Workgroup: BESS Workgroup
Internet-Draft:
draft-ietf-bess-evpn-pref-df-10
Published: 2 September 2022
Intended Status: Standards Track
Expires: 6 March 2023
Authors: J. Rabadan, Ed. Nokia
S. Sathappan Nokia
W. Lin Juniper Networks
J. Drake Juniper Networks
A. Sajassi Cisco Systems

Preference-based EVPN DF Election

Abstract

The Designated Forwarder (DF) in Ethernet Virtual Private Networks (EVPN) is defined as the PE responsible for sending Broadcast, Unknown unicast and Broadcast traffic (BUM) to a multi-homed device/network in the case of an all-active multi-homing Ethernet Segment (ES), or BUM and unicast in the case of single-active multi-homing. The Designated Forwarder is selected out of a candidate list of PEs that advertise the same Ethernet Segment Identifier (ESI) to the EVPN network, according to the Default Designated Forwarder Election algorithm. While the Default Algorithm provides an efficient and automated way of selecting the Designated Forwarder across different Ethernet Tags in the Ethernet Segment, there are some use cases where a more 'deterministic' and user-controlled method is required. At the same time, Service Providers require an easy way to force an on-demand Designated Forwarder switchover in order to carry out some maintenance tasks on the existing Designated Forwarder or control whether a new active PE can preempt the existing Designated Forwarder PE.

This document proposes a Designated Forwarder Election algorithm that meets the requirements of determinism and operation control.

Status of This Memo

This Internet-Draft will expire on 6 March 2023.

Table of Contents

1. Introduction

1.1. Problem Statement

[RFC7432] defines the Designated Forwarder (DF) in EVPN networks as the PE responsible for sending Broadcast, Multicast and Unknown unicast traffic (BUM) to a multi-homed device/network in the case of an all-active multi-homing Ethernet Segment or BUM and unicast traffic to a multi-homed device or network in case of single-active multi-homing. The Designated Forwarder is selected out of a candidate list of PEs that advertise the Ethernet Segment Identifier

(ESI) to the EVPN network and according to the Designated Forwarder Election Algorithm, or DF Alg as per [RFC8584].

While the Default Designated Forwarder Algorithm [RFC7432] or the Highest Random Weight algorithm (HRW) [RFC8584] provide an efficient and automated way of selecting the Designated Forwarder across different Ethernet Tags in the Ethernet Segment, there are some use-cases where a more 'deterministic' and user-controlled method is required. At the same time, Service Providers require an easy way to force an on-demand Designated Forwarder switchover in order to carry out some maintenance tasks on the existing Designated Forwarder or control whether a new active PE can preempt the existing Designated Forwarder PE.

This document proposes two new DF Algs (Highest-Preference and Lowest-Preference) which provide the deterministic Designated Forwarder method required, as well as the "Don't Preempt" capability to address the need to control whether a PE can take over an existing Designated Forwarder PE.

1.2. Solution requirements

The procedures described in this document meet the following requirements:

a. The solution provides an administrative preference option so that the user can control in what order the candidate PEs may become Designated Forwarder, assuming they are all operationally ready to take over as Designated Forwarder. The operator can determine whether the Highest-Preference or Lowest-Preference PE among the PEs in the Ethernet Segment will be elected as Designated Forwarder, based on the DF Algs described in this document.

b. The extensions in this document work for [RFC7432] Ethernet Segments and virtual Ethernet Segments, as defined in [I-D.ietf-bess-evpn-virtual-eth-segment].

c. The user may force a PE to preempt the existing Designated Forwarder for a given Ethernet Tag without re-configuring all the PEs in the Ethernet Segment, by simply modifying the existing administrative preference in that PE.

d. The solution allows an option to NOT preempt the current Designated Forwarder ("Don't Preempt" capability), even if the former Designated Forwarder PE comes back up after a failure. This is also known as "non-revertive" behavior, as opposed to the [RFC7432] Designated Forwarder election procedures that are always revertive (because the winner PE of the default

Designated Forwarder election algorithm always takes over as the operational Designated Forwarder).

e. The procedures described in this document support single-active and all-active multi-homing Ethernet Segments.

2. Requirements Language and Terminology

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.

*AC - Attachment Circuit. An AC has an Ethernet Tag associated to it.

*BUM - refers to the Broadcast, Unknown unicast and Multicast traffic.

*DF, NDF and BDF - Designated Forwarder, Non-Designated Forwarder and Backup Designated Forwarder.

*DF Alg - refers to Designated Forwarder Election Algorithm. This is sometimes shortened to "Alg" in this document.

*HRW - Highest Random Weight, as per [RFC8584].

*ES, vES and ESI - Ethernet Segment, virtual Ethernet Segment and Ethernet Segment Identifier.

*EVI - EVPN Instance.

*ISID - refers to Service Instance Identifiers in Provider Backbone Bridging (PBB) networks.

*MAC-VRF - A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

*BD - Broadcast Domain. An EVI may be comprised of one (VLAN-Based or VLAN Bundle services) or multiple (VLAN-Aware Bundle services) Broadcast Domains.

*EVC - Ethernet Virtual Circuit.

*DP - refers to the "Don't Preempt me" capability in the Designated Forwarder Election extended community.

*OAM - refers to Operations And Maintenance protocols.

*Ethernet A-D per ES route - refers to [RFC7432] route type 1 or Auto-Discovery per Ethernet Segment route.

*Ethernet A-D per EVI route - refers to [RFC7432] route type 1 or Auto-Discovery per EVPN Instance route.

*Ethernet Tag - used to represent a Broadcast Domain that is configured on a given Ethernet Segment for the purpose of Designated Forwarder election. Note that any of the following may be used to represent a Broadcast Domain: VIDs (including Q-in-Q tags), configured IDs, VNI (VXLAN Network Identifiers), normalized VID, I-SIDs (Service Instance Identifiers), etc., as long as the representation of the broadcast domains is configured consistently across the multi-homed PEs attached to that Ethernet Segment. The Ethernet Tag value MUST be different from zero.

3. EVPN BGP Attributes Extensions

This solution reuses and extends the Designated Forwarder Election Extended Community defined in [RFC8584] that is advertised along with the Ethernet Segment route, by replacing the last two reserved octets of the DF Election Extended Community when the DF Alg is set to Highest-Preference or Lowest-Preference. This document also defines a new capability referred to as "Don't Preempt" capability, that MAY be used with DF Algs Highest-Preference or Lowest-Preference. The format of the DF Election Extended Community that is used in this document follows:

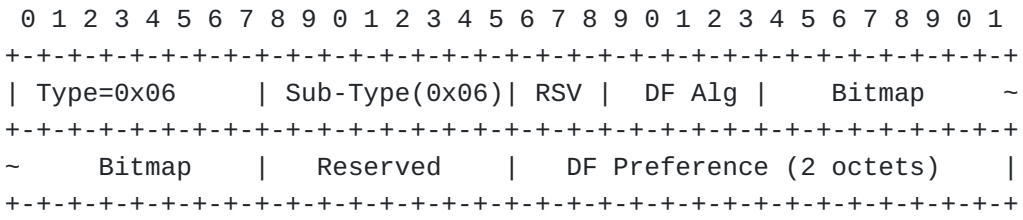

Figure 1: DF Election Extended Community

Where the above fields are defined as follows:

*DF Alg can have the following values:

-Alg 0 - Default Designated Forwarder Election algorithm, or modulus-based algorithm as per [RFC7432].

-Alg 1 - HRW algorithm as per [RFC8584].

-Alg 2 - Highest-Preference algorithm (this document Section 4.1).

-Alg TBD - Lowest-Preference algorithm (this document Section 4.1). TBD will be replaced by the allocated value at the time of publication.

*Bitmap (2 octets) encodes "capabilities" [RFC8584], where this document defines the "Don't Preempt" capability, used to indicate if a PE supports a non-revertive behavior:

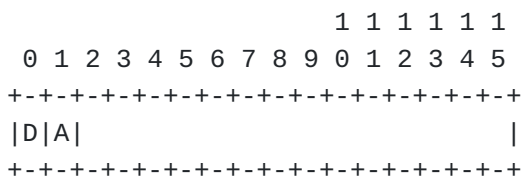

Figure 2: Bitmap field in the DF Election Extended Community

-Bit 0 (corresponds to Bit 24 of the Designated Forwarder Election Extended Community and it is defined by this document): D bit or 'Don't Preempt' bit (DP hereafter), determines if the PE advertising the Ethernet Segment route requests the remote PEs in the Ethernet Segment not to preempt it as Designated Forwarder. The default value is DP=0, which is compatible with the 'preempt' or 'revertive' behavior in the Default DF Alg [RFC7432]. The DP capability is supported by DF Algs Highest-Preference or Lowest-Preference, and MAY be used with the default DF Alg or HRW [RFC8584]. The procedures of the "Don't Preempt" capability for the default DF Alg or HRW are out of the scope of this document. The procedures of the "Don't Preempt" capability for DF Algs Highest-Preference and Lowest-Preference are described in Section 4.1.

-Bit 1: AC-DF or AC-Influenced Designated Forwarder Election is described in [RFC8584]. When set to 1, it indicates the desire to use AC-Influenced Designated Forwarder Election with the rest of the PEs in the Ethernet Segment. The AC-DF capability bit MAY be set along with the DP capability and DF Algs Highest-Preference or Lowest-Preference.

*Designated Forwarder (DF) Preference (described in this document): defines a 2-octet value that indicates the PE preference to become the Designated Forwarder in the Ethernet Segment, as described in Section 4.1. The allowed values are within the range 0-65535, and the default value MUST be 32767. This value is the midpoint in the allowed Preference range of values, which gives the operator the flexibility of choosing a significant number of values, above or below the default Preference. A numerically higher or lower value of this field is more preferred for Designated Forwarder election depending on the

DF Alg being used, as explained in Section 4.1. The Designated Forwarder Preference field is specific to DF Algs Highest-Preference and Lowest-Preference, and this document does not define any meaning for other algorithms. If the DF Alg is different from Highest-Preference or Lowest-Preference, these two octets can be encoded differently.

*RSV and Reserved fields (from bit 16 to bit 18, and from bit 40 to 47): when DF Alg is set to Highest-Preference or Lowest-Preference algorithm, the values are set to zero when advertising the Ethernet Segment route, and they are ignored when receiving the Ethernet Segment route.

4. Solution description

Figure 3 illustrates an example that will be used in the description of the solution.

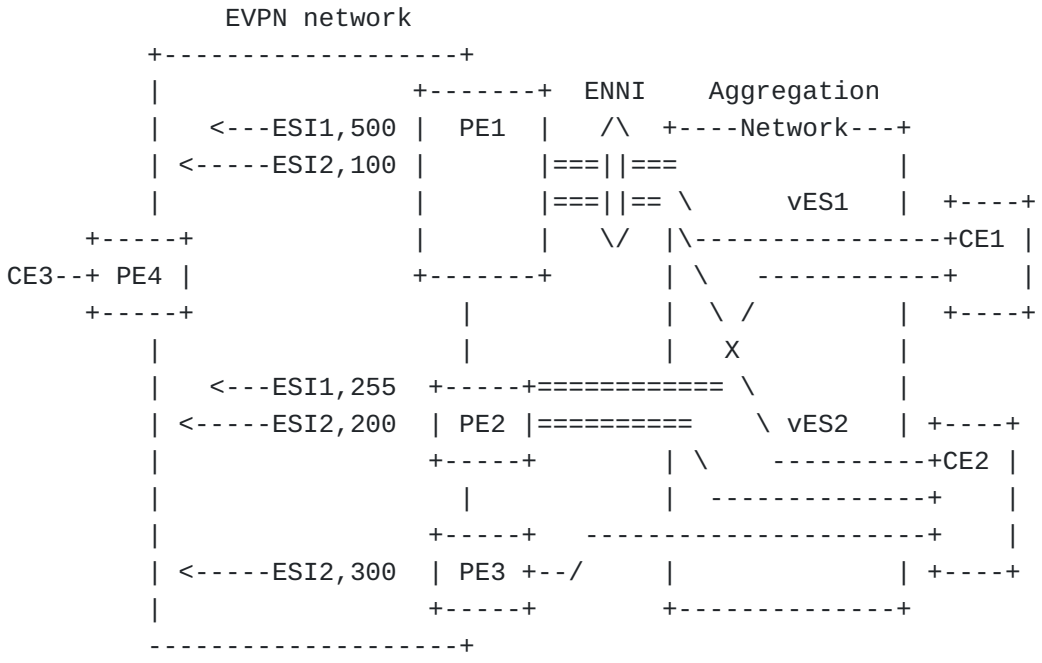

Figure 3: Preference-based DF Election

Figure 3 shows three PEs that are connecting EVCs coming from the Aggregation Network to their EVIs in the EVPN network. CE1 is connected to vES1 - that spans PE1 and PE2 - and CE2 is connected to vES2, that is attached to PE1, PE2 and PE3.

If the algorithm chosen for vES1 and vES2 is DF Alg Highest-Preference or Lowest-Preference, the PEs may become Designated Forwarder irrespective of their IP address and based on the administrative Preference value. The following sections provide some

examples of the procedures and how they are applied in the use-case of Figure 3.

4.1. Use of the Highest-Preference and Lowest Preference Algorithm

Assuming the operator wants to control - in a flexible way - what PE becomes the Designated Forwarder for a given virtual Ethernet Segment and the order in which the PEs become Designated Forwarder in case of multiple failures, the following procedure may be used:

a. vES1 and vES2 are now configurable with three optional parameters that are signaled in the Designated Forwarder Election extended community. These parameters are the Preference, Preemption option (or "Don't Preempt" option) and DF Alg. We will represent these parameters as (Pref,DP,Alg). For instance, vES1 (Pref,DP,Alg) is configured as (500,0,Highest-Preference) in PE1, and (255,0,Highest-Preference) in PE2. vES2 is configured as (100,0,Highest-Preferance), (200,0,Highest-Preference) and (300,0,Highest-Preference) in PE1, PE2 and PE3 respectively.

b. The PEs advertise an Ethernet Segment route for each virtual Ethernet Segment, including the three parameters indicated in 'a' above, in the Designated Forwarder Election Extended Community Section 3.

c. According to [RFC8584], each PE will run the Designated Forwarder election algorithm upon expiration of the DF Wait timer. Each PE runs the Highest-Preference or Lowest-Preference DF Alg for each Ethernet Segment as follows:

   * The PE will check the DF Alg value in each Ethernet Segment route, and assuming all the Ethernet Segment routes (including the local route) are consistent in this DF Alg (that is, all are configured for Highest-Preference or Lowest-Preference, but not a mix), the PE runs the procedure in this section. Otherwise, the procedure falls back to [RFC7432] Default Alg. The Highest-Preference and Lowest-Preference Algs are different Algs, therefore if two PEs configured for Highest-Preference and Lowest-Preference respectively, are attached to the same Ethernet Segment, the operational Designated Forwarder Election Alg will fall back to the Default Alg.

   * If all the PEs attached to the Ethernet Segment advertise Highest-Preference Alg, each PE builds a list of candidate PEs, ordered by Preference value from the numerically highest value to lowest value. E.g., PE1 builds a list of candidate PEs for vES1 ordered by the Preference, from high

to low: <PE1, PE2> (since PE1's preference is more preferred than PE2's). Hence PE1 becomes the Designated Forwarder for vES1. In the same way, PE3 becomes the Designated Forwarder for vES2.

*If all the PEs attached to the Ethernet Segment advertise Lowest-Preference Alg, then the candidate lits is ordered from the numerically lowest Preference value to the highest Preference value. E.g., PE1's ordered list for vES1 is <PE2, PE1>. Hence PE2 becomes the Designated Forwarder for vES1. In the same way, PE1 becomes the Designated Forwarder for vES2.

d. Assuming some maintenance tasks had to be executed on a PE the operator may want to make sure the PE is not the Designated Forwarder for the Ethernet Segment so that the impact on the service is minimized. E.g., if PE3 is going on maintenance and the DF Alg is Highest-Preference, the operator could change vES2's Preference on PE3 from 300 to E.g., 50 (hence the Ethernet Segment route from PE3 is updated with the new preference value) so that PE2 is forced to take over as Designated Forwarder for vES2 (irrespective of the DP capability). Once the maintenance task on PE3 is over, the operator could decide to leave the latest configured preference value or configure the initial preference value back. A similar procedure can be used for DF Alg Lowest-Preference too.

e. In case of equal Preference in two or more PEs in the Ethernet Segment, the DP bit and the numerically lowest IP address of the candidate PEs are used as tie-breakers. If more that one PE is advertising itself as the preferred Designated Forwarder, an implementation MUST first select the PE advertising the DP bit set, and then select the PE with the lowest IP address (if the DP bit selection does not yield a unique candidate). The PE's IP address is the address used in the candidate list and it is derived from the Originating Router's IP address of the Ethernet Segment route. In case PEs use Originating Router's IP address of different families, an IPv4 address is always considered numerically lower than an IPv6 address. Some examples of the use of the DP bit and IP address tie-breakers follow:

*If vES1 parameters were (500,0,Highest-Preference) in PE1 and (500,1,Highest-Preference) in PE2, PE2 would be elected due to the DP bit. Same example applies if PE1 and PE2 advertise Lowest-Preference DF Alg instead.

*If vES1 parameters were (500,0,Highest-Preference) in PE1 and (500,0,Highest-Preference) in PE2, PE1 would be elected,

if PE1's IP address is lower than PE2's. Or PE2 would be elected if PE2's IP addres is lower than PE1's. Same example applies if PE1 and PE2 advertise Lowest-Preference DF Alg instead.

f. The Preference is an administrative option that MUST be configured on a per-Ethernet Segment basis, and it is normally configured from the management plane. The Preference value MAY also be dynamically changed based on the use of local policies that react to events on the PE. The following examples illustrate the use of local policy to change the Preference value in a dynamic way.

   E.g., on PE1, if the DF Alg is Highest-Preference, ES1's Preference value can be lowered from 500 to 100 in case the bandwidth on the ENNI port is decreased a 50% (that could happen if e.g. the 2-port LAG between PE1 and the Aggregation Network loses one port).

   Local policy MAY also trigger dynamic Preference changes based on the PE's bandwidth availability in the core, specific ports going operationally down, etc.

   The definition of the actual local policies is out of scope of this document.

The Highest-Preference and Lowest-Preference Algs MAY be used along with the AC-DF capability. Assuming all the PEs in the Ethernet Segment are configured consistently with Highest-Preference or Lowest-Preference Alg and AC-DF capability, a given PE in the Ethernet Segment is not considered as candidate for Designated Forwarder Election until its corresponding Ethernet A-D per ES and Ethernet A-D per EVI routes are not received, as described in [RFC8584].

The Highest-Preference and Lowest-Preference DF Algs can be used in different virtual Ethernet Segments on the same PE. For instance, PE1 and PE2 can use Highest-Preference for vES1 and PE1, PE2 and PE3 Lowest-Preference for vES2. The use of one DF Alg over the other is the operator's choice. The existence of both provide flexibility and full control to the operator.

The procedures in this document can be used in [RFC7432] based Ethernet Segment or virtual Ethernet Segment as in [I-D.ietf-bess-evpn-virtual-eth-segment], and including EVPN networks as in [RFC8214], [RFC7623] or [RFC8365].

4.2. Use of the Highest-Preference algorithm in [RFC7432] Ethernet Segments

While the Highest-Preference or Lowest-Preference DF Alg described in Section 4.1 is typically used in virtual Ethernet Segment scenarios where there is normally an individual Ethernet Tag per virtual Ethernet Segment, the existing [RFC7432] definition of an Ethernet Segment allows potentially up to thousands of Ethernet Tags on the same Ethernet Segment. If this is the case, if Highest-Preference or Lowest-Preference Alg is configured in all the PEs of the Ethernet Segment, the same PE will be the elected Designated Forwarder for all the Ethernet Tags of the Ethernet Segment. A potential way to achive a more granular load balancing is decribed below.

The Ethernet Segment is configured with an administrative Preference value and an administrative DF Alg, i.e., Highest-Preference or Lowest-Preference Alg. However, the administrative DF Alg (which is used to signal the DF Alg for the Ethernet Segment) MAY be overridden to a different operational DF Alg for a range of Ethernet Tags. With this option, the PE builds a list of candidate PEs ordered by Preference, however the Designated Forwarder for a given Ethernet Tag will be determined by the local overridden DF Alg.

For instance:

- Assuming ES3 is defined in PE1 and PE2, PE1 may be configured as (500,0,Highest-Preference) for ES3 and PE2 as (100,0,Highest-Preference). Both PEs will advertise the Ethernet Segment routes for ES3 with the indicated parameters in the DF Election Extended Community.

- In addition, assuming VLAN-based service interfaces and that the PEs are attached to all Ethernet Tags in the range 1-4000, both PE1 and PE2 may be configured with (Ethernet Tag-range,Lowest-Preference), E.g., (2001-4000, Lowest-Preference).

- This will result in PE1 being Designated Forwarder for Ethernet Tags 1-2000 (since they use the default Highest-Preference Alg) and PE2 being Designated Forwarder for Ethernet Tags 2001-4000, due to the local policy overriding the Highest-Preference Alg.

For Ethernet Segments attached to three or more PEs, any other logic that provides a fair distribution of the Designated Forwarder function among the PEs is valid, as long as that logic is consistent in all the PEs in the Ethernet Segment. It is important to note that, when a local policy overrides the Highest-Preference or Lowest-Preference signaled by all the PEs in the Ethernet Segment, this local policy MUST be consistent in all the PEs of the Ethernet

Segment. If the local policy is inconsistent for a given Ethernet Tag in the Ethernet Segment, black-holes or packet duplication may occur on that Ethernet Tag.

4.3. The Non-Revertive Capability

As discussed in Section 1.2 (d), a capability to NOT preempt the existing Designated Forwarder (for all the Ethernet Tags in the Ethernet Segment) is required and therefore added to the Designated Forwarder Election extended community. This option allows a non-revertive behavior in the Designated Forwarder election.

Note that, when a given PE in an Ethernet Segment is taken down for maintenance operations, before bringing it back, the Preference may be changed in order to provide a non-revertive behavior. The DP bit and the mechanism explained in this section will be used for those cases when a former Designated Forwarder comes back up without any controlled maintenance operation, and the non-revertive option is desired in order to avoid service impact.

In Figure 3, we assume that based on the Highest-Preference Alg, PE3 is the Designated Forwarder for ESI2.

If PE3 has a link, EVC or node failure, PE2 would take over as Designated Forwarder. If/when PE3 comes back up again, PE3 will take over, causing some unnecessary packet loss in the Ethernet Segment.

The following procedure avoids preemption upon failure recovery (please refer to Figure 3). The procedure supports a non-revertive mode that can be used along with:

- *Highest-Preference Alg

- *Highest-Preference Alg, where a local policy overrides the Highest-Preference tie-breaker for a range of Ethernet Tags

- *Lowest-Preference Alg

The procedure is described assuming Highest-Preference Alg in the Ethernet Segment, where local policy overrides the tie-breaker for a given Ethernet Tag, since this is the most complex case. The other two cases above are a sub-set of this one and the differences will be explained later.

1. A "Don't Preempt" capability is defined on a per-PE/per-Ethernet Segment basis, as described in Section 3. If "Don't Preempt" is disabled (default behavior), the PE sets DP to zero and advertises it in an Ethernet Segment route. If "Don't Preempt Me" is enabled, the Ethernet Segment route from the PE will indicate the desire of not being preempted by the other

PEs in the Ethernet Segment. All the PEs in an Ethernet Segment SHOULD be consistent in their configuration of the DP capability, however this document does not enforce the consistency across all the PEs. In case of inconsistency in the support of the DP capability in the PEs of the same Ethernet Segment, non-revertive behavior is not guaranteed. However, PEs supporting this capability will still attempt this procedure.

2. We assume we want to avoid 'preemption' in all the PEs in the Ethernet Segment, the three PEs are configured with the "Don't Preempt" capability. In this example, we assume ESI2 is configured as 'DP=enabled' in the three PEs.

3. We also assume vES2 is attached to Ethernet Tag-1 and Ethernet Tag-2. vES2 uses Highest-Preference as DF Alg and a local policy is configured in the three PEs to use Lowest-Preference for Ethernet Tag-2. When vES2 is enabled in the three PEs, the PEs will exchange the Ethernet Segment routes and select PE3 as Designated Forwarder for Ethernet Tag-1 (due to the Highest-Preference), and PE1 as Designated Forwarder for Ethernet Tag-2 (due to the Lowest-Preference).

4. If PE3's vES2 goes down (due to EVC failure - detected by OAM, or port failure or node failure), PE2 will become the Designated Forwarder for Ethernet Tag-1. No changes will occur for Ethernet Tag-2.

5. When PE3's vES2 comes back up, PE3 will start a boot-timer (if booting up) or hold-timer (if the port or EVC recovers). That timer will allow some time for PE3 to receive the Ethernet Segment routes from PE1 and PE2. This timer is applied between the INIT and the DF_WAIT states in the Designated Forwarder Election Finite State Machine described in [RFC8584]. PE3 will then:

   *Select two "reference-PEs" among the Ethernet Segment routes in the virtual Ethernet Segment, the "Highest-PE" and the "Lowest-PE":

     -The Highest-PE is the PE with higher Preference, using the DP bit first (with DP=1 being better) and, after that, the lower PE-IP address as tie-breakers. PE3 will select PE2 as Highest-PE over PE1, since, when comparing (Pref,DP,PE-IP), (200,1,PE2-IP) wins over (100,1,PE1-IP).

     -The Lowest-PE is the PE with lower Preference, using the DP bit first (with DP=1 being better) and, after that, the lower PE-IP address as tie-breakers. PE3 will select

PE1 as Lowest-PE over PE2, since (100,1,PE1-IP) wins over (200,1,PE2-IP).

- Note that if there were only one remote PE in the Ethernet Segment, Lowest and Highest PE would be the same PE.

* Check its own administrative Pref and compare it with the one of the Highest-PE and Lowest-PE that have the DP capability set in their Ethernet Segment routes. Depending on this comparison PE3 will send the Ethernet Segment route with a (Pref,DP) that may be different from its administrative (Pref,DP):

  - If PE3's Pref value is higher or equal than the Highest-PE's, PE3 will send the Ethernet Segment route with an 'in-use' operational Pref equal to the Highest-PE's and DP=0.

  - If PE3's Pref value is lower or equal than the Lowest-PE's, PE3 will send the Ethernet Segment route with an 'in-use' operational Preference equal to the Lowest-PE's and DP=0.

  - If PE3's Pref value is not higher or equal than the Highest-PE's and is not lower or equal than the Lowest-PE's, PE3 will send the Ethernet Segment route with its administrative (Pref,DP)=(300,1).

  - In this example, PE3's administrative Pref=300 is higher than the Highest-PE with DP=1, that is, PE2 (Pref=200). Hence PE3 will inherit PE2's preference and send the Ethernet Segment route with an operational 'in-use' (Pref,DP)=(200,0).

* Note that, a PE will always send its DP capability set to zero as long as the advertised Pref is the 'in-use' operational Pref (as opposed to the 'administrative' Pref).

* This Ethernet Segment route update sent by PE3, with (200,0,PE3-IP), will not cause any Designated Forwarder switchover for any Ethernet Tag. PE2 will continue being Designated Forwarder for Ethernet Tag-1. This is because the DP bit will be used as a tie-breaker in the Designated Forwarder election. That is, if a PE has two candidate PEs with the same Pref, it will pick up the one with DP=1. There are no Designated Forwarder changes for Ethernet Tag-2 either.

6. For any subsequent received update/withdraw in the Ethernet Segment, the PEs will go through the process described in (5) to select Highest and Lowest-PEs, now considering themselves as candidates. For instance, if PE2 fails, upon receiving PE2's Ethernet Segment route withdrawal, PE3 and PE1 will go through the selection of new Highest and Lowest-PEs (considering their own active Ethernet Segment route) and then they will run the Designated Forwarder Election.

   *If a PE selects itself as new Highest or Lowest-PE and it was not before, the PE will then compare its operational 'in-use' Pref with its administrative Pref. If different, the PE will send an Ethernet Segment route update with its administrative Pref and DP values. In the example, PE3 will be the new Highest-PE, therefore it will send an Ethernet Segment route update with (Pref,DP)=(300,1).

   *After running the Designated Forwarder Election, PE3 will become the new Designated Forwarder for Ethernet Tag-1. No changes will occur for Ethernet Tag-2.

If the Ethernet Segment uses Highest-Preference Alg (for all the Ethernet Tags, no local policy), the PEs only need to select the "Highest-PE" as the "reference-PE" (i.e., no need to select the "Lowest-PE"). If the Ethernet Segment uses Lowest-Preference Alg for all the Ethernet Tags, the PEs only need to select the "Lowest-PE" as the "reference-PE". The rest of the procedure remains the same.

Note that, irrespective of the DP bit, when a PE or Ethernet Segment comes back and the PE advertises a Designated Forwarder Election Alg different than the one configured in the rest of the PEs in the Ethernet Segment, all the PEs in the Ethernet Segment MUST fall back to the Default [RFC7432] Alg.

This document does not modify the use of the P and B bits in the Ethernet A-D per EVI routes [RFC8214] advertised by the PEs in the Ethernet Segment after running the Designated Forwarder Election, irrespective of the revertive or non-revertive behavior in the PE.

5. Security Considerations

This document describes a Designated Forwarder Election Algorithm that provides absolute control (by configuration) over what PE is the Designated Forwarder for a given Ethernet Tag. While this control is desired in many situations, a malicious user that gets access to the configuration of a PE in the Ethernet Segment may change the behavior of the network. In other DF Algs such as HRW, the Designated Forwarder Election is more automated and cannot be determined by configuration. With Highest-Preference or Lowest-

Preference as DF Alg, an attacker may change the configuration of the Preference value on a PE and Ethernet Segment, and impact the traffic going through that PE and Ethernet Segment.

The non-revertive capability described in this document may be seen as a security improvement over the regular EVPN revertive Designated Forwarder Election: an intentional link (or node) "flapping" on a PE will only cause service disruption once, when the PE goes to Non-Designated Forwarder state.

The document also describes how a local policy can override the Highest-Preference Alg for a range of Ethernet Tags in the Ethernet Segment. If the local policy is not consistent across all PEs in the Ethernet Segment and there is an Ethernet Tag that ends up with an inconsistent use of Highest-Preference or Lowest-Preference in different PEs, black-holing or packet duplication may occur for that Ethernet Tag.

6. IANA Considerations

This document solicits:

* The allocation of two new values in the "DF Alg" registry created by [RFC8584] as follows:

Alg	Name	Reference
2	Highest-Preference Algorithm	This document
TBD	Lowest-Preference Algorithm	This document

* The allocation of a new value in the "DF Election Capabilities" registry created by [RFC8584] for the 2-octet Bitmap field in the DF Election Extended Community (Border gateway Protocol (BGP) Extended Communities registry), as follows:

Bit	Name	Reference
0	D (Don't Preempt) Capability	This document

7. Acknowledgments

The authors would like to thank Kishore Tiruveedhula and Sasha Vainshtein for their review and comments. Also thank you to Luc Andre Burdet and Stephane Litkowski for their thorough review and suggestions for a new DF Alg for lowest-preference.

8. Contributors

In addition to the authors listed, the following individuals also contributed to this document:

Tony Przygienda, Juniper

Satya Mohanty, Cisco

Kiran Nagaraj, Nokia

Vinod Prabhu, Nokia

Selvakumar Sivaraj, Juniper

Sami Boutros, VMWare

9. References

9.1. 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, <https://www.rfc-editor.org/info/rfc7432>.

[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

[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>.

[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>.

[I-D.ietf-bess-evpn-virtual-eth-segment]
Sajassi, A., Brissette, P., Schell, R., Drake, J. E., and J. Rabadan, "EVPN Virtual Ethernet Segment", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-virtual-eth-segment-07, 6 July 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-virtual-eth-segment-07.txt>.

9.2. Informative References

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC7623] Sajassi, A., Ed., Salam, S., Bitar, N., Isaac, A., and W. Henderickx, "Provider Backbone Bridging Combined with Ethernet VPN (PBB-EVPN)", RFC 7623, DOI 10.17487/RFC7623, September 2015, <https://www.rfc-editor.org/info/rfc7623>.

Authors' Addresses

J. Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America

Email: jorge.rabadan@nokia.com

S. Sathappan
Nokia

Email: senthil.sathappan@nokia.com

W. Lin
Juniper Networks

Email: wlin@juniper.net

J. Drake
Juniper Networks

Email: jdrake@juniper.net

A. Sajassi
Cisco Systems

Email: sajassi@cisco.com