BESS Working Group A. Sajassi
Internet-Draft G. Badoni
Intended Status: Standards Track D. Rao
P. Brissette
Cisco
J. Drake
Juniper
J. Rabadan
Nokia

Expires: September 19, 2018 March 19, 2018

Fast Recovery for EVPN DF Election
draft-sajassi-bess-evpn-fast-df-recovery-02

Abstract

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] describes DF election procedures for multi-homing Ethernet Segments. These procedures are enhanced further in [DF-FRAMEWORK] by applying Highest Random Weight Algorithm for DF election in order to avoid DF status unnecessarily upon a failure. This draft makes further improvement to DF election procedures in [DF-FRAMEWORK] by providing two options for fast DF election upon recovery of the failed link or node associated with the multi-homing Ethernet Segment. This fast DF election is achieved independent of number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each PE in the multi-homing group.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC 7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

EVPN solution [RFC 7432] describes DF election procedures for multi-homing Ethernet Segments. These procedures are enhanced further in [DF-FRAMEWORK] by applying Highest Random Weight Algorithm for DF election in order to avoid DF status change unnecessarily upon a link or node failure associated with the multi-homing Ethernet Segment. This draft makes further improvement to DF election procedures in [DF-FRAMEWORK] by providing two options for a fast DF election upon recovery of the failed link or node associated with the multi-homing Ethernet Segment. This DF election is achieved independent of number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each PE in the multi-homing group. The draft presents two signaling options. The first option is based on a bidirectional handshake procedure whereas the second option is based on simple one-way signaling mechanism.

1.1 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 RFC 2119 [KEYWORDS].

Provider Edge (PE) : A device that sits in the boundary of Provider and Customer networks and performs encap/decap of data from L2 to L3 and vice-versa.

Designated Forwarder (DF): An PE that is currently forwarding (encapsulating/decapsulating) traffic for a given VLAN in and out of a site.

2 Challenges with Existing Solution

In EVPN technology, multiple PE devices have the ability to encap and decap data belonging to the same VLAN. In certain situations, this may cause L2 duplicates and even loops if there is a momentary overlap of forwarding roles between two or more PE devices, leading to broadcast storms.

EVPN [RFC 7432] currently uses timer based synchronization among PE devices in redundancy group that can result in duplications (and even loops) because of multiple DFs if the timer is too short or blackholing if the timer is too long.

Using site-of-origin Split Horizon filtering can prevent loops (but not duplicates), however if there are overlapping DFs in two different sites at the same time for the same VLAN, the site identifier will be different upon re-entry of the packet and hence the split horizon check will fail, leading to L2 loops.

The current state of art [DF-FRAMEWORK] uses the well known HRW (Highest Random Weight) algorithm to avoid reshuffling of VLANs among PE devices in the redundancy group upon failure/recovery and thus reducing the impact of failure/recovery to VLANs not on the failed/recovered ports. This eliminates loops/duplicates in failure scenarios.

However, upon PE insertion or port bring-up, HRW cannot help as a transfer of DF role need to happen to the newly inserted device/port while the old DF is still active.

```
                                     +---------+
                  +-------------+    |         |
                  |             |    |         |
                / |    PE1      |----|         |   +-------------+
               /  |             |    |  MPLS/  |   |             |---H3
              /   +-------------+    |  VxLAN/ |   |     PE10    |
         CE1 -                       |  Cloud  |   |             |
              \   +-------------+    |         |---|             |
               \  |             |    |         |   +-------------+
                \ |     PE2     |----|         |
                  |             |    |         |
                  +-------------+    |         |
                                     +---------+
```

Figure 1: CE1 multi-homed to PE1 and PE2. Potential for duplicate DF.

In the Figure 1, when PE2 is inserted or booted up, PE1 will transfer DF role of some VLANs to PE2 to achieve load balancing. However, because there is no handshake mechanism between PE1 and PE2, duplication of DF roles for a give VLAN is possible. Duplication of DF roles may eventually lead to L2 loops as well as duplication of traffic.

Current state of EVPN art relies on a blackholing timer for transferring the DF role to the newly inserted device. This can cause the following issues:

* Loops/Duplicates if the timer value is too short

* Prolonged Traffic Blackholing if the timer value is too long

This draft is proposing solutions that deterministically eliminates loops/duplicates and at the same time provides fast convergence upon PE/port insertion.

3 Operation

Here we describe two signaling mechanisms between the newly inserted PE and remaining PEs. The signaling is only possible once the newly inserted PE has reliably discovered the other PEs and vice versa. The first option is referred to as DF Election Handshake solution and is described in section 3.1. The second option is referred to as DF Election Synchronization Solution and is described in section 3.2.

3.1 DF Election Handshake Solution

Due to HRW, the handshake will only be one per PE device and independent of EVI/VNI scale. Therefore, this solution is divided into three steps:

Phase 1: Discovery

Phase 2: DF Candidate Determination; HRW or Preference-based

Phase 3: Handshake

Following is the description each step in detail.

3.1.1 Discovery

Each PE needs to have a consistent view of the network including the newly inserted PE.

Newly inserted device PE will advertise it's Ethernet Segment route and start a flood/wait timer. This timer should be large enough to guarantee the dissemination and receipt of this advertisement by previously inserted PEs.

As the old DF is continuously forwarding traffic while the new PE is running this timer, this timer can be made as long as required without impacting traffic convergence. The timer value can be the BGP session hold time in the worst case to ensure proper discovery.

3.1.2 DF candidates Determination

After the discovery timer has elapsed, each PE would have an imported list of the Ethernet Segment Routes from other PEs. The resultant database will comprise of all the DF candidates on a per ES basis and will be used for DF election. Each PE will independently run the selected DF algorithm - i.e., HRW algorithm (or Preference-based) for all VLANs in a given Ethernet Segment. Since the discovery phase guarantees uniform network view between the participating devices, the VLAN distribution results based on HRW (or Preference-based) will be consistent.

3.1.3 DF Election Handshake

The DF Election handshake will be accomplished in the following steps:

- The newly inserted PE will send the DF Request to previously inserted PEs with a new sequence number.

- The previously inserted PE(s) will receive the DF Request, will validate this request as per own discovery state and HRW (or Preference-based) results.

- The previously inserted PE(s) will program hardware to block the VLANs that must be transferred to the newly inserted PE.

- The previously inserted PE(s) will send DF Response (W/ ACK OR NACK) to the newly inserted PE with the same sequence number that was contained in the DF Request.

- Newly inserted PE will receive DF Response and validate it using the sequence number. It will take action per received DF Response message and will not wait for all previously inserted devices for faster convergence. The received DF Response is interpreted as an indication from the previously inserted PE to give up the DF role on those VLANs for which the newly inserted PE should be DF. In other words, the newly inserted PE will only take over as DF for a given VLAN/ISID if (a) it is the DF Election winner AND (b) it gets the ACK from the previous DF.

- In case of Preference-based DF Election, the above procedure should only be followed if there is at least one previously inserted PE that signals DP=0 in its ES route (there is no need for handshake in case of non-revertive mode).

- In case of a DF Response ACK, newly inserted PE will program its hardware to assume the DF responsibility.

We don't need to have a handshake on a per VLAN/EVI basis but rather per pair of PEs in the redundancy group - i.e., if a new PE is added to an existing redundancy group of 3 PE devices, then we need only to have 3 handshakes. This is because the devices already are in sync about which VLANs to give-up/takeover (HRW).

At the end of these three phases, the VLAN DF role transfer would have happened in a deterministic way while ensuring minimum traffic loss. Device recovery and device insertion scenarios are identical in terms of the handshaking procedure. In next section, we describe the procedure details for device insertion.

3.1.4 Node Insertion

Consider the scenario where PE3 is inserted in the network, while PE1 and PE2 are already in stable state. PE3 will send/receive the following flags along with the EVPN Type 4 route:

- DF Request: Upon completing the DF Election, PE3 will send DF Request with a new sequence number. PE1 and PE2 will receive this message and respond with DF Response ACK or NACK with the same sequence number that was generated by PE3.

- DF Response ACK: When PE3 receives DF Response ACK from PE1 with the same sequence number as DF Request, it will take over the DF role for the appropriate VLANS that are being transferred from PE1. When DF Response ACK from PE2 arrives, the rest of the VLANS to be transferred from PE2 to PE3 are then taken over by PE3.

- DF Response NACK: If PE3 receives DF Response NACK from at least one of PE1 or PE2, it will not take over DF role and will start over.

Consider the scenario where two nodes PE3 and PE4 are being inserted at the same time. Both of them will send a DF Request to PE1 and PE2 at around the same time with possibly the same sequence number. When PE1 and PE2 respond with DF Response ACK, it is important to signify exactly whom the response is meant for as it could be for either requester (PE3 or PE4). To remove any ambiguity and false positives, the IP address of the requester MUST be included in the response message to specify who the response is meant for.

3.1.5 BGP Encoding

The EVPN NLRI comprises of Route Type (1B), Length (1B) and Route Type specific variable encoding. Here we propose the creation of two new EVPN route types:

+ 0x0C - DF Election Handshake Request Route
+ 0x0D - DF Election Handshake Response Route

3.1.5.1 DF Election Handshake Request Route

A DF Election Handshake Request Type NLRI consists of the following:

```
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-++-+-+-+-+-+-++
       | RD (8 octets)                           |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Ethernet Segment Identifier (10 octets) |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | DF-Flags (1 octet)                      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Sequence Number (1 octet)               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Originating Router's IP Address         |
       |          (4 or 16 octets)               |
       +-----------------------------------------+
```

The DF-Flags can have the following values:

```
DF-INIT  : Sent initially upon boot-up; bootstraps the network
DF-REQUEST : Sent to request DF takeover
```

For the purpose of BGP route key processing, the Ethernet Segment Identifier and Originating Router's IP address fields are considered to be part of the prefix in the NLRI. The DF-Flag and Sequence number is to be treated as a route attribute as opposed to being part of the route. This route is sent along with ESI-Import route target.

3.5.1.2 DF Election Handshake Response Route

A DF Election Handshake Response Type NLRI consists of the following:

```
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-++-+-+-+-+-+-++
       | RD (8 octets)                             |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Ethernet Segment Identifier (10 octets) |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | IP-Address Length (1 octet)             |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Destination Router's IP Address         |
       |      (4 or 16 octets)                   |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | DF-Flags (1 octet)                      |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Sequence Number (1 octet)               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | Originating Router's IP Address         |
       |          (4 or 16 octets)               |
       +-----------------------------------------+
```

The DF-Flags can have the following values:

```
   DF-ACK       : Sent to Acknowledge DF-REQUEST
   DF-NACK      : Sent to Reject DF-Request
```

For the purpose of BGP route key processing, the Ethernet Segment Identifier, IP Address Length and Destination Router's IP Address fields, and Originating Router's IP address fields are considered to be part of the prefix in the NLRI. The DF-Flag and Sequence number is to be treated as a route attribute as opposed to being part of the route. This route is sent along with ESI-Import route target.

This document introduces a new flag called "H" (for Handshake) to the bitmap field of the DF Election Extended Community defined in [DF-FRAMWORK].

```
                        1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06     | Sub-Type(0x06)|   DF Type     |P|A|H|T| Bitmap|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          Reserved = 0                         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

H: This flag is located in bit position 26 as shown above. When set to 1, it indicates the desire to use Handshaking capability with the rest of the PEs in the ES. This capability can only be used with a selected number of DF election algorithms such as HRW and Preference-based.

3.1.6 DF Handshake Scenarios

Consider the scenario where PE3 is freshly inserted into the network with PE1 and PE2 in steady state (as shown below). As shown in the sequence diagram below, at time = T0, PE3 will send Type 4 ES route and that will cause PE1 and PE2 to discover PE3.

Post the discovery timer, at time = T1, PE3 will send DF Request containing [ESI, DF-REQ, SEQ1].

PE2 responds via DF Response ACK at time = T2, with the same sequence number SEQ1. [ESI, DF-ACK, PE3, SEQ1]. Note that the sequence number is the same as is contained in the DF Request from PE3. PE3 will receive the DF Response ACK and take over the appropriate VLANs based on HRW only if the sequence number matches.

PE1 responds via DF Response ACK at time = T3, with the same sequence number SEQ1; [ESI, DF-ACK, PE3, SEQ1]. PE3 will receive the DF Response ACK and take over the appropriate VLANs based on HRW only if the sequence number matches.

By the end of the handshake, all appropriate VLANs for the ES are transferred from PE1 and PE2 to PE3 with a single per-ES handshake.

```
      PE1                  PE2                          PE3
       |                   |                           |
       |                   |      Type 4 (Discovery)   |
       |                   |<<-------------------------| T0
       |<<---------------------------------------------|
       |                   |                           |
       |                   |                           |
       |                   |      Type C (DF Request)  | .br
       |<<-----------------|<<-------------------------| T1
       |                   |                           |
       |                   |      Type D (DF Response) | .br
       |                   |------------------------->>| T2
       | Type D(DF Resp)   |                           |
       |--------------------------------------------->>| T3
       |                   |                           |
       |<<###########################################>>|
       |            PE3 freshly inserted               |
       |<<###########################################>>|
       .                   .                           .
```

Consider the scenario where PE2 and PE3 are inserted simultaneously in the network where PE1 is in steady state (as shown below). PE2 and PE3 will send the Type 4 ES routes and start the discovery timer. This will cause PE1, PE2 and PE3 to discover each other.

PE2 and PE3 will then simultaneously and separately send DF Request. PE1 will receive these requests and respond to them.

To avoid any ambiguity, PE1 will explicitly specify in the DF Request route the destination for which the DF-ACK is meant for. That is why the responses from PE1 will contain [ES1, DF-ACK, PE2, SEQ] and [ESI, DF-ACK, PE3, SEQ] to specify that the response is meant for PE2 and PE3 respectively.

Upon receiving the Type-D response message, PE2 and PE3 will take over the respective VLANs.

```
      PE1                 PE2                            PE3
       |                   |                           |
       |                   |      Type 4 (Discovery)   |
       |<<-----------------|                           | T0
       |<<-----------------|<<-------------------------|
       |                   |                           |
       |                   |                           |
       |                   |      Type C (DF Request)  |
       |                   |<<-------------------------| T1
       |                   |                           |
       | Type C(DF Request)|                           |
       |<<-----------------|                           | T2
       |                   |                           |
       |                   |      Type D (DF Response) |
       |--------------------------------------------->>| T3
       |                   |                           |
       |Type D(DF Response)|                           |
       |----------------->>|                           | T4
       |                   |                           |
       |<<###########################################>>|
       |     PE2 and PE3 inserted simultaneously       |
       |<<###########################################>>|
       .                   .                           .
```

When PE3 is booted down or removed from the network, the routes formerly advertised by PE3 will be withdrawn, including the Type 4 route (as shown below). When PE1 and PE2 process the deletion of PE3's Type 4 route, they will clean up any DF handshake state pertaining to PE3. This means that PE1 and PE2 will withdraw the DF Response routes that they had earlier sent with PE3 as the destination.

```
      PE1                  PE2                          PE3
       |                   |                           |
       |                   | Type 4 Route Withdrawal   |
       |                   |<<-------------------------| T0
       |<<---------------------------------------------|
       |                   |                           |
       |       PE2 purges Type D(DF Resp) sent to PE3  | T2
       |                   |                           |
       |       PE1 purges Type D(DF Resp) sent to PE3  | T3
       |                   |                           |
       |<<###########################################>>|
       | PE3 booted down/removed from the network      |
       |<<###########################################>>|
       .                   .                           .
```

3.1.7 Interoperability

Per redundancy group (per ES), for the DF election procedures to be globally convergent and unanimous, it is necessary that all the participating PEs agree on the DF Election algorithm to be used. It is, however, possible that some PEs continue to use the existing modulus based DF election and do not rely on the new handshake/sync procedures. PEs running an old versions of draft/RFC shall simply discard unrecognized new BGP extended communities.

A PE can indicate its willingness to support new Handshake and/or Time Synchronization capabilities by signaling them in the DF Election Extended Community defined in [DF-FRAMEWORK] sent along with the Ethernet-Segment Route (Type-4).

Considering that all the PE devices support the HRW election algorithm, but only a subset of them may have the capability of performing the handshake or synchronization mechanism. In such a situation, the following procedure are exercised.

If some PEs in the redundancy group signal both Handshake and Time Synchronization capabilities (both H & T set to 1), then Time Synchronization capability SHALL be chosen over Handshake capability with the HRW (or Preference-based) DF election algorithm.

If some PEs in the redundancy group signal Time Synchronization (T=1) but not Handshaking (H=0); whereas, some other PEs in the same redundancy group signal Handshaking (H=1) but not Time

Synchronization (T=0), then the PEs that have handshaking ability, SHALL perform HRW with handshaking among themselves and the PEs that Time Synchronization capability SHALL perform HRW (or Preference-based) with time synchronization among themselves.

If some PEs in the redundancy group don't signal either Time Synchronization or Handshaking capabilities, then these PEs SHALL perform HRW (or Preference-based) with default timer based mechanism defined in [RFC 7432].

In the illustration below, PE1, PE2 and PE3 send their respective Type 4 routes indicating their DF capabilities at time T1, T2 and T3 respectively. Only PE2 and PE3 are Handshake capable, hence only PE2 and PE3 partake in DF Handshaking procedure described here at time T4 and T5. PE1 on the other hand, runs the DF election timer and takes over the DF role upon timer expiry at time T6.

```
      PE1                  PE2                          PE3.br
       |                   |                           |
       |                   |                           |
       |       Type 4 (0x0 Default Capability)         |
       |----------------->>|------------------------->>| T1
       |                   |                           |
       |       Type 4 (H=1 Handshake Capable)          |
       |<<-----------------|------------------------->>| T2
       |                   |                           |
       |       Type 4 (H=1 Handshake Capable)          |
       |<<-----------------|<<-------------------------| T3
       |                   |                           |
       |                   |                           |
       |                   |      Type C (DF Request)  |
       |                   |<<-------------------------| T4
       |                   |                           |
       |                   |      Type D (DF Response) |
       |                   |------------------------->>| T5
       | PE1  Timer Expiry (DF Takeover)               | T6
       |<<###########################################>>|
       |     Only PE2 and PE3 Handshake Capable        |
       |<<###########################################>>|
       .                   .                           .
```

3.2 DF Election Synchronization Solution

If all PE devices attached to a given Ethernet Segment are clock-synchronized with each other, then the above handshaking procedures can be simplified and packet loss can be reduced from BGP-propagation time (between recovered PE and the DF PE) to very small time (e.g., milliseconds or less).

The simplified procedure is as follow:

First, the DF election procedure, described in RFC7432, is applied as before.

All PEs attached to a given Ethernet-Segment are clock-synchronized; using a networking protocol for clock synchronization (e.g. NTP, PTP, etc).

Newly inserted device PE or during failure recovery of a PE, that PE communicates the current time to peering partners plus the remaining peering timer time left. This constitute an "endtime" as see from local PE. That "endtime" is called "Service Carving Time" (SCT).

A new BGP Extended Community is advertised along with RT-4 to communicate to other partners the Service Carving Time.

Upon reception of that new BGP Extended Community, partner PEs know exactly its carving time. The notion of skew is introduced to eliminate any potential duplicate traffic or loops. They add a skew (default = -10ms) to the Service Carving Time to enforce this; basically partner PEs must carve first.

To summarize, all peering PEs carve almost simultaneously at the time announced by newly added / recovered PE. The newly added/recovered PE initiates the SCT, carves immediately on peering timer expiry. Other PE receiving RT-4 with a SCT BGP ExtComm, carve shortly before "SCT time".

3.2.3 Advantages

There are multiples advantages of using the approach. Here is a non-exhaustive list:

- A simple uni-directional signaling is all needed

- Backwards-compatible: old versions of draft/RFC shall simply discard unrecognized new SCT BGP ExtComm

- Multiple DF Election algorithms can be supported:
    * RFC7432's default ordered list ordinal algorithm (modulo)

        * HRW in [DF-FRAMEWORK], etc
   - Independent of BGP transmission delay for RT-4
   - Solutions is agnostic of the time synchronization mechanisms (e.g. NTP, PTP, ...)

3.2.4 Interoperability

Per redundancy group, for the DF election procedures to be globally convergent and unanimous, it is necessary that all the participating PEs agree on the DF Election algorithm to be used. It is, however, possible that some PEs continue to use the existing modulus based DF election and do not rely on the new SCT BGP extended community. PEs running an baseline DF election mechanism shall simply discard unrecognized new SCT BGP extended community.

A PE can indicate its willingness to support clock-synched carving by signaling the new SCT BGP extended community along with the Ethernet-Segment Route (Type-4).

3.2.5 BGP Encoding

A new BGP extended community needs to be defined to communicate the Service Carving Expected Timestamp for each Ethernet Segment.

A new transitive extended community where the Type field is 0x06, and the Sub-Type is <to be defined> is advertised along with Ethernet Segment route. Timestamp for expected Service carving is encoded as a 8-octet value as follows:

```
                        1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06   | Sub-Type(TBD) |              Timestamp(upper 16)|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                       Timestamp (lower 32)                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

This document introduces a new flag called "T" (for Time Synchronization) to the bitmap field of the DF Election Extended Community defined in [DF-FRAMWORK].

```
                        1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | Type=0x06     | Sub-Type(0x06)|   DF Type     |P|A|H|T| Bitmap|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                          Reserved = 0                         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

T: This flag is located in bit position 27 as shown above. When set to 1, it indicates the desire to use Time Synchronization capability with the rest of the PEs in the ES. This capability is used in conjunction with the agreed upon DF Type (DF Election Type). For example if all the PEs in the ES indicated that they have Time Synchronization capability and they want the DF type be of HRW, then HRW algorithm is used in conjunction with this capability.

3.2.6 Note on NTP-based synchronization

The 64-bit timestamp used by NTP protocol consists of a 32-bit part for seconds and a 32-bit part for fractional second. Giving a time scale that rolls over every 2^32 seconds (136 years) and a theoretical resolution of 2^32 seconds (233 picoseconds). The recommendation is to keep the top 32 bits and carry lower MSB 16 bits of fractional second.

3.2.7 An example

Let's take figure 1 as an example where initially PE2 had failed and PE1 had taken over.

Based on RFC-7432:

- Initial state: PE1 is in steady-state, PE2 is recovering
- PE2 recovers at (absolute) time t=99
- PE2 advertises RT-4 (sent at t=100) to partner PE1.
- PE2, it starts its 3sec peering timer as per RFC7432
- PE1 carves immediately on RT-4 reception. PE2 carves at time t=103.

With following procedure, there is a high chance to generate a traffic black hole or traffic loop. The peering timer value has a direct effect of this behavior. A short peering timer may generate loop whereas a long peering timer provide a prolong blackout.

Based on the SCT approach:

- Initial state: PE1 is in steady-state, PE2 is recovering

- PE2 recovers at (absolute) time t=99
- PE2 advertises RT-4 (sent at t=100) with target SCT value t=103 to partner PE1
- PE2 starts its 3sec peering timer as per RFC7432
- Both PE1 and PE2 carves at (absolute) time t=103; In fact, PE1 should carve slightly before PE2 (skew).

Using SCT approach, the effect of the peering timer is gone. Also, the BGP RT-4 transmission delay (from PE2 to PE1) becomes a no-op.

4 Acknowledgement Authors would like to acknowledge helpful comments and contributions of Satya Mohanty and Luc Andre Burdet.

5 Security Considerations

The mechanisms in this document use EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable. This document uses MPLS and IP-based tunnel technologies to support data plane transport. Security considerations described in [R7432] and in [ietf-evpn-overlay] are equally applicable.

6 IANA Considerations

Allocation of Extended Community Type and Sub-Type for EVPN.

7 References

7.1 Normative References

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

[RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", February, 2015.

[DF-FRAMEWORK] Rabadan, Mohanty et al., "Framework for EVPN Designated Forwarder Election Extensibility", draft-ietf-bess-evpn-df-election-framework-00, work in progress, March 5, 2018.

7.2 Informative References

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Gaurav Badoni
Cisco
Email: gbadoni@cisco.com

Patrice Brissette
Cisco
Email: pbrisset@cisco.com

Dhananjaya Rao
Cisco
Email: dhrao@cisco.com

John Drake
Juniper
Email: jdrake@juniper.net

Jorge Rabadan
Juniper
Email: jorge.rabadan@nokia.com