

BESS Workgroup  
Internet-Draft  
Intended status: Standards Track  
Expires: March 27, 2022

J. Rabadan, Ed.  
S. Sathappan  
Nokia  
T. Przygienda  
W. Lin  
J. Drake  
Juniper Networks  
A. Sajassi  
S. Mohanty  
Cisco Systems  
September 23, 2021

Preference-based EVPN DF Election  
draft-ietf-bess-evpn-pref-df-08

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 DF 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 DF Election algorithm. While the Default Algorithm provides an efficient and automated way of selecting the DF across different Ethernet Tags in the ES, 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 DF switchover in order to carry out some maintenance tasks on the existing DF or control whether a new active PE can preempt the existing DF PE.

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

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Draft

Preference-based EVPN DF Election

September 2021

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."

This Internet-Draft will expire on March 27, 2022.

## Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

<a href="#">1.</a>	Introduction . . . . .	<a href="#">3</a>
<a href="#">1.1.</a>	Problem Statement . . . . .	<a href="#">3</a>
<a href="#">1.2.</a>	Solution requirements . . . . .	<a href="#">3</a>
<a href="#">2.</a>	Requirements Language and Terminology . . . . .	<a href="#">4</a>
<a href="#">3.</a>	EVPN BGP Attributes Extensions . . . . .	<a href="#">5</a>
<a href="#">4.</a>	Solution description . . . . .	<a href="#">6</a>
<a href="#">4.1.</a>	Use of the Highest-Preference Algorithm . . . . .	<a href="#">7</a>
<a href="#">4.2.</a>	Use of the Lowest-Preference Algorithm . . . . .	<a href="#">9</a>
4.3.	Use of the Highest-Preference algorithm in <a href="#">[RFC7432]</a> Ethernet Segments . . . . .	<a href="#">9</a>
<a href="#">4.4.</a>	The Non-Revertive Capability . . . . .	<a href="#">10</a>
<a href="#">5.</a>	Security Considerations . . . . .	<a href="#">14</a>
<a href="#">6.</a>	IANA Considerations . . . . .	<a href="#">14</a>
<a href="#">7.</a>	Acknowledgments . . . . .	<a href="#">14</a>
<a href="#">8.</a>	Contributors . . . . .	<a href="#">15</a>
<a href="#">9.</a>	References . . . . .	<a href="#">15</a>
<a href="#">9.1.</a>	Normative References . . . . .	<a href="#">15</a>
<a href="#">9.2.</a>	Informative References . . . . .	<a href="#">16</a>
	Authors' Addresses . . . . .	<a href="#">16</a>

## [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 ES or BUM and unicast traffic to a multi-homed device or network in case of single-active multi-homing. The DF is selected out of a candidate list of PEs that advertise the Ethernet Segment Identifier (ESI) to the EVPN network and according to the DF Election Algorithm, or DF Alg as per [[RFC8584](#)].

While the Default DF Alg [[RFC7432](#)] or HRW [[RFC8584](#)] provide an efficient and automated way of selecting the DF across different Ethernet Tags in the ES, 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 DF switchover in order to carry out some maintenance tasks on the existing DF or control whether a new active PE can preempt the existing DF PE.

This document proposes a new DF Alg and capability to address the above needs.

### [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 DF, assuming they are all operationally ready to take over as DF.
- b. This extension works for [[RFC7432](#)] Ethernet Segments and virtual ES, as defined in [[I-D.ietf-bess-evpn-virtual-eth-segment](#)].

- c. The user may force a PE to preempt the existing DF for a given Ethernet Tag without re-configuring all the PEs in the ES.
- d. The solution allows an option to NOT preempt the current DF, even if the former DF PE comes back up after a failure. This is also known as "non-revertive" behavior, as opposed to the [\[RFC7432\]](#) DF election procedures that are always revertive.
- e. The solution works for 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 [BCP14](#) [\[RFC2119\]](#) [\[RFC8174\]](#) when, and only when, they appear in all capitals, as shown here.

- o AC - Attachment Circuit. An AC has an Ethernet Tag associated to it.
- o BUM - refers to the Broadcast, Unknown unicast and Multicast traffic.
- o DF, NDF and BDF - Designated Forwarder, Non-Designated Forwarder and Backup Designated Forwarder.
- o DF Alg or simply Alg - refers to Designated Forwarder Election Algorithm.
- o HRW - Highest Random Weight, as per [\[RFC8584\]](#).
- o ES, vES and ESI - Ethernet Segment, virtual Ethernet Segment and Ethernet Segment Identifier.
- o EVI - EVPN Instance.
- o ISID - refers to Service Instance Identifiers in Provider Backbone Bridging (PBB) networks.

- o MAC-VRF - A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.
- o BD - Broadcast Domain. An EVI may be comprised of one (VLAN-Based or VLAN Bundle services) or multiple (VLAN-Aware Bundle services) Broadcast Domains.
- o EVC - Ethernet Virtual Circuit.
- o DP - refers to the "Don't Preempt me" capability in the DF Election extended community.
- o OAM - refers to Operations And Maintenance protocols.
- o Ethernet A-D per ES route - refers to [[RFC7432](#)] route type 1 or Auto-Discovery per Ethernet Segment route.

- o Ethernet A-D per EVI route - refers to [[RFC7432](#)] route type 1 or Auto-Discovery per EVPN Instance route.
- o Ethernet Tag - used to represent a Broadcast Domain that is configured on a given ES for the purpose of DF 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 ES. The Ethernet Tag value MUST be different from zero.

### 3. EVPN BGP Attributes Extensions

This solution reuses and extends the DF Election Extended Community defined in [[RFC8584](#)] that is advertised along with the ES route:

```

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)| RSV | DF Alg |      Bitmap      ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

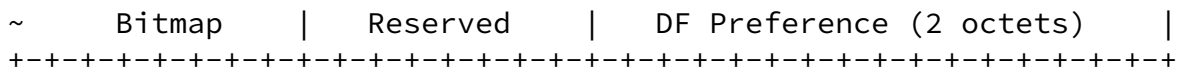


Figure 1: DF Election Extended Community

Where the following fields are defined as follows:

- o DF Alg can have the following values:
  - Alg 0 - Default DF Election algorithm, or modulus-based algorithm as per [[RFC7432](#)].
  - Alg 1 - HRW algorithm as per [[RFC8584](#)].
  - Alg 2 - Highest-Preference algorithm (this document).
  - Alg TBD - Lowest-Preference algorithm (this document). TBD will be replaced by the allocated value at the time of publication.
- o Bitmap (2 octets) can have the following values:

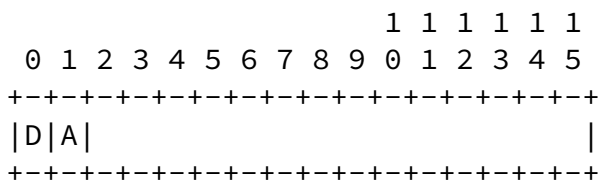


Figure 2: Bitmap field in the DF Election Extended Community

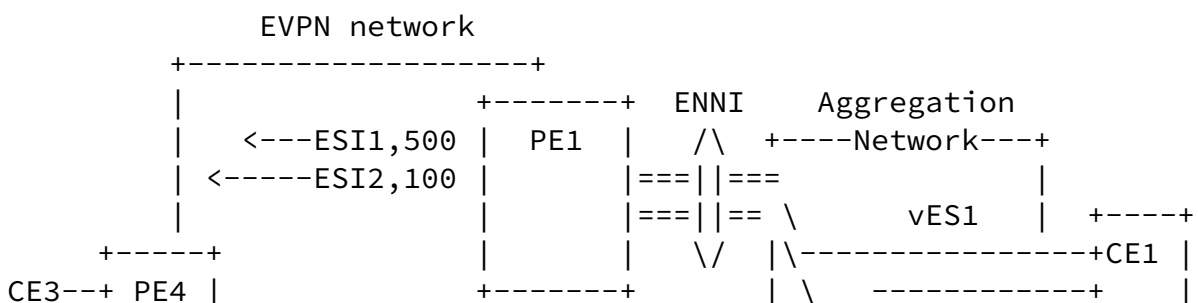
- Bit 0 (corresponds to Bit 24 of the DF Election Extended Community and it is defined by this document): D bit or 'Don't Preempt' bit (DP hereafter), determines if the PE advertising the ES route requests the remote PEs in the ES not to preempt it as DF. The default value is DP=0, which is compatible with

the 'preempt' or 'revertive' behavior in the Default DF Alg [RFC7432]. The DP capability is supported by Alg 2 and Alg TBD, and MAY be used with DF Alg 0 or 1. The procedures of the DP capability for DF Alg 0 or 1 are out of the scope of this document.

- Bit 1: AC-DF or AC-Influenced DF Election, as explained in [RFC8584]. When set to 1, it indicates the desire to use AC-Influenced DF Election with the rest of the PEs in the ES. The AC-DF capability bit MAY be set along with the DP capability and DF Alg 2 or Alg TBD.
- o DF Preference (defined in this document): defines a 2-octet value that indicates the PE preference to become the DF in the ES. 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. The DF Preference field is specific to DF Alg 2 and DF Alg TBD, and does not represent any Preference value for other Algs. If the DF Alg is different than Alg 2 or Alg TBD, these two octets can be encoded differently.

#### 4. Solution description

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



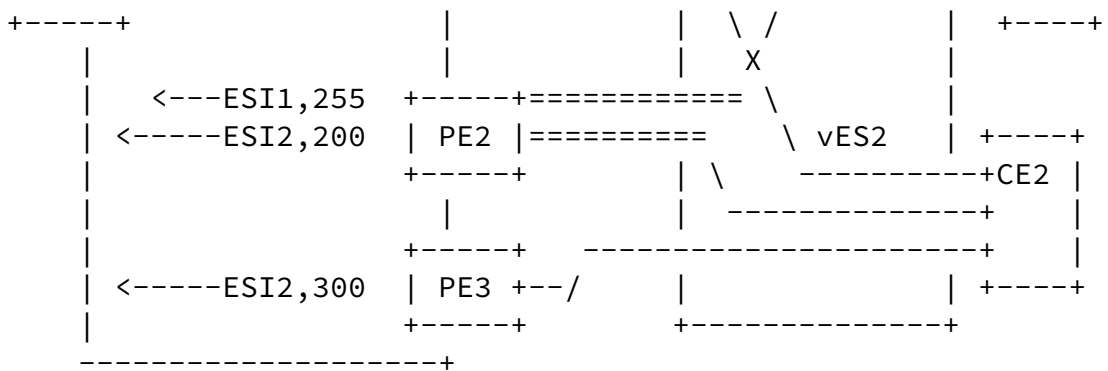


Figure 3: Preference-based DF Election

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 defined in PE1, PE2 and PE3.

If the algorithm chosen for vES1 and vES2 is Alg 2 or Alg TBD, i.e., Highest-Preference or Lowest-Preference, the PEs may become DF irrespective of their IP address and based on an 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 Algorithm

Assuming the operator wants to control - in a flexible way - what PE becomes the DF for a given vES and the order in which the PEs become DF 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 DF Election extended community. These parameters are the Preference, Preemption option (or "Don't Preempt Me" option) and DF Alg. We will represent these parameters as (Pref,DP,Alg). Let's assume vES1 is configured as (500,0,Highest-Pref) in PE1, and (255,0,Highest-Pref) in PE2. vES2 is configured as (100,0,Highest-Pref), (200,0,Highest-Pref) and (300,0,Highest-Pref) in PE1, PE2 and PE3 respectively.

- b. The PEs will advertise an ES route for each vES, including the 3



parameters in the DF Election Extended Community.

- c. According to [[RFC8584](#)], each PE will run the DF election algorithm upon expiration of the DF Wait timer. In this case, each PE runs the Highest-Preference DF Alg for each ES as follows:
  - The PE will check the DF Alg value in each ES route, and assuming all the ES routes are consistent in this DF Alg and the value is 2 (Highest-Preference), the PE will run the procedure in this section. Otherwise, the procedure will fall back to [[RFC7432](#)] Default Alg.
  - In this Highest-Preference Alg, each PE builds a list of candidate PEs, ordered by Preference. E.g. PE1 will build a list of candidate PEs for vES1 ordered by the Preference, from high to low: PE1>PE2. Hence PE1 will become the DF for vES1. In the same way, PE3 becomes the DF for vES2.
- d. Assuming some maintenance tasks had to be executed on, E.g., PE3, the operator could set vES2's Preference to E.g., 50 so that PE2 is forced to take over as DF for vES2 (irrespective of the DP capability). Once the maintenance task on PE3 is over, the operator could decide to leave the existing preference or configure the old preference back.
- e. In case of equal Preference in two or more PEs in the ES, the DP bit and the lowest IP of the candidate PEs are used as tie-breakers. After selecting the PEs with the highest Preference value, 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 ES route. Some examples of the use of the DP bit and IP address tie-breakers follow:
  - If vES1 parameters were (500,0,Highest-Pref) in PE1 and (500,1,Highest-Pref) in PE2, PE2 would be elected due to the DP bit.
  - If vES1 parameters were (500,0,Highest-Pref) in PE1 and (500,0,Highest-Pref) in PE2, PE1 would be elected, assuming PE1's IP address is lower than PE2's.
- f. The Preference is an administrative option that MUST be configured on a per-ES basis from the management plane, but MAY

also be dynamically changed based on the use of local policies. For instance, on PE1, ES1's Preference 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). Policies 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 default Preference value is 32767.

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

The procedures in this document can be used in [\[RFC7432\]](#) based ES or vES 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 Lowest-Preference Algorithm

In addition to the Highest-Preference Alg described in [Section 4.1](#) this document defines the Lowest-Preference Alg. In this case, and using the example of vES1 in Figure 3, if the Lowest-Preference Alg is configured in all the PEs in the ES, PE2 will be the DF due to its lower Preference.

All the procedures described in [Section 4.1](#) apply to the Lowest-Preference Alg, only replacing the Highest-Preference tie-breaker with the Lowest-Preference tie-breaker. 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 ES, the operational DF Election Alg will fall back to the Default Alg.

#### [4.3.](#) Use of the Highest-Preference algorithm in [\[RFC7432\]](#) Ethernet Segments

While the Highest-Preference (or Lowest-Preference for that matter) DF Alg described in [Section 4.1](#) is typically used in virtual ES scenarios where there is normally an individual Ethernet Tag per vES, the existing [\[RFC7432\]](#) definition of an ES allows potentially up to thousands of Ethernet Tags on the same ES. If this is the case, if Highest-Preference (or Lowest-Preference) Alg is configured in all

the PEs of the ES, the same PE will be the elected DF for all the

Ethernet Tags of the ES. A potential way to achieve a more granular load balancing is described below.

The ES is configured with an administrative Preference value and E.g., Highest-Preference Alg, but then a range of Ethernet Tags can be defined to use the Lowest-Preference depending on the desired behavior. With this option, the PE will build a list of candidate PEs ordered by Preference, however the DF for a given Ethernet Tag will be determined by the local configuration.

For instance:

- o 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).
- o 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 will be configured with (Ethernet Tag-range,low), E.g., (2001-4000, low).
- o This will result in PE1 being DF for Ethernet Tags 1-2000 (since they use the default Highest-Preference Alg) and PE2 being DF 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 DF function among the PEs is valid, as long as that logic is consistent in all the PEs in the ES. It is important to note that, when a local policy overrides the Highest-Preference or Lowest-Preference signaled by all the PEs in the ES, this local policy MUST be consistent in all the PEs of the ES. If the local policy is inconsistent for a given Ethernet Tag in the ES, black-holes or packet duplication may occur on that Ethernet Tag.

#### [4.4.](#) The Non-Revertive Capability

As discussed in [Section 1.2](#) (d), a capability to NOT preempt the

existing DF for a given Ethernet Tag is required and therefore added to the DF Election extended community. This option will allow a non-revertive behavior in the DF election.

Note that, when a given PE in an ES 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 DF 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 DF for ESI2.

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

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:

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

The procedure is described assuming Highest-Preference Alg in the ES, 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 Me" capability is defined on a per-PE/per-ES basis, as described in [Section 3](#). If "Don't Preempt Me" is disabled (default behavior), the advertised DP bit will be 0. If "Don't Preempt Me" is enabled, the ES route will be advertised with DP=1 ("Don't Preempt Me"). All the PEs in an ES 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 ES, 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 ES, the three PEs are configured with the "Don't Preempt Me" 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 ES routes and select PE3 as DF for Ethernet Tag-1 (due to the Highest-Preference), and PE1 as DF 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 DF 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 ES routes from PE1 and PE2. This timer is applied between the INIT and the DF\_WAIT states in the DF Election Finite State Machine described in [\[RFC8584\]](#). PE3 will then:
  - Select two "reference-PEs" among the ES routes in the vES, 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 ES, Lowest and Highest PE would be the same PE.
- Check its own administrative Pref and compares it with the one of the Highest-PE and Lowest-PE that have DP=1 in their ES routes. Depending on this comparison PE3 will send the ES route with a (Pref,DP) that may be different from its administrative (Pref,DP):
  - \* If PE3's Pref value is higher than the Highest-PE's, PE3 will send the ES route with an 'in-use' operational Pref equal to the Highest-PE's and DP=0.
  - \* If PE3's Pref value is lower than the Lowest-PE's, PE3 will send the ES route with an 'in-use' operational Preference equal to the Lowest-PE's and DP=0.

- \* If PE3's Pref value is neither higher nor lower than the Highest-PE's or the Lowest-PE's respectively, PE3 will send the ES 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 ES route with an operational 'in-use' (Pref,DP)=(200,0).
- Note that, a PE will always send DP=0 as long as the advertised Pref is the 'in-use' operational Pref (as opposed to the 'administrative' Pref).
- This ES route update sent by PE3, with (200,0,PE3-IP), will not cause any DF switchover for any Ethernet Tag. PE2 will continue being DF for Ethernet Tag-1. This is because the DP bit will be used as a tie-breaker in the DF 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 DF changes for Ethernet Tag-2 either.

6. For any subsequent update/withdraw in the ES, the PEs will go through the process described in (5) to select Highest and Lowest-PEs. For instance, if PE2 fails, upon receiving PE2's ES route withdrawal, PE3 and PE1 will go through the selection of new Highest and Lowest-PEs (considering their own active ES route) and then they will run the DF 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 ES route update with its administrative Pref and DP values. In the example, PE3 will be the new Highest-PE, therefore it will send an ES route update with (Pref,DP)=(300,1).
  - After running the DF Election, PE3 will become the new DF for Ethernet Tag-1. No changes will occur for Ethernet Tag-2.

If the ES 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 ES 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 ES comes back and the PE advertises a DF Election Alg different than the one configured

in the rest of the PEs in the ES, all the PEs in the ES 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 ES after running the DF Election, irrespective of the revertive or non-revertive behavior in the PE.

## 5. Security Considerations

This document describes a DF Election Algorithm that provides

absolute control (by configuration) over what PE is the DF 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 ES may change the behavior of the network. In other DF Algs such as HRW, the DF Election is more automated and cannot be determined by configuration.

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

The document also describes how a local policy can override the Highest-Preference Alg for a range of Ethernet Tags in the ES. If the local policy is not consistent across all PEs in the ES 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 the following values:

- o DF Alg = 2 in the [[RFC8584](#)] "DF Alg" registry, with name "Highest-Preference Algorithm".
- o DF Alg = TBD in the same "DF Alg" registry, with name "Lowest-Preference Algorithm".
- o Bit 0 in the [[RFC8584](#)] DF Election Capabilities registry, with name "D (Don't Preempt) Capability" for Non-revertive ES.

## 7. Acknowledgments

The authors would like to thank Kishore Tiruveedhula for his 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:

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", [draft-ietf-bess-evpn-virtual-eth-segment-07](#) (work in progress), July 2021.

## 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  
777 Middlefield Road  
Mountain View, CA 94043  
USA

Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

S. Sathappan  
Nokia

Email: [senthil.sathappan@nokia.com](mailto:senthil.sathappan@nokia.com)

T. Przygienda  
Juniper Networks

Email: [prz@juniper.net](mailto:prz@juniper.net)

W. Lin  
Juniper Networks

Email: [wlin@juniper.net](mailto:wlin@juniper.net)

Internet-Draft

Preference-based EVPN DF Election

September 2021

J. Drake  
Juniper Networks

Email: [jdrake@juniper.net](mailto:jdrake@juniper.net)

A. Sajassi  
Cisco Systems

Email: [sajassi@cisco.com](mailto:sajassi@cisco.com)

S. Mohanty  
Cisco Systems

Email: [satyamoh@cisco.com](mailto:satyamoh@cisco.com)

