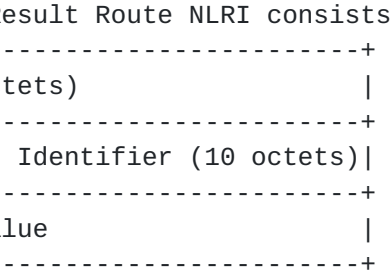
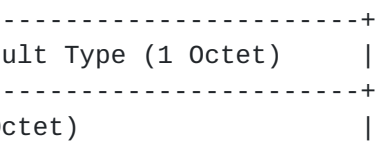


Figure 2: DF Election Result Router NLRI

o RD: The Route Distinguisher (RD) MUST be a Type 1 RD [RFC4364]. The value field comprises an IP address of the Controller (typically, the loopback address) followed by a number unique to the Controller.

o Ethernet Segment Identifier: Is a non-zero 10-byte identifier for an Ethernet Segment.

o Value (variable): Information in the Value field is encoded in Type/Length/Value triplets. Multiple TLVs can be comprised.



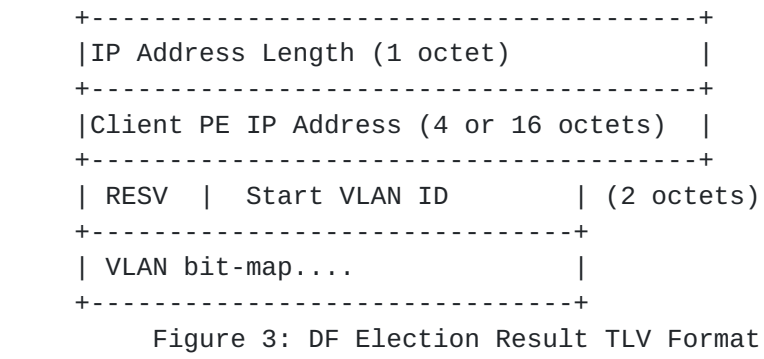


Figure 3: DF Election Result TLV Format

o DF Election Result Type (1 octets): identifies the type of DF Election. This document defines the following one type:

- VLAN Bitmap: DF Election Type = 1

Unknown types are to be ignored and skipped upon receipt.

o Length (2 octets): the total number of octets of the value field.

o TLV Value (variable): encodings of the value field depend on the sub-TLV type as enumerated above. The following sub-sections define the encoding in detail.

o The IP Prefix Length can be set to a value between 0 and 32 (bits) for ipv4 and between 0 and 128 for ipv6.

o The Client PE IP Address will be a 32 or 128-bit field (ipv4 or ipv6) as PE's identification.

o Start VLAN ID: The 12-bit VLAN ID that is represented by the high order bit of the first byte of the VLAN bit-map.

o VLAN bit-map: The highest order bit indicates the VLAN equal to the start VLAN ID, the next highest bit indicates the VLAN equal to start VLAN ID + 1, continuing to the end of the VLAN bit-map field. Bit 1 indicates DF, Bit 0 indicates non-DF.

4.2. Centralized DF Election procedures

The controller has all ES's related configuration in entire EVPN network. After the controller boots up, the controller can start a boot-timer to allow the establishment of BGP EVPN Sessions with all multi-homed EVPN PEs. The controller also needs to receive all ES

routes from those PEs before the boot-timer timeout. The controller will preserve all EVPN PE's ES routes.

Based on local algorithm for each ES, it can start to perform DF election calculation. The default algorithm is the VLAN modulus method defined in section 8.5 [RFC7432] relying on local VLAN configuration on each ES. The algorithm should allow user defined.

After DF election calculation is finished on the controller, it will notify each multi-homed PE using the new defined DF Election Result Route. The DF Election Result Route is per ES, i.e., the DF election results of all PEs connecting to same ES are carried in one route. The controller that advertises the Ethernet Segment route must carry an ES-Import Route Target, the DF Election Result filtering procedure is same as Ethernet Segment route filtering defined in [RFC7432], i.e., the DF Election Result Route filtering MUST be imported only by the PEs that are Multi-homed to the same Ethernet segment. Each Multi-homed PE compares Client PE IP Address with its local IP Address, if the two IP addresses are same, then it gets corresponding start VLAN and VLAN Bitmap as the DF election result.

When a multi-homed PE failure occurs and is detected by the controller, the controller will initiate DF re-election process. Because it's the controller making decision which PE is DF or non-DF, the controller should ensure that the DF re-election won't cause unnecessary service disruption. The controller should only redistribute the DF VLAN on PE3 to PE1 and PE2, the existing DF VLAN on PE1 and PE2 should remain unchanged to avoid service disruption.

When the access link failure on one multi-homed PE occurs, the PE will advertise Ethernet Segment Withdraw message to the controller, then it will trigger the DF re-election on the controller, the re-election principle is same as node failure to avoid service disruption as little as possible.

5. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. The communications between SDN Controller and EVPN PE should be protected to ensure security. BGP peerings are not automatic and require configuration, thus it is the responsibility of the network operator to ensure that they are trusted entities.

6. IANA Considerations

This document requests a new BGP capability code - Centralized DF Election Capability.

This document requests the allocation of value TBD in the "EVPN Route Types" registry defined by [RFC7432] and modification of the registry as follows:

Value	Description	Reference
TBD	DF Election Result Route	[this document]
TBD-255	Unassigned	

IANA set up a registry for "DF Election Result Type". This is a registry of two-octet values (0-65535), to be assigned on a first-come, first-served basis. The initial assignments are as follows:

Tunnel Name	Type
VLAN Bitmap	1

7. Normative References

[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, <http://www.rfc-editor.org/info/rfc7432>.

8. Informative References

[EVPN-HRW-DF] Mohanty S. et al. "A new Designated Forwarder Election for the EVPN", draft-mohanty-bess-evpn-df-election-02, work-in-progress, October 19, 2015.

Acknowledgments

The authors wish to acknowledge the important contributions of Qiandeng Liang.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012, China
Email: haoweiguo@huawei.com

Lili Wang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lily.wong@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: liyizhou@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com