**Centralized EVPN DF Election**
**draft-hao-bess-evpn-centralized-df-01.txt**

Abstract

   This document proposes a centralized DF Designated Forwarder election
   mechanism to be used between an SDN (Software Defined Network)
   controller and each PE (Provider Edge) device in an EVPN network.
   Such a mechanism overcomes the issues of current standalone DF
   election defined in RFC 7432. A new BGP capability and an additional
   DF Election Result Route Type are specified to support this
   centralized DF mechanism.

Status of this Memo

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
   and VLANs (or VLAN bundles) 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 must drop such traffic. This DF
   based mechanism is used to prevent duplicated packet injection into
   the multi-homed access network via multiple PEs.

   In [RFC7432] the 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]. The algorithm
   can ensure each participating PE independently and unambiguously
   determines which one of the participating PEs is the DF; however, use
   of this algorithm has some drawbacks as follows [EVPN-HRW-DF].

   1. Uneven load balancing in some VLAN configuration cases 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 when PEs join or leave a redundancy
      group. In Figure 1 below, 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. Lack of user control over DF election. In some cases, the user may
      want to flexibly control the load balancing 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 specifies a centralized DF election method to overcome
   the issues aforementioned. A physically distributed but logically
   centralized controller is deployed to perform the DF election
   calculation for all multi-homed PEs. Each individual multi-homed PE
   should disable its own DF election process and listen to the DF
   election result from the SDN controller. [RFC7432] DF election
   procedures are extended for the interaction between the 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", "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.

The following terms and acronyms are used:

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

   DF: Designated Forwarder.

   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".

   ESI: Ethernet Segment Identifier: A unique non-zero identifier
        that identifies an Ethernet segment.

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

   EVPN: Ethernet Virtual Private Network.

   PE: Provider Edge device.

   NLRI: Network Layer Reachability Information.

   SDN: Software Defined Networking.

   VLAN: Virtual Local Area Network.

## 3. Solution Overview

```
                     ------------------
                     | SDN Controller |
                     ------------------
                             |
            ---------------------------------------------
          /                                               \
          |              MPLS EVPN Network                |
          \                                               /
            ---------------------------------------------
            |         |         |           |         |
         -------   -------   -------     -------   -------
         | PE1 |   | PE2 |   | PE3 |     | PE4 |   | PE5 |
         -------   -------   -------     -------   -------
            \         |         /           \         /
             \        |        /             \       /
              \       |       /               \     /
               -------               -------
               | CE1 |               | CE2 |
               -------               -------
```
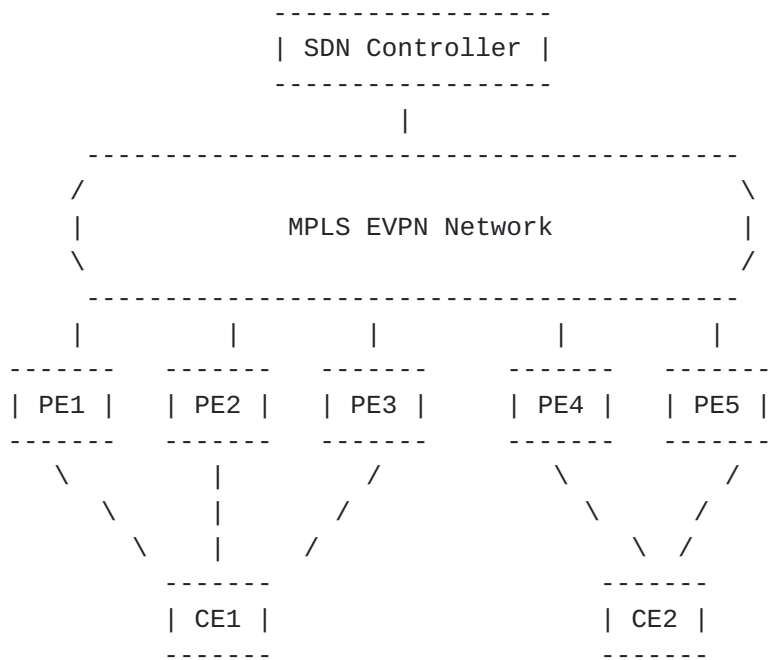
Figure 1. Centralized DF Election Scenario

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 SDN controller
will be pre-provisioned with the entire network's ESI related
configuration.  This 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 the SDN controller exchange BGP route information
for DF election, the SDN controller and each PE MUST negotiate a new
BGP centralized DF election capability and role when OPEN messages
are first exchanged; each multi-homed PE is the client for DF
election while the SDN controller is the server. For the DF election
Client, the regular DF election process as per [RFC7432] will be
disabled, and each PE listens 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 a local
algorithm and will notify each EVPN PE of the election result through
a new EVPN route type.

### 3.1 Centralized DF Election Capability

The centralized DF election capability is a new BGP capability

[RFC5492] 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: TBD1

Capability length: 2 octets

Capability value: Consists of the "Election Flags" field and
"Holding Time" field as follows:

```
| 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15|
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
| Election  | Holding Time in seconds           |
|  Flags    |                                   |
| (4 bits)  |  (12 bits)                        |
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
```

The use and meaning of these fields are as follows:

Election Flags: This field contains bit flags related to
restart as follows:

```
| 0   1   2   3|
+---+---+---+---+
| C | S | Rsv   |
+---+---+---+---+
```

C: The most significant bit is the election Client bit. When
set to 1 it indicates the BGP speaker is a Client which
will await the DF election result from the Controller
(Server).

S: When set to 1, this bit indicates that the BGP speaker is
the Server (Controller) that has the DF election
calculation capability for all multi-homed PEs in the
entire EVPN network.

Rsv: Reserved bits that MUST be sent as zero and ignored on
receipt.

Holding Time: This is the estimated time in seconds it will
take for the client to get DF election result from the
controller after the BGP session is established. When no
result for the DF election is received after the holding
time, PEs will revert to the traditional EVPN DF election
process as per [RFC7432].

## 4. DF Election Result Route Type

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

```
+----------------------------------+
|      Route Type (1 octet)        |
+----------------------------------+
|        Length (1 octet)          |
+----------------------------------+
|  Route Type specific (variable)  |
+----------------------------------+
```

This document defines an additional Route Type used for the server
(SDN Controller) to send DF election results 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

The route type specific information for a DF Election Result Route
NLRI consists of the following fields:

```
+--------------------------------------+
|      RD    (8 octets)                |
+--------------------------------------+-------+
|  Ethernet Segment Identifier (10 octets)     |
+----------------------------------------------+
|               TLVs               ...
+-----------------------------------
```

Figure 2: DF Election Result Router NLRI

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.

ESI: Ethernet Segment Identifier: Is a non-zero 10-octet
   identifier for an Ethernet Segment.

TLVs: Information in the TLVs field is encoded in
   Type/Length/Value triplets. Multiple TLVs can be included. This
   document specifies type 1, the VLAN Bitmap type, whose
   structure is as follows:

```
+-----------------------------------------+
| DF Election Result Type = 1             | (2 octets)
+-----------------------------------------+
| Length                                  | (2 octets)
+-------------------------+---------------+
|IP Address Prefix Length|     (1 octet)
+------------------------..--------------------+
| Client PE IP Address         (4 or 16 octets)  |
+------------------------..--------------------+
| RESV  |   Start VLAN ID                  | (2 octets)
+-----------------------------------------+
| VLAN bit-map....             ...
+------------------------------
```

              Figure 3. DF Election Result TLV Format

    o  DF Election Result Type (2 octets): Identifies the type
       of DF Election result.  This document defines type 1 as
       the "VLAN Bitmap" Type. TLVs withe unknown types are
       ignored and skipped upon receipt.

    o  Length (2 octets): The total number of octets of the
       value part of the TLV.

    The type and length are followed by the variable length value.
    This value, for the VLAN Bitmap type, consists of the following
    fields:

    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  RESV is a 4-bit reserved field that MUST be sent as zero
       and ignored on receipt.

    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. A bit value of 1
       indicates DF and a bit value of 0 indicates non-DF.

## 4.2 Centralized DF Election procedures

The controller has all ES related configuration information for the entire EVPN network. After the controller boots up, it 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 a local algorithm for each ES, it can start to perform the 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. A user defined algorithm should be allowed.

After the DF election calculation is finished on the controller, it will notify each multi-homed PE using the newly defined DF Election Result Route. The DF Election Result Route is per ES, i.e., the DF election results for all PEs connecting to the 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 the 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 the Client PE IP Address with its local IP Address, if the two IP addresses are same, then it gets the corresponding start VLAN and VLAN Bitmap as the DF election results.

When the failure of a multi-homed PE is detected by the controller, the controller will initiate the DF re-election process.  Because it's the controller making decisions as to which PE is DF or non-DF, the controller should ensure that the DF re-election won't cause unnecessary service disruption. In the example above, 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 fails on one multi-homed PE, the PE will advertise an Ethernet Segment Withdraw message to the controller, which will trigger the DF re-election on the controller, the re-election principle is same as in the node failure case to minimize service disruption.

[5](). **Security Considerations**

   Procedures and protocol extensions defined in this document do not
   affect the BGP security model. The communications between the SDN
   Controller and EVPN PEs 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

Three IANA actions are requested as below.

IANA is requested to assign a new BGP Capability Code in the
Capability Code registry as follows:

```
Value    Description              Reference
------   ----------------------   ---------------
 TBD1    Centralized DF Election  [this document]
```

This document requested the assignment of value TBD2 in the "EVPN
Route Types" registry created by [RFC7432] and modification of the
registry to add the following:

```
Value    Description          Reference
------   ------------------   ---------------
 TBD2    DF Election Result   [this document]
```

IANA is requested to create a registry for "DF Election Result Types"
as follows:

Name: DF Election Result Types
Registration Procedure: First Come First Served
Reference: [this document]

```
  Type       Description      Reference
 --------   -------------    ---------
      0      (Reserved)
      1      VLAN Bitmap      [this document]
 2-65534     unassigned
   65535     (reserved)
```

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, <http://www.rfc-editor.org/info/rfc2119>.

    [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>.

    [RFC5492] - Scudder, J. and R. Chandra, "Capabilities Advertisement
          with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009,
          <https://www.rfc-editor.org/info/rfc5492>.

    [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>.

    [RFC8174] - Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
          2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May
          2017, <https://www.rfc-editor.org/info/rfc8174>.


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.

Authors' Addresses

       Donald Eastlake, 3rd
       Huawei Technologies
       155 Beaver Street
       Milford, MA 01757 USA

       Email: d3e3e3@gmail.com


       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

Copyright, Disclaimer, and Additional IPR Provisions