Internet Engineering Task Force (IETF)
Request for Comments: 9785
Updates: 8584
Category: Standards Track
ISSN: 2070-1721

J. Rabadan, Ed.
S. Sathappan
Nokia
W. Lin
Juniper Networks
J. Drake
Independent
A. Sajassi
Cisco Systems

June 2025

Preference-Based EVPN Designated Forwarder (DF) Election

Abstract

The Designated Forwarder (DF) in Ethernet Virtual Private Networks (EVPNs) is defined as the Provider Edge (PE) router responsible for sending Broadcast, Unknown Unicast, and Multicast (BUM) traffic to a multihomed device/network in the case of an All-Active multihoming Ethernet Segment (ES) or BUM and unicast in the case of Single-Active multihoming. 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 DF 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, Network Operators 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 use of a DF election algorithm that meets the requirements of determinism and operation control. This document updates RFC 8584 by modifying the definition of the DF Election Extended Community.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9785.

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, Unknown Unicast, and Multicast (BUM) traffic to a multihomed device/network in the case of an All-Active multihoming Ethernet Segment or BUM and unicast traffic to a multihomed device or network in the case of Single-Active multihoming. 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 DF election algorithm or to DF Alg as per [RFC8584].

While the default DF algorithm or the Highest Random Weight (HRW) algorithm [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 user-controlled method is required. At the same time, Network Operators 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.

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 the Designated Forwarder, assuming they are all operationally ready to take over as the 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 the Designated Forwarder, based on the DF algorithms described in this document.

b. The extensions described in this document work for Ethernet Segments [RFC7432] and virtual Ethernet Segments as defined in [RFC9784].

c. The user may force a PE to preempt the existing Designated Forwarder for a given Ethernet Tag without reconfiguring 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 (the "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 DF election procedures [RFC7432] that are always revertive (because the winner PE of the default DF election algorithm always takes over as the operational Designated Forwarder).

e. The procedures described in this document support Single-Active and All-Active multihoming Ethernet Segments.

1.3. Solution Overview

To provide a solution that satisfies the above requirements, we introduce two new DF algorithms that can be advertised in the DF Election Extended Community (Section 3). Carried with the new DF Election Extended Community variants is a DF election preference advertised for each PE that influences which PE will become the DF (Section 4.1). The advertised DF election preference can dynamically vary from the administratively configured preference to provide non-revertive behavior (Section 4.3). In Section 4.2, an optional solution is discussed for use in Ethernet Segments that supports large numbers of Ethernet Tags and therefore needs to balance load among multiple DFs.

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.

CE: Customer Equipment router

DF: Designated Forwarder

DF Alg: Refers to the DF election algorithm, which is sometimes shortened to "Alg" in this document.

DP: Refers to the "Don’t Preempt" Capability in the DF Election Extended Community.

ENNI: External Network-Network Interface

ES and vES: Ethernet Segment and virtual Ethernet Segment.

Ethernet A-D per EVI route: Refers to Route Type 1 or Auto-Discovery per EVPN Instance route [RFC7432].

EVC: Ethernet Virtual Circuit

EVI: EVPN Instance

Ethernet Tag: Used to represent a broadcast domain that is configured on a given Ethernet Segment for the purpose of DF election. Note that any of the following may be used to represent a broadcast domain: VLAN IDs (VIDs) (including Q-in-Q tags), configured IDs, VXLAN Network Identifiers (VNIs), normalized VIDs, Service Instance Identifiers (I-SIDs), etc., as long as the representation of the broadcast domains is configured consistently across the multihomed PEs attached to that Ethernet Segment. The Ethernet Tag value MUST NOT be zero.

HRW: Highest Random Weight, as per [RFC8584].

OAM: Operations, Administration, and Maintenance.

3. EVPN BGP Attribute Extensions

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

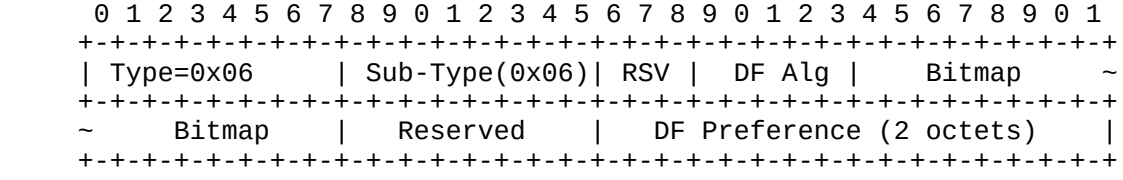

Figure 1: DF Election Extended Community

Where the above fields are defined as follows:

* The DF algorithm can have the following values:

  - Alg 0 - Default DF election algorithm, i.e., the modulus-based algorithm as per [RFC7432].

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

  - Alg 2 - Highest-Preference Algorithm (Section 4.1).

  - Alg 3 - Lowest-Preference Algorithm (Section 4.1).

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

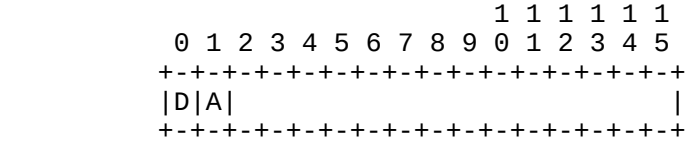

Figure 2: Bitmap Field in the DF Election Extended Community

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

'preempt' or 'revertive' behavior in the default DF algorithm [RFC7432]. The "Don't Preempt" Capability is supported by the Highest-Preference or Lowest-Preference Algorithms. The procedures of the "Don't Preempt" Capability for other DF algorithms are out of the scope of this document. The procedures of the "Don't Preempt" Capability for the Highest-Preference and Lowest-Preference Algorithms are described in Section 4.1.

- Bit 1: AC-Influenced DF (AC-DF) election is described in [RFC8584]. When set to 1, it indicates the desire to use AC-DF with the rest of the PEs in the Ethernet Segment. The AC-DF capability bit MAY be set along with the "Don't Preempt" Capability and Highest-Preference or Lowest-Preference Algorithms.

* Designated Forwarder (DF) Preference: 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 DF election depending on the DF algorithm being used, as explained in Section 4.1. The Designated Forwarder Preference field is specific to Highest-Preference and Lowest-Preference Algorithms, and this document does not define any meaning for other algorithms. If the DF algorithm is different from Highest-Preference or Lowest-Preference, these 2 octets can be encoded differently.

* RSV and Reserved fields (from bit 16 to bit 18, and from bit 40 to 47): When the DF algorithm is set to Highest-Preference or Lowest-Preference, 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.

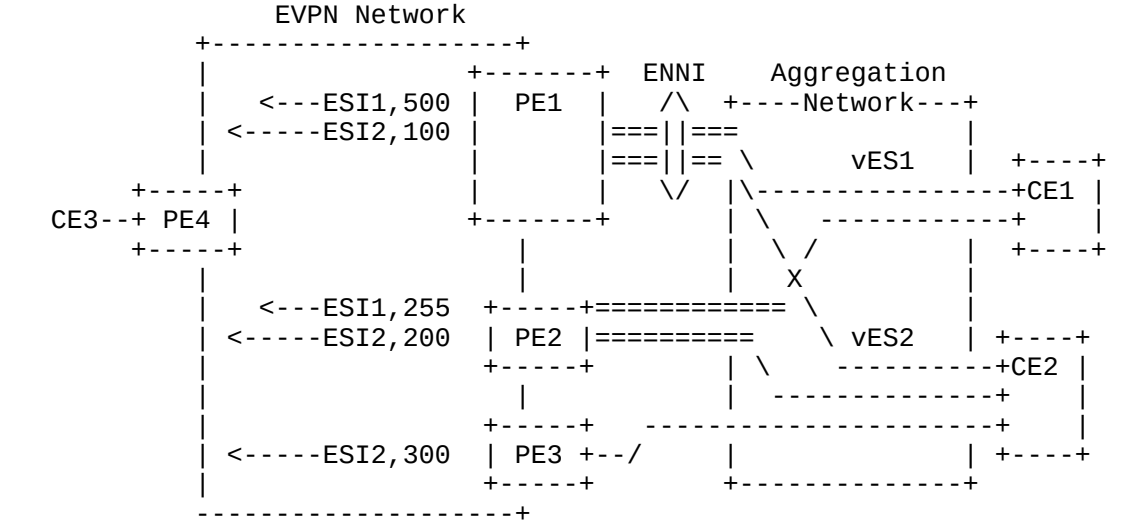

Figure 3: Preference-Based DF Election

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, which spans PE1 and PE2, and CE2 is connected to vES2, which is attached to PE1, PE2, and PE3.

If the algorithm chosen for vES1 and vES2 is the Highest-Preference or Lowest-Preference Algorithm, the PEs may become the 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 a Designated Forwarder in case of multiple failures, the Highest-Preference or Lowest-Preference Algorithms can be used. Per the example in Figure 3, these algorithms are used as follows:

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 (or "Don't Preempt" Capability) option, and DF algorithm. We will represent these parameters as (Pref, DP, Alg). For instance, vES1 (Pref, DP, Alg) is configured as:

   (500, 0, Highest-Preference) in PE1,
   (255, 0, Highest-Preference) in PE2.

   vES2 is configured as:

   (100, 0, Highest-Preference) in PE1,
   (200, 0, Highest-Preference) in PE2, and
   (300, 0, Highest-Preference) in PE3.

b. The PEs advertise an Ethernet Segment route for each virtual Ethernet Segment, including the three parameters indicated in (a) above, in the DF Election Extended Community (encoded as described in Section 3).

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

   * The PE will check the DF algorithm value in each Ethernet Segment route, and assuming all the Ethernet Segment routes (including the local route) are consistent in this DF algorithm (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 the default DF algorithm [RFC7432]. The Highest-Preference and Lowest-Preference Algorithms are different algorithms; therefore, if two PEs configured for Highest-Preference and Lowest-Preference, respectively, are attached to the same Ethernet Segment, the operational DF election algorithm will fall back to the default DF algorithm.

   * If all the PEs attached to the Ethernet Segment advertise the Highest-Preference Algorithm, each PE builds a list of candidate PEs, ordered by preference value from the numerically highest value to lowest value. For example, 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 the Lowest-Preference Algorithm, then the candidate list is ordered from the numerically lowest preference value to the highest preference value. For example, 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. For example, if PE3 is going on maintenance and the DF algorithm 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 "Don't Preempt" 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 the Lowest-Preference Algorithm too. For example, suppose the algorithm for vES2 is Lowest-Preference, and PE1 (the DF) goes on maintenance mode. The operator could change vES2's Preference on PE1 from 100 to, e.g., 250, so that PE2 is forced to take over as the Designated Forwarder for vES2.

e. In case of equal Preference in two or more PEs in the Ethernet Segment, the "Don't Preempt" Capability and the numerically lowest IP address of the candidate PE(s) are used as tiebreakers. The procedures for the use of the "Don't Preempt" Capability are specified in Section 4.3. If more than one PE is advertising itself as the preferred Designated Forwarder, an implementation MUST first select the PE advertising the "Don't Preempt" Capability set, and then select the PE with the lowest IP address (if the "Don't Preempt" Capability 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 the Originating Router's IP address of different families, an IPv4 address is always considered numerically lower than an IPv6 address. Some examples of using the "Don't Preempt" Capability and IP address tiebreakers 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 "Don't Preempt" Capability. The same example applies if PE1 and PE2 advertise the Lowest-Preference Algorithm 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 address is lower than PE1's. The same example applies if PE1 and PE2 advertise the Lowest-Preference Algorithm 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.

* On PE1, if the DF algorithm is Highest-Preference, ES1's preference value can be lowered from 500 to 100 in case the bandwidth on the ENNI port is decreased by 50% (that could happen if, e.g., the 2-port Link Aggregation Group 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.

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

Highest-Preference and Lowest-Preference Algorithms 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, and PE2 and PE3 can use Lowest-Preference for vES2. The use of one DF algorithm over the other is the operator's choice. The existence of both provides flexibility and full control to the operator.

The procedures in this document can be used in an Ethernet Segment as defined in [RFC7432] or a virtual Ethernet Segment per [RFC9784] and also in EVPN networks as described in [RFC8214], [RFC7623], or [RFC8365].

4.2. Use of the Highest-Preference or Lowest-Preference Algorithm in Ethernet Segments

While the Highest-Preference or Lowest-Preference Algorithm 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 definition of an Ethernet Segment [RFC7432] allows potentially up to thousands of Ethernet Tags on the same Ethernet Segment. If this is the case, and if the Highest-Preference or Lowest-Preference Algorithm 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 achieve a more granular load balancing is described below.

The Ethernet Segment is configured with an administrative preference value and an administrative DF algorithm, i.e., Highest-Preference or Lowest-Preference Algorithm. However, the administrative DF algorithm (which is used to signal the DF algorithm for the Ethernet Segment) MAY be overridden to a different operational DF algorithm 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 locally overridden DF algorithm.

For instance:

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

* In addition, assuming there are 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 the Designated Forwarder for Ethernet Tags 1-2000 (since they use the default Highest-Preference Algorithm) and PE2 being the Designated Forwarder for Ethernet Tags 2001-4000, due to the local policy overriding the Highest-Preference Algorithm.

While the above logic provides a perfect load-balancing distribution of Ethernet Tags per Designated Forwarder when there are only two PEs, for Ethernet Segments attached to three or more PEs, there would be only two Designated Forwarder PEs for all the Ethernet Tags. Any other logic that provides a fair distribution of the Designated Forwarder function among the three or more 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, packet drops or packet duplication may occur on that Ethernet Tag. For all these reasons, the use of virtual Ethernet Segments is RECOMMENDED for cases where more than two PEs per Ethernet Segment exist and a good load-balancing distribution per Ethernet Tag of the Designated Forwarder function is desired.

4.3. The Non-Revertive Capability

As discussed in item d of Section 1.2, 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 DF Election Extended Community. This option allows a non-revertive behavior in the DF 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 "Don’t Preempt" Capability 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 Algorithm, PE3 is the Designated Forwarder for ESI2.

If PE3 has a link, EVC, or node failure, PE2 would take over as the 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 (see Figure 3). The procedure supports a non-revertive mode that can be used along with:

* Highest-Preference Algorithm

* Lowest-Preference Algorithm

* Highest-Preference or Lowest-Preference Algorithm, where a local policy overrides the Highest-/Lowest-Preference tiebreaker for a range of Ethernet Tags (Section 4.2)

The procedure is described, assuming the Highest-Preference Algorithm

in the Ethernet Segment, where local policy overrides the tiebreaker for a given Ethernet Tag. The other cases above are a subset of this one, and the differences are explained.

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" is enabled, the Ethernet Segment route from the PE indicates 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 "Don’t Preempt" Capability; however, this document does not enforce the consistency across all the PEs. In case of inconsistency in the support of the "Don’t Preempt" Capability in the PEs of the same Ethernet Segment, non-revertive behavior is not guaranteed. However, PEs supporting this capability still attempt this procedure.

2. Assuming 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 the DF algorithm, 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 the Designated Forwarder for Ethernet Tag-1 (due to the Highest-Preference) and PE1 as the Designated Forwarder for Ethernet Tag-2 (due to the Lowest-Preference).

4. If PE3’s vES2 goes down (due to an EVC failure (as detected by OAM protocols), a port failure, or a 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 DF election Finite State Machine described in [RFC8584]. PE3 will then:

   * Select a "reference-PE" among the Ethernet Segment routes in the virtual Ethernet Segment. If the Ethernet Segment uses the Highest-Preference Algorithm, select a "Highest-PE". If it uses the Lowest-Preference Algorithm, select a "Lowest-PE". If a local policy is in use, to override the Highest-/Lowest-Preference for a range of Ethernet Tags (as discussed in Section 4.2), it is necessary to select both a Highest-PE and a Lowest-PE. They are selected as follows:

     - The Highest-PE is the PE with higher Preference, using the "Don’t Preempt" Capability first (with DP=1 being better) and, after that, the lower PE-IP address as tiebreakers.

     - The Lowest-PE is the PE with lower Preference, using the "Don’t Preempt" Capability first (with DP=1 being better) and, after that, the lower PE-IP address as tiebreakers.

     - In our example, the Highest-Preference Algorithm is used, with a local policy to override it to use Lowest-Preference for a range of Ethernet Tags. Therefore, PE3 selects a Highest-PE and a Lowest-PE. PE3 will select PE2 as the

Highest-PE over PE1, because when comparing (Pref, DP, PE-IP), (200, 1, PE2-IP) wins over (100, 1, PE1-IP). PE3 will select PE1 as the Lowest-PE over PE2, because (100, 1, PE1-IP) wins over (200, 1, PE2-IP). Note that if there were only one remote PE in the Ethernet Segment, the 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 "Don't Preempt" Capability set in their Ethernet Segment routes. Depending on this comparison, PE3 sends the Ethernet Segment route with a (Pref, DP) that may be different from its administrative (Pref, DP):

  - If PE3's Pref value is higher than or equal to 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 than or equal to 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 than or equal to the Highest-PE's and is not lower than or equal to 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).

* Send its "Don't Preempt" Capability set to zero, as long as the advertised Pref is the 'in-use' operational Pref (as opposed to the 'administrative' Pref).

* Not trigger any Designated Forwarder changes for Ethernet Tag-1. This Ethernet Segment route update sent by PE3, with (200, 0, PE3-IP), will not cause any Designated Forwarder switchover for any Ethernet Tag. This is because the "Don't Preempt" Capability will be used as a tiebreaker in the DF election. That is, if a PE has two candidate PEs with the same Pref, it will pick 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 the Highest-PEs 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 the new Highest-PEs and Lowest-PEs (considering their own active Ethernet Segment route), and then they will run the DF election.

   * If a PE selects itself as the new Highest-PE 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 DF election, PE3 will become the new Designated Forwarder for Ethernet Tag-1. No changes will

occur for Ethernet Tag-2.

Note that, irrespective of the "Don’t Preempt" Capability, when a PE or Ethernet Segment comes back and the PE advertises a DF election algorithm different from 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 DF algorithm [RFC7432].

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 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 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 algorithms such as HRW, the DF election is more automated and cannot be determined by configuration. If the DF algorithm is Highest-Preference or Lowest-Preference, an attacker may change the configuration of the preference value on a PE and Ethernet Segment to 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 DF 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. However, an attacker who gets access to the configuration of a PE in the Ethernet Segment will be able to disable the non-revertive behavior, by advertising a conflicting DF election algorithm and thereby forcing fallback to the default DF algorithm.

The document also describes how a local policy can override the Highest-Preference or Lowest-Preference Algorithms 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, packet drop or packet duplication may occur for that Ethernet Tag.

Finally, the two DF election algorithms specified in this document (Highest-Preference and Lowest-Preference) do not change the way the PEs share their Ethernet Segment information, compared to the algorithms in [RFC7432] and [RFC8584]. Therefore, the security considerations in [RFC7432] and [RFC8584] apply to this document as well.

6. IANA Considerations

Per this document, IANA has:

* Allocated two new values in the "DF Alg" registry created by [RFC8584], as follows:

Alg	Name	Reference
2	Highest-Preference Algorithm	RFC 9785
3	Lowest-Preference Algorithm	RFC 9785

Table 1

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

Bit	Name	Reference
0	D (Don’t Preempt) Capability	RFC 9785

Table 2

* Listed this document as an additional reference for the DF Election Extended Community field in the "EVPN Extended Community Sub-Types" registry, as follows:

Sub-Type Value	Name	Reference
0x06	DF Election Extended Community	[RFC8584] and RFC 9785

Table 3

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[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>.

[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>.

[RFC9784] Sajassi, A., Brissette, P., Schell, R., Drake, J., and J. Rabadan, "Virtual Ethernet Segments for EVPN and Provider Backbone Bridge EVPN", RFC 9784, DOI 10.17487/9784, June 2025, <https://www.rfc-editor.org/info/rfc9784>.

7.2. Informative References

[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>.

[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>.

Acknowledgements

The authors would like to thank Kishore Tiruveedhula and Sasha Vainshtein for their reviews and comments. Also, thank you to Luc Andr Burdet and Stephane Litkowski for their thorough reviews and suggestions for a new Lowest-Preference Algorithm.

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

Authors’ Addresses

Jorge Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia
Email: senthil.sathappan@nokia.com

Wen Lin
Juniper Networks
Email: wlin@juniper.net

John Drake

Independent
Email: je_drake@yahoo.com

Ali Sajassi
Cisco Systems
Email: sajassi@cisco.com