BESS Workgroup
Internet Draft
Updates: 7432
Intended status: Standards Track

J. Rabadan, Ed.
Nokia
S. Mohanty, Ed.
A. Sajassi
Cisco
J. Drake
Juniper
K. Nagaraj
S. Sathappan
Nokia

Expires: July 28, 2019 January 24, 2019

Framework for EVPN Designated Forwarder Election Extensibility
draft-ietf-bess-evpn-df-election-framework-09

Abstract

An alternative to the Default Designated Forwarder (DF) selection algorithm in Ethernet VPN (EVPN) networks is defined. The DF is the Provider Edge (PE) router responsible for sending broadcast, unknown unicast and multicast (BUM) traffic to multi-homed Customer Equipment (CE) on a particular Ethernet Segment (ES) within a VLAN. In addition, the capability to influence the DF election result for a VLAN based on the state of the associated Attachment Circuit (AC) is specified. This document clarifies the DF Election Finite State Machine in EVPN, therefore it updates the EVPN specification.

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html

This Internet-Draft will expire on July 28, 2019.

Copyright Notice

Copyright (c) 2019 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 (http://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

1. Introduction 3
1.1. Default Designated Forwarder (DF) Election in EVPN 3
1.2. Problem Statement 6
1.2.1. Unfair Load-Balancing and Service Disruption 6
1.2.2. Traffic Black-Holing on Individual AC Failures 7
1.3. The Need for Extending the Default DF Election in EVPN . . 10
2. Conventions and Terminology 11
3. Designated Forwarder Election Protocol and BGP Extensions . . . 12
3.1. The DF Election Finite State Machine (FSM) 12
3.2. The DF Election Extended Community 15
3.2.1. Backward Compatibility 18
3.3. Auto-Derivation of ES-Import Route Target 18
4. The Highest Random Weight DF Election Algorithm 18
4.1. HRW and Consistent Hashing 19
4.2. HRW Algorithm for EVPN DF Election 19
5. The Attachment Circuit Influenced DF Election Capability . . . 21
5.1. AC-Influenced DF Election Capability For VLAN-Aware Bundle Services 23

6. Solution Benefits 24
7. Security Considerations 25
8. IANA Considerations 25
9. References 26
9.1. Normative References 26
9.2. Informative References 27
10. Acknowledgments 27
11. Contributors 28
Authors' Addresses 28

1. Introduction

The Designated Forwarder (DF) in EVPN networks is the Provider Edge (PE) router responsible for sending broadcast, unknown unicast and multicast (BUM) traffic to a multi-homed Customer Equipment (CE) device, on a given VLAN on a particular Ethernet Segment (ES). The DF is selected out of a list of candidate PEs that advertise the same Ethernet Segment Identifier (ESI) to the EVPN network. By default, EVPN uses a DF Election algorithm referred to as "Service Carving" and it is based on a modulus function (V mod N) that takes the number of PEs in the ES (N) and the VLAN value (V) as input. This Default DF Election algorithm has some inefficiencies that this document addresses by defining a new DF Election algorithm and a capability to influence the DF Election result for a VLAN, depending on the state of the associated Attachment Circuit (AC). In order to avoid any ambiguity with the identifier used in the DF Election Algorithm, this document uses the term Ethernet Tag instead of VLAN. This document also creates a registry with IANA, for future DF Election Algorithms and Capabilities. It also presents a formal definition and clarification of the DF Election Finite State Machine (FSM), therefore the document updates [RFC7432] and EVPN implementations MUST conform to the prescribed FSM.

The procedures described in this document apply to DF election in all EVPN solutions including [RFC7432] and [RFC8214]. Apart from the FSM formal description, this document does not intend to update other [RFC7432] procedures. It only aims to improve the behavior of the DF Election on PEs that are upgraded to follow the described procedures.

1.1. Default Designated Forwarder (DF) Election in EVPN

[RFC7432] defines the Designated Forwarder (DF) as the EVPN PE

responsible for:

- Flooding Broadcast, Unknown unicast and Multicast traffic (BUM), on a given Ethernet Tag on a particular Ethernet Segment (ES), to the CE. This is valid for single-active and all-active EVPN multi-homing.

- Sending unicast traffic on a given Ethernet Tag on a particular ES to the CE. This is valid for single-active multi-homing.

Figure 1 illustrates an example that we will use to explain the Designated Forwarder function.

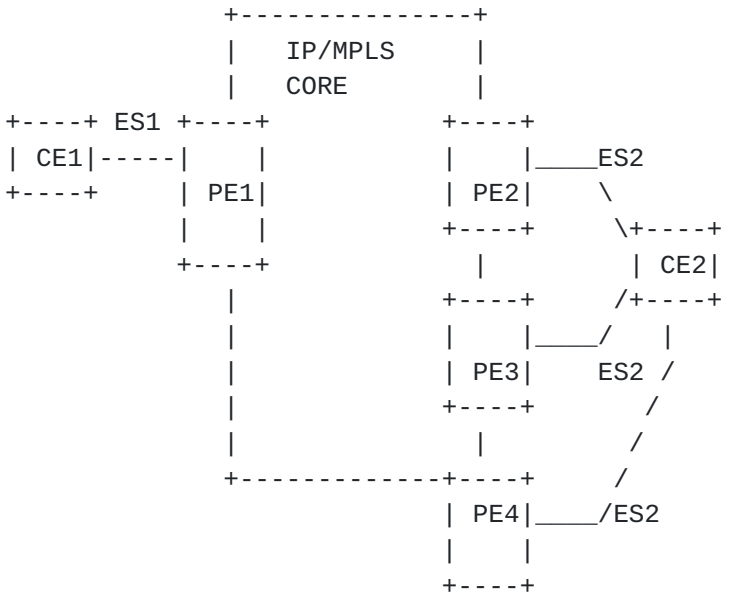


Figure 1 Multi-homing Network of EVPN

Figure 1 illustrates a case where there are two Ethernet Segments, ES1 and ES2. PE1 is attached to CE1 via Ethernet Segment ES1 whereas PE2, PE3 and PE4 are attached to CE2 via ES2 i.e. PE2, PE3 and PE4 form a redundancy group. Since CE2 is multi-homed to different PEs on the same Ethernet Segment, it is necessary for PE2, PE3 and PE4 to agree on a DF to satisfy the above mentioned requirements.

The effect of forwarding loops in a Layer-2 network is particularly severe because of the broadcast nature of Ethernet traffic and the lack of a Time-To-Live (TTL). Therefore it is very important that in the case of a multi-homed CE only one of the PEs be used to send BUM traffic to it.

One of the pre-requisites for this support is that participating PEs

must agree amongst themselves as to who would act as the Designated Forwarder (DF). This needs to be achieved through a distributed algorithm in which each participating PE independently and unambiguously selects one of the participating PEs as the DF, and the result should be consistent and unanimous.

The default algorithm for DF election defined by [RFC7432] at the granularity of (ESI,EVI) is referred to as "service carving". In this document, service carving and Default DF Election algorithm are used interchangeably. With service carving, it is possible to elect multiple DFs per Ethernet Segment (one per EVI) in order to perform load-balancing of traffic destined to a given Segment. The objective is that the load-balancing procedures should carve up the BD space among the redundant PE nodes evenly, in such a way that every PE is the DF for a distinct set of EVIs.

The DF Election algorithm as described in [RFC7432] (Section 8.5) is based on a modulus operation. The PEs to which the ES (for which DF election is to be carried out per EVI) is multi-homed form an ordered (ordinal) list in ascending order of the PE IP address values. For example, there are N PEs: PE0, PE1,... PEN-1 ranked as per increasing IP addresses in the ordinal list; then for each VLAN with Ethernet Tag V, configured on the Ethernet Segment ES1, PEx is the DF for VLAN V on ES1 when x equals (V mod N). In the case of VLAN Bundle only the lowest VLAN is used. In the case when the planned density is high (meaning there are significant number of VLANs and the Ethernet Tags are uniformly distributed), the thinking is that the DF Election will be spread across the PEs hosting that Ethernet Segment and good load-balancing can be achieved.

However, the described Default DF Election algorithm has some undesirable properties and in some cases can be somewhat disruptive and unfair. This document describes some of those issues and defines a mechanism for dealing with them. These mechanisms do involve changes to the Default DF Election algorithm, but they do not require any changes to the EVPN Route exchange and have minimal changes in the EVPN routes.

In addition, there is a need to extend the DF Election procedures so that new algorithms and capabilities are possible. A single algorithm (the Default DF Election algorithm) may not meet the requirements in all the use-cases.

Note that while [RFC7432] elects a DF per <ES, EVI>, this document elects a DF per <ES, BD>. This means that unlike [RFC7432], where for a VLAN-Aware Bundle service EVI there is only one DF for the EVI, this document specifies that there will be multiple DFs, one for each BD configured in that EVI.

1.2. Problem Statement

This section describes some potential issues with the Default DF Election algorithm.

1.2.1. Unfair Load-Balancing and Service Disruption

There are three fundamental problems with the current Default DF Election algorithm.

1- First, the algorithm will not perform well when the Ethernet Tag follows a non-uniform distribution, for instance when the Ethernet Tags are all even or all odd. In such a case let us assume that the ES is multi-homed to two PEs; one of the PEs will be elected as DF for all of the VLANs. This is very sub-optimal. It defeats the purpose of service carving as the DFs are not really evenly spread across. In fact, in this particular case, one of the PEs does not get elected as DF at all, so it does not participate in the DF responsibilities at all. Consider another example where, referring to Figure 1, lets assume that PE2, PE3, PE4 are in ascending order of the IP address; and each VLAN configured on ES2 is associated with an Ethernet Tag of the form (3x+1), where x is an integer. This will result in PE3 always be selected as the DF.

2- The Ethernet tag that identifies the BD can be as large as 2^24; however, it is not guaranteed that the tenant BD on the ES will conform to a uniform distribution. In fact, it is up to the customer what BDs they will configure on the ES. Quoting [Knuth], "In general, we want to avoid values of M that divide r^k+a or r^k-a, where k and a are small numbers and r is the radix of the alphabetic character set (usually r=64, 256 or 100), since a remainder modulo such a value of M tends to be largely a simple superposition of key digits. Such considerations suggest that we choose M to be a prime number such that r^k!=a(modulo)M or r^k!=?a(modulo)M for small k & a."

   In our case, N is the number of PEs in [RFC7432] which corresponds to M above. Since N, N-1 or N+1 need not satisfy the primality properties of the M above; as per the [RFC7432] modulo based DF assignment, whenever a PE goes down or a new PE boots up (hosting the same Ethernet Segment), the modulo scheme will not necessarily map BDs to PEs uniformly.

3- The third problem is one of disruption. Consider a case when the same Ethernet Segment is multi-homed to a set of PEs. When the ES is down in one of the PEs, say PE1, or PE1 itself reboots, or the BGP process goes down or the connectivity between PE1 and an RR goes down, the effective number of PEs in the system now becomes

N-1, and DFs are computed for all the VLANs that are configured on that Ethernet Segment. In general, if the DF for a VLAN v happens not to be PE1, but some other PE, say PE2, it is likely that some other PE (different from PE1 and PE2) will become the new DF. This is not desirable. Similarly when a new PE hosts the same Ethernet Segment, the mapping again changes because of the modulus operation. This results in needless churn. Again referring to Figure 1, say v1, v2 and v3 are VLANs configured on ES2 with associated Ethernet Tags of value 999, 1000 and 1001 respectively. So PE1, PE2 and PE3 are the DFs for v1, v2 and v3 respectively. Now when PE3 goes down, PE2 will become the DF for v1 and PE1 will become the DF for v2.

One point to note is that the Default DF election algorithm assumes that all the PEs who are multi-homed to the same Ethernet Segment (and interested in the DF Election by exchanging EVPN routes) use an Originating Router's IP Address of the same family. This does not need to be the case as the EVPN address-family can be carried over an IPv4 or IPv6 peering, and the PEs attached to the same ES may use an address of either family.

Mathematically, a conventional hash function maps a key k to a number i representing one of m hash buckets through a function h(k) i.e. i=h(k). In the EVPN case, h is simply a modulo-m hash function viz. h(v) = v mod N, where N is the number of PEs that are multi-homed to the Ethernet Segment in discussion. It is well-known that for good hash distribution using the modulus operation, the modulus N should be a prime-number not too close to a power of 2 [CLRS2009]. When the effective number of PEs changes from N to N-1 (or vice versa); all the objects (VLAN V) will be remapped except those for which V mod N and V mod (N-1) refer to the same PE in the previous and subsequent ordinal rankings respectively. From a forwarding perspective, this is a churn, as it results in re-programming the PE ports as either blocking or non-blocking at the PEs where the DF state changes.

This document addresses this problem and furnishes a solution to this undesirable behavior.

1.2.2. Traffic Black-Holing on Individual AC Failures

As discussed in section 2.1 the Default DF Election algorithm defined by [RFC7432] takes into account only two variables in the modulus function for a given ES: the existence of the PE's IP address on the candidate list and the locally provisioned Ethernet Tags.

If the DF for an <ESI, EVI> fails (due to physical link/node failures) an ES route withdrawal will make the Non-DF (NDF) PEs re-

elect the DF for that <ESI, EVI> and the service will be recovered.

However, the Default DF election procedure does not provide a protection against "logical" failures or human errors that may occur at service level on the DF, while the list of active PEs for a given ES does not change. These failures may have an impact not only on the local PE where the issue happens, but also on the rest of the PEs of the ES. Some examples of such logical failures are listed below:

a) A given individual Attachment Circuit (AC) defined in an ES is accidentally shutdown or even not provisioned yet (hence the Attachment Circuit Status - ACS - is DOWN), while the ES is operationally active (since the ES route is active).

b) A given MAC-VRF - with a defined ES - is shutdown or not provisioned yet, while the ES is operationally active (since the ES route is active). In this case, the ACS of all the ACs defined in that MAC-VRF is considered to be DOWN.

Neither (a) nor (b) will trigger the DF re-election on the remote multi-homed PEs for a given ES since the ACS is not taken into account in the DF election procedures. While the ACS is used as a DF election tie-breaker and trigger in VPLS multi-homing procedures [VPLS-MH], there is no procedure defined in EVPN [RFC7432] to trigger the DF re-election based on the ACS change on the DF.

Figure 2 illustrates the described issue with an example.

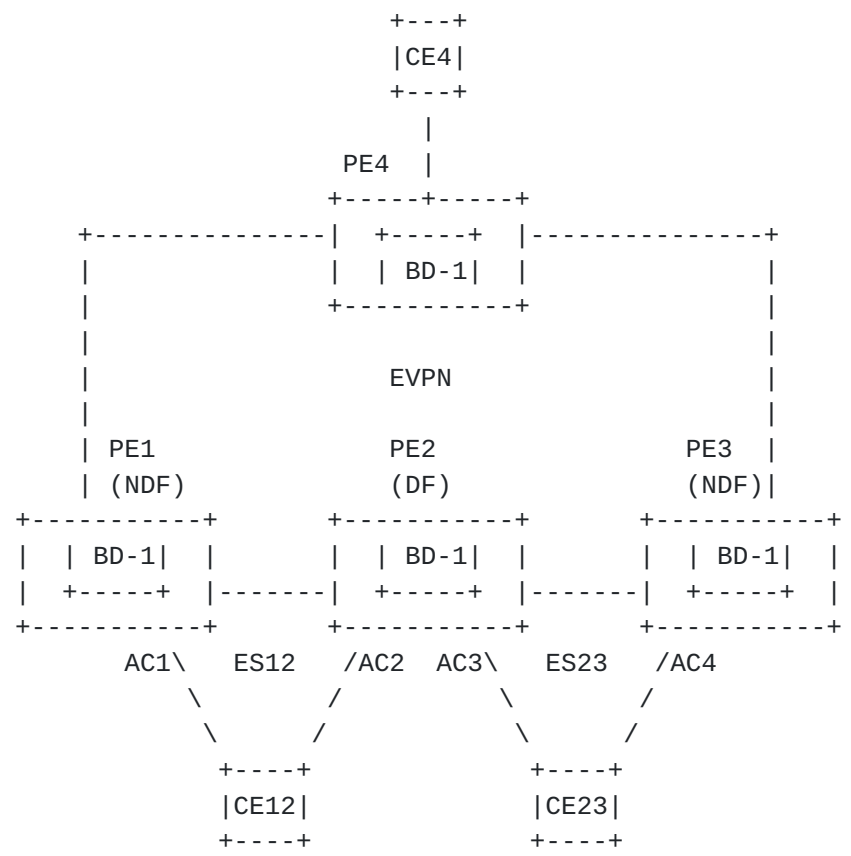


Figure 2 Default DF Election and Traffic Black-Holing

BD-1 is defined in PE1, PE2, PE3 and PE4. CE12 is a multi-homed CE connected to ES12 in PE1 and PE2. Similarly CE23 is multi-homed to PE2 and PE3 using ES23. Both, CE12 and CE23, are connected to BD-1 through VLAN-based service interfaces: CE12-VID 1 (VLAN ID 1 on CE12) is associated to AC1 and AC2 in BD-1, whereas CE23-VID 1 is associated to AC3 and AC4 in BD-1. Assume that, although not represented, there are other ACs defined on these ES mapped to different BDs.

After executing the [RFC7432] Default DF election algorithm, PE2 turns out to be the DF for ES12 and ES23 in BD-1. The following issues may arise:

a) If AC2 is accidentally shutdown or even not configured, CE12 traffic will be impacted. In case of all-active multi-homing, the BUM traffic to CE12 will be "black-holed", whereas for single-active multi-homing, all the traffic to/from CE12 will be discarded. This is due to the fact that a logical failure in PE2's AC2 may not trigger an ES route withdrawn for ES12 (since there are still other ACs active on ES12) and therefore PE1 will not re-

run the DF election procedures.

b) If the Bridge Table for BD-1 is administratively shutdown or even not configured yet on PE2, CE12 and CE23 will both be impacted: BUM traffic to both CEs will be discarded in case of all-active multi-homing and all traffic will be discarded to/from the CEs in case of single-active multi-homing. This is due to the fact that PE1 and PE3 will not re-run the DF election procedures and will keep assuming PE2 is the DF.

Quoting [RFC7432], "when an Ethernet Tag is decommissioned on an Ethernet Segment, then the PE MUST withdraw the Ethernet A-D per EVI route(s) announced for the <ESI, Ethernet Tags> that are impacted by the decommissioning", however, while this A-D per EVI route withdrawal is used at the remote PEs performing aliasing or backup procedures, it is not used to influence the DF election for the affected EVIs.

This document adds an optional modification of the DF Election procedure so that the ACS may be taken into account as a variable in the DF election, and therefore EVPN can provide protection against logical failures.

1.3. The Need for Extending the Default DF Election in EVPN

Section 1.2 describes some of the issues that exist in the Default DF Election procedures. In order to address those issues, this document introduces a new DF Election framework. This framework allows the PEs to agree on a common DF election algorithm, as well as the capabilities to enable during the DF Election procedure. Generally, 'DF election algorithm' refers to the algorithm by which a number of input parameters are used to determine the DF PE, while 'DF election capability' refers to an additional feature that can be used prior to the invocation of the DF election algorithm, such as modifying the inputs (or list of candidate PEs).

Within this framework, this document defines a new DF Election algorithm and a new capability that can influence the DF Election result:

- The new DF Election algorithm is referred to as "Highest Random Weight" (HRW). The HRW procedures are described in section 4.

- The new DF Election capability is referred to as "AC-Influenced DF Election" (AC-DF). The AC-DF procedures are described in section 5.

- HRW and AC-DF mechanisms are independent of each other. Therefore,

a PE may support either HRW or AC-DF independently or may support both of them together. A PE may also support AC-DF capability along with the Default DF election algorithm per [RFC7432].

In addition, this document defines a way to indicate the support of HRW and/or AC-DF along with the EVPN ES routes advertised for a given ES. Refer to section 3.2 for more details.

2. Conventions 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 and ACS - Attachment Circuit and Attachment Circuit Status. 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
- 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.
- ES and ESI - Ethernet Segment and Ethernet Segment Identifier.
- EVI - EVPN Instance.
- 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.
- Bridge Table - An instantiation of a broadcast domain on a MAC-VRF.
- HRW - Highest Random Weight
- VID and CE-VID - VLAN Identifier and Customer Equipment VLAN Identifier.

- 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.

- Ethernet Tag ID - refers to the identifier used in the EVPN routes defined in [RFC7432]. Its value may be the same as the Ethernet Tag value (see Ethernet Tag definition) when advertising routes for VLAN-aware Bundle services. Note that in case of VLAN-based or VLAN Bundle services, the Ethernet Tag ID is zero.

- DF Election Procedure and DF Algorithm - The Designated Forwarder Election Procedure or simply DF Election, refers to the process in its entirety, including the discovery of the PEs in the ES, the creation and maintenance of the PE candidate list and the selection of a PE. The Designated Forwarder Algorithm is just a component of the DF Election Procedure and strictly refers to the selection of a PE for a given <ES,Ethernet Tag>.

- TTL - Time To Live

This document also assumes familiarity with the terminology of [RFC7432].

3. Designated Forwarder Election Protocol and BGP Extensions

This section describes the BGP extensions required to support the new DF Election procedures. In addition, since the EVPN specification [RFC7432] does leave several questions open as to the precise final state machine behavior of the DF election, section 3.1 describes precisely the intended behavior.

3.1. The DF Election Finite State Machine (FSM)

Per [RFC7432], the FSM described in Figure 3 is executed per <ESI,VLAN> in case of VLAN-based service or <ESI,[VLANs in VLAN Bundle]> in case of VLAN Bundle on each participating PE.

Observe that currently the VLANs are derived from local configuration and the FSM does not provide any protection against misconfiguration where the same (EVI,ESI) combination has different set of VLANs on

different participating PEs or one of the PEs elects to consider VLANs as VLAN Bundle and another as separate VLANs for election purposes (service type mismatch).

The FSM is conceptual and any design or implementation MUST comply with a behavior equivalent to the one outlined in this FSM.

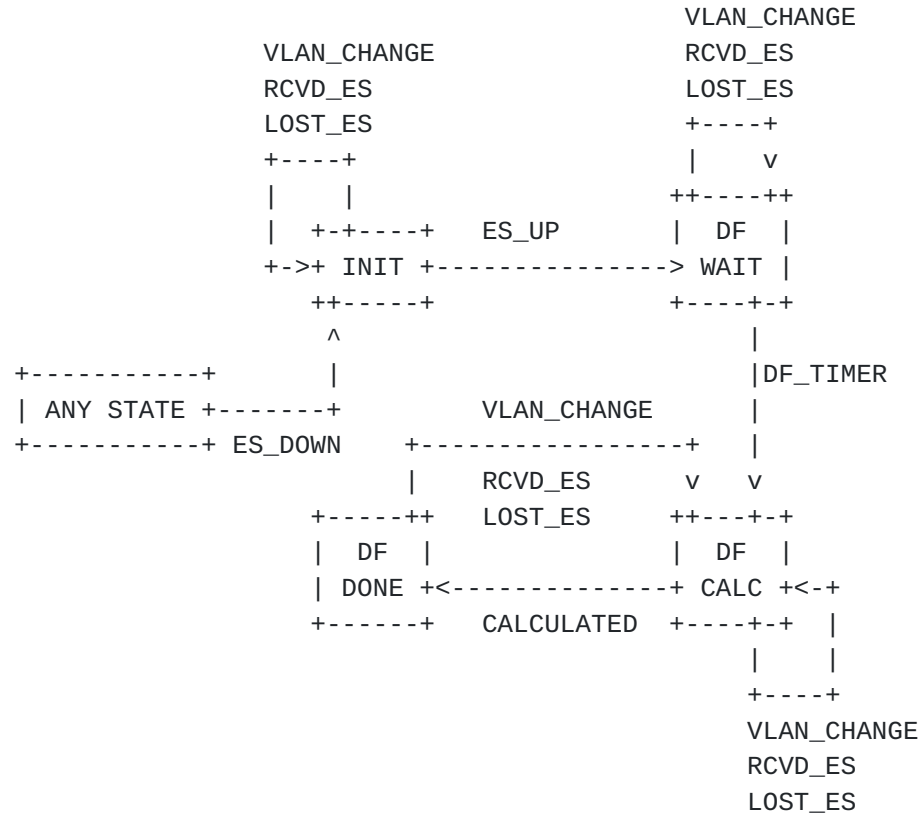


Figure 3 DF Election Finite State Machine

States:

1. INIT: Initial State

2. DF_WAIT: State in which the participant waits for enough information to perform the DF election for the EVI/ESI/VLAN combination.

3. DF_CALC: State in which the new DF is recomputed.

4. DF_DONE: State in which the according DF for the EVI/ESI/VLAN combination has been elected.

5. ANY_STATE: Refers to any of the above states.

Events:

1. ES_UP: The ESI has been locally configured as 'up'.

2. ES_DOWN: The ESI has been locally configured as 'down'.

3. VLAN_CHANGE: The VLANs configured in a bundle (that uses the ESI) changed. This event is necessary for VLAN Bundles only.

4. DF_TIMER: DF Wait timer [RFC7432] has expired.

5. RCVD_ES: A new or changed Ethernet Segment route is received in a BGP REACH UPDATE. Receiving an unchanged UPDATE MUST NOT trigger this event.

6. LOST_ES: A BGP UNREACH UPDATE for a previously received Ethernet Segment route has been received. If an UNREACH is seen for a route that has not been advertised previously, the event MUST NOT be triggered.

7. CALCULATED: DF has been successfully calculated.

According actions when transitions are performed or states entered/exited:

1. ANY_STATE on ES_DOWN: (i) stop DF wait timer (ii) assume NDF for local PE.

2. INIT on ES_UP: transition to DF_WAIT.

3. INIT on VLAN_CHANGE, RCVD_ES or LOST_ES: do nothing.

4. DF_WAIT on entering the state: (i) start DF wait timer if not started already or expired (ii) assume NDF for local PE.

5. DF_WAIT on VLAN_CHANGE, RCVD_ES or LOST_ES: do nothing.

6. DF_WAIT on DF_TIMER: transition to DF_CALC.

7. DF_CALC on entering or re-entering the state: (i) rebuild candidate list, hash and perform election (ii) Afterwards FSM generates CALCULATED event against itself.

8. DF_CALC on VLAN_CHANGE, RCVD_ES or LOST_ES: do as in transition 7.

9. DF_CALC on CALCULATED: mark election result for VLAN or bundle,

and transition to DF_DONE.

11. DF_DONE on exiting the state: if there is a new DF election triggered and the current DF is lost, then assume NDF for local PE for VLAN or VLAN Bundle.

12. DF_DONE on VLAN_CHANGE, RCVD_ES or LOST_ES: transition to DF_CALC.

The above events and transitions are defined for the Default DF Election Algorithm. As described in Section 5, the use of the AC-DF capability introduces additional events and transitions.

3.2. The DF Election Extended Community

For the DF election procedures to be consistent and unanimous, it is necessary that all the participating PEs agree on the DF Election algorithm and capabilities to be used. For instance, it is not possible that some PEs continue to use the Default DF Election algorithm and some PEs use HRW. For brown-field deployments and for interoperability with legacy PEs, it is important that all PEs need to have the capability to fall back on the Default DF Election. A PE can indicate its willingness to support HRW and/or AC-DF by signaling a DF Election Extended Community along with the Ethernet Segment route (Type-4).

The DF Election Extended Community is a new BGP transitive extended community attribute [RFC4360] that is defined to identify the DF election procedure to be used for the Ethernet Segment. Figure 4 shows the encoding of the DF Election Extended Community.

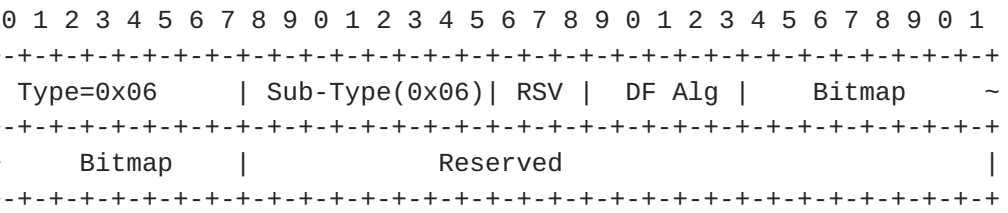


Figure 4 DF Election Extended Community

Where:

o Type is 0x06 as registered with IANA for EVPN Extended Communities.

o Sub-Type is 0x06 - "DF Election Extended Community" as requested by this document to IANA.

- RSV / Reserved - Reserved bits for DF Alg specific information.

- DF Alg (5 bits) - Encodes the DF Election algorithm values (between 0 and 31) that the advertising PE desires to use for the ES. This document requests IANA to set up a registry called "DF Alg Registry" and solicits the following values:

  - Type 0: Default DF Election algorithm, or modulus-based algorithm as in [RFC7432].

  - Type 1: HRW algorithm (explained in this document).

  - Types 2-30: Unassigned.

  - Type 31: Reserved for Experimental Use.

- Bitmap (2 octets) - Encodes "capabilities" to use with the DF Election algorithm in the field "DF Alg". This document requests IANA to create a registry for the Bitmap field, with values 0-15, called "DF Election Capabilities" and solicits the following values:

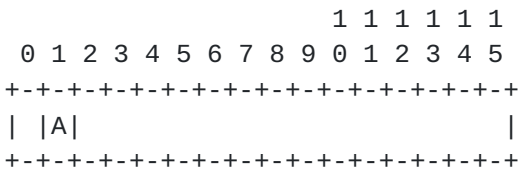


Figure 5 Bitmap field in the DF Election Extended Community

  - Bit 0 (corresponds to Bit 24 of the DF Election Extended Community): Unassigned.

  - Bit 1: AC-DF (AC-Influenced DF Election, explained in this document). When set to 1, it indicates the desire to use AC-Influenced DF Election with the rest of the PEs in the ES.

  - Bits 2-15: Unassigned.

The DF Election Extended Community is used as follows:

- A PE SHOULD attach the DF Election Extended Community to any advertised ES route and the Extended Community MUST be sent if the

ES is locally configured with a DF election algorithm other than the Default Election algorithm or if a capability is required to be used. In the Extended Community, the PE indicates the desired "DF Alg" algorithm and "Bitmap" capabilities to be used for the ES.

- Only one DF Election Extended Community can be sent along with an ES route. Note that the intent is not for the advertising PE to indicate all the supported DF election algorithms and capabilities, but signal the preferred one.

- DF Algs 0 and 1 can be both used with bit AC-DF set to 0 or 1.

- In general, a specific DF Alg SHOULD determine the use of the reserved bits in the Extended Community, which may be used in a different way for a different DF Alg. In particular, for DF Algs 0 and 1, the reserved bits are not set by the advertising PE and SHOULD be ignored by the receiving PE.

o When a PE receives the ES Routes from all the other PEs for the ES in question, it checks to see if all the advertisements have the extended community with the same DF Alg and Bitmap:

- In the case that they do, this particular PE MUST follow the procedures for the advertised DF Alg and capabilities. For instance, if all ES routes for a given ES indicate DF Alg HRW and AC-DF set to 1, the receiving PE and by induction all the other PEs in the ES will proceed to do DF Election as per the HRW Algorithm and following the AC-DF procedures.

- Otherwise if even a single advertisement for the type-4 route is received without the locally configured DF Alg and capability, the Default DF Election algorithm (modulus) algorithm MUST be used as in [RFC7432]. This procedure handles the case where participating PEs in the ES disagree about the DF algorithm and capability to apply.

- The absence of the DF Election Extended Community or the presence of multiple DF Election Extended Communities (in the same route) MUST be interpreted by a receiving PE as an indication of the Default DF Election algorithm on the sending PE, that is, DF Alg 0 and no DF Election capabilities.

o When all the PEs in an ES advertise DF Type 31, they will rely on the local policy to decide how to proceed with the DF Election.

o For any new capability defined in the future, the applicability/compatibility of this new capability to the existing DF Algs must be assessed on a case by case basis.

o Likewise, for any new DF Alg defined in future, its applicability/compatibility to the existing capabilities must be assessed on a case by case basis.

3.2.1. Backward Compatibility

[RFC7432] implementations (i.e., those that predate this specification) will not advertise the DF Election Extended Community. That means that all other participating PEs in the ES will not receive DF preferences and will revert to the Default DF Election algorithm without AC-Influenced DF Election.

Similarly, a [RFC7432] implementation receiving a DF Election Extended Community will ignore it and will continue to use the Default DF Election algorithm.

3.3. Auto-Derivation of ES-Import Route Target

Section 7.6 of [RFC7432] describes how the value of the ES-Import Route Target for ESI types 1, 2, and 3 can be auto-derived by using the high-order six bytes of the nine byte ESI value. The same auto-derivation procedure can be extended to ESI types 0, 4, and 5 as long as it is ensured that the auto-derived values for ES-Import RT among different ES types don't overlap. As in [RFC7432], the mechanism to guarantee that the auto-derived ESI or ES-import RT values for different ESIs do not match is out of scope of this document.

4. The Highest Random Weight DF Election Algorithm

The procedure discussed in this section is applicable to the DF Election in EVPN Services [RFC7432] and EVPN Virtual Private Wire Services [RFC8214].

Highest Random Weight (HRW) as defined in [HRW1999] is originally proposed in the context of Internet Caching and proxy Server load balancing. Given an object name and a set of servers, HRW maps a request to a server using the object-name (object-id) and server-name (server-id) rather than the server states. HRW forms a hash out of the server-id and the object-id and forms an ordered list of the servers for the particular object-id. The server for which the hash value is highest, serves as the primary responsible for that particular object, and the server with the next highest value in that hash serves as the backup server. HRW always maps a given object name to the same server within a given cluster; consequently it can be used at client sites to achieve global consensus on object-server mappings. When that server goes down, the backup server becomes the

responsible designate.

Choosing an appropriate hash function that is statistically oblivious to the key distribution and imparts a good uniform distribution of the hash output is an important aspect of the algorithm. Fortunately many such hash functions exist. [HRW1999] provides pseudo-random functions based on the Unix utilities rand and srand and easily constructed XOR functions that satisfy the desired hashing properties. HRW already finds use in multicast and ECMP [RFC2991],[RFC2992].

4.1. HRW and Consistent Hashing

HRW is not the only algorithm that addresses the object to server mapping problem with goals of fair load distribution, redundancy and fast access. There is another family of algorithms that also addresses this problem; these fall under the umbrella of the Consistent Hashing Algorithms [CHASH]. These will not be considered here.

4.2. HRW Algorithm for EVPN DF Election

This section describes the application of HRW to DF election. Let DF(v) denote the Designated Forwarder and BDF(v) the Backup Designated forwarder for the Ethernet Tag v, where v is the VLAN, Si is the IP address of PE i, Es denotes the Ethernet Segment Identifier and weight is a function of v, Si, and Es.

Note that while the DF election algorithm in [RFC7432] uses PE address and vlan as inputs, this document uses Ethernet Tag, PE address and ESI as inputs. This is because if the same set of PEs are multi-homed to the same set of ESes, then the DF election algorithm used in [RFC7432] would result in the same PE being elected DF for the same set of broadcast domains on each ES, which can have adverse side-effects on both load balancing and redundancy. Including ESI in the DF election algorithm introduces additional entropy which significantly reduces the probability of the same PE being elected DF for the same set of broadcast domains on each ES. Therefore, when using the HRW Algorithm for EVPN DF Election, the ESI value in the Weight function below SHOULD be set to that of the corresponding ES.

In case of a VLAN Bundle service, v denotes the lowest VLAN similar to the 'lowest VLAN in bundle' logic of [RFC7432].

1. DF(v) = Si| Weight(v, Es, Si) >= Weight(v, Es, Sj), for all j. In case of a tie, choose the PE whose IP address is numerically the least. Note 0 <= i,j < Number of PEs in the redundancy group.

2. BDF(v) = Sk| Weight(v, Es, Si) >= Weight(v, Es, Sk) and Weight(v, Es, Sk) >= Weight(v, Es, Sj). In case of tie choose the PE whose IP address is numerically the least.

Where:

DF(v): is defined to be the address Si (index i) for which weight(v, Es, Si) is the highest, 0 <= i < N-1

BDF(v) is defined as that PE with address Sk for which the computed weight is the next highest after the weight of the DF. j is the running index from 0 to N-1, i, k are selected values.

Since the Weight is a pseudo-random function with domain as the three-tuple (v, Es, S), it is an efficient and deterministic algorithm that is independent of the Ethernet Tag v sample space distribution. Choosing a good hash function for the pseudo-random function is an important consideration for this algorithm to perform better than the Default algorithm. As mentioned previously, such functions are described in the HRW paper. We take as candidate hash function the first one out of the two that are preferred in [HRW1999]:

Wrand(v, Es, Si) = (1103515245((1103515245.Si+12345) XOR D(v,Es))+12345)(mod 2^31)

Here D(v,Es) is the 31-bit digest (CRC-32 and discarding the MSB as in [HRW1999]) of the 14-byte stream, the Ethernet Tag v (4 bytes) followed by the Ethernet Segment Identifier (10 bytes). It is mandated that the 14-byte stream is formed by concatenation of the Ethernet tag and the Ethernet Segment identifier in network byte order. The CRC should proceed as if the stream is in network byte order (big-endian). Si is address of the ith server. The server's IP address length does not matter as only the low-order 31 bits are modulo significant.

A point to note is that the Weight function takes into consideration the combination of the Ethernet Tag, Ethernet Segment and the PE IP-address, and the actual length of the server IP address (whether IPv4 or IPv6) is not really relevant. The Default algorithm in [RFC7432] cannot employ both IPv4 and IPv6 PE addresses, since [RFC7432] does not specify how to decide on the ordering (the ordinal list) when both IPv4 and IPv6 PEs are present.

HRW solves the disadvantages pointed out in Section 1.2.1 and ensures:

o with very high probability that the task of DF election for the

VLANs configured on an ES is more or less equally distributed among the PEs even for the 2 PE case.

- o If a PE that is not the DF or the BDF for that VLAN, goes down or its connection to the ES goes down, it does not result in a DF or BDF reassignment. This saves computation, especially in the case when the connection flaps.

- o More importantly it avoids the needless disruption case of Section 1.2.1 (3), that is inherent in the existing Default DF Election.

- o In addition to the DF, the algorithm also furnishes the BDF, which would be the DF if the current DF fails.

5. The Attachment Circuit Influenced DF Election Capability

The procedure discussed in this section is applicable to the DF Election in EVPN Services [RFC7432] and EVPN Virtual Private Wire Services [RFC8214].

The AC-DF capability is expected to be of general applicability with any future DF Algorithm. It modifies the DF Election procedures by removing from consideration any candidate PE in the ES that cannot forward traffic on the AC that belongs to the BD. This section is applicable to VLAN-Based and VLAN Bundle service interfaces. Section 5.1 describes the procedures for VLAN-Aware Bundle interfaces.

In particular, when used with the Default DF Alg, the AC-DF capability modifies the Step 3 in the DF Election procedure described in [RFC7432] Section 8.5, as follows:

3. When the timer expires, each PE builds an ordered "candidate" list of the IP addresses of all the PE nodes attached to the Ethernet Segment (including itself), in increasing numeric value. The candidate list is based on the Originator Router's IP addresses of the ES routes, but excludes any PE from whom no Ethernet A-D per ES route has been received, or from whom the route has been withdrawn. Afterwards, the DF Election algorithm is applied on a per <ES, Ethernet Tag>, however, the IP address for a PE will not be considered candidate for a given <ES, Ethernet Tag> until the corresponding Ethernet A-D per EVI route has been received from that PE. In other words, the ACS on the ES for a given PE must be UP so that the PE is considered as candidate for a given BD.

   If the Default DF Alg is used, every PE in the resulting candidate list is then given an ordinal indicating its position in the ordered list, starting with 0 as the ordinal for the PE with the

numerically lowest IP address. The ordinals are used to determine which PE node will be the DF for a given Ethernet Tag on the Ethernet Segment, using the following rule:

Assuming a redundancy group of N PE nodes, for VLAN-based service, the PE with ordinal i is the DF for an <ES, Ethernet Tag V> when (V mod N)= i. In the case of VLAN-(aware) bundle service, then the numerically lowest VLAN value in that bundle on that ES MUST be used in the modulo function as Ethernet Tag.

It should be noted that using the "Originating Router's IP address" field in the Ethernet Segment route to get the PE IP address needed for the ordered list allows for a CE to be multihomed across different ASes if such a need ever arises.

The above three paragraphs differ from [RFC7432] Section 8.5, Step 3, in two aspects:

- o Any DF Alg algorithm can be used, and not only the described modulus-based DF Alg (referred to as the Default DF Election, or DF Alg 0 in this document).

- o The candidate list is pruned based upon non-receipt of Ethernet A-D routes: a PE's IP address MUST be removed from the ES candidate list if its Ethernet A-D per ES route is withdrawn. A PE's IP address MUST NOT be considered as candidate DF for a <ES, Ethernet Tag>, if its Ethernet A-D per EVI route for the <ES, Ethernet Tag> is withdrawn.

The following example illustrates the AC-DF behavior applied to the Default DF election algorithm, assuming the network in Figure 2:

a) When PE1 and PE2 discover ES12, they advertise an ES route for ES12 with the associated ES-import extended community and the DF Election Extended Community indicating AC-DF=1; they start a DF Wait timer (independently). Likewise, PE2 and PE3 advertise an ES route for ES23 with AC-DF=1 and start a DF Wait timer.

b) PE1/PE2 advertise an Ethernet A-D per ES route for ES12, and PE2/PE3 advertise an Ethernet A-D per ES route for ES23.

c) In addition, PE1/PE2/PE3 advertise an Ethernet A-D per EVI route for AC1, AC2, AC3 and AC4 as soon as the ACs are enabled. Note that the AC can be associated to a single customer VID (e.g. VLAN-based service interfaces) or a bundle of customer VIDs (e.g. VLAN Bundle service interfaces).

d) When the timer expires, each PE builds an ordered "candidate" list

of the IP addresses of all the PE nodes connected to the Ethernet Segment (including itself) as explained above in [RFC7432] Step 3. Any PE from which an Ethernet A-D per ES route has not been received is pruned from the list.

e) When electing the DF for a given BD, a PE will not be considered candidate until an Ethernet A-D per EVI route has been received from that PE. In other words, the ACS on the ES for a given PE must be UP so that the PE is considered as candidate for a given BD. For example, PE1 will not consider PE2 as candidate for DF election for <ES12,VLAN-1> until an Ethernet A-D per EVI route is received from PE2 for <ES12,VLAN-1>.

f) Once the PEs with ACS = DOWN for a given BD have been removed from the candidate list, the DF Election can be applied for the remaining N candidates.

Note that this procedure only modifies the existing EVPN control plane by adding and processing the DF Election Extended Community, and by pruning the candidate list of PEs that take part in the DF election.

In addition to the events defined in the FSM in Section 3.1, the following events SHALL modify the candidate PE list and trigger the DF re-election in a PE for a given <ES, Ethernet Tag>. In the FSM of Figure 3, the events below MUST trigger a transition from DF_DONE to DF_CALC:

i. Local AC going DOWN/UP.

ii. Reception of a new Ethernet A-D per EVI update/withdraw for the <ES, Ethernet Tag>.

iii. Reception of a new Ethernet A-D per ES update/withdraw for the ES.

5.1. AC-Influenced DF Election Capability For VLAN-Aware Bundle Services

The procedure described in section 5 works for VLAN-based and VLAN Bundle service interfaces since, for those service types, a PE advertises only one Ethernet A-D per EVI route per <ES,VLAN> or <ES,VLAN Bundle>. In Section 5, an Ethernet Tag represents a given VLAN or VLAN Bundle for the purpose of DF Election. The withdrawal of such route means that the PE cannot forward traffic on that particular <ES,VLAN> or <ES,VLAN Bundle>, therefore the PE can be removed from consideration for DF.

According to [RFC7432], in VLAN-aware Bundle services, the PE advertises multiple Ethernet A-D per EVI routes per <ES,VLAN Bundle> (one route per Ethernet Tag), while the DF Election is still performed per <ES,VLAN Bundle>. The withdrawal of an individual route only indicates the unavailability of a specific AC but not necessarily all the ACs in the <ES,VLAN Bundle>.

This document modifies the DF Election for VLAN-Aware Bundle services in the following way:

- After confirming that all the PEs in the ES advertise the AC-DF capability, a PE will perform a DF Election per <ES,VLAN>, as opposed to per <ES,VLAN Bundle> in [RFC7432]. Now, the withdrawal of an Ethernet A-D per EVI route for a VLAN will indicate that the advertising PE's ACS is DOWN and the rest of the PEs in the ES can remove the PE from consideration for DF in the <ES,VLAN>.

- The PEs will now follow the procedures in section 5.

For example, assuming three Bridge Tables in PE1 for the same MAC-VRF (each one associated to a different Ethernet Tag, e.g. VLAN-1, VLAN-2 and VLAN-3), PE1 will advertise three Ethernet A-D per EVI routes for ES12. Each of the three routes will indicate the status of each of the three ACs in ES12. PE1 will be considered as a valid candidate PE for DF Election in <ES12,VLAN-1>, <ES12,VLAN-2>, <ES12,VLAN-3> as long as its three routes are active. For instance, if PE1 withdraws the Ethernet A-D per EVI routes for <ES12,VLAN-1>, the PEs in ES12 will not consider PE1 as a suitable DF candidate for <ES12,VLAN-1>. PE1 will still be considered for <ES12,VLAN-2> and <ES12,VLAN-3> since its routes are active.

6. Solution Benefits

The solution described in this document provides the following benefits:

a) Extends the DF Election in [RFC7432] to address the unfair load-balancing and potential black-holing issues of the Default DF Election algorithm. The solution is applicable to the DF Election in EVPN Services [RFC7432] and EVPN Virtual Private Wire Services [RFC8214].

b) It defines a way to signal the DF Election algorithm and capabilities intended by the advertising PE. This is done by defining the DF Election Extended Community, which allow signaling of the capabilities supported by this document as well as any other future DF Election algorithms and capabilities.

c) The solution is backwards compatible with the procedures defined in [RFC7432]. If one or more PEs in the ES do not support the new procedures, they will all follow the [RFC7432] DF Election.

7. Security Considerations

This document addresses some identified issues in the DF Election procedures described in [RFC7432] by defining a new DF Election framework. In general, this framework allows the PEs that are part of the same Ethernet Segment to exchange additional information and agree on the DF Election Type and Capabilities to be used.

Following the procedures in this document, the operator will minimize undesired situations such as unfair load-balancing, service disruption and traffic black-holing. Since those situations may have been purposely created by a malicious user with access to the configuration of one PE, this document enhances also the security of the network. Note that the network will not benefit of the new procedures if the DF Election Alg is not consistently configured on all the PEs in the ES (if there is no unanimity among all the PEs, the DF Election Alg falls back to the Default [RFC7432] DF Election). This behavior could be exploited by an attacker that manages to modify the configuration of one PE in the Ethernet Segment so that the DF Election Alg and capabilities in all the PEs in the Ethernet Segment fall back to the Default DF Election. If that is the case, the PEs will be exposed to the unfair load-balancing, service disruption and black-holing that were mentioned earlier.

In addition, the new framework is extensible and allows for future new security enhancements that are out of the scope of this document. Finally, since this document extends the procedures in [RFC7432], the same Security Considerations described in [RFC7432] are valid for this document.

8. IANA Considerations

IANA is requested to:

o Allocate Sub-Type value 0x06 in the "EVPN Extended Community Sub-Types" registry defined in [RFC7153] as follows:

SUB-TYPE VALUE	NAME	Reference
0x06	DF Election Extended Community	This document

o Set up a registry called "DF Alg" for the DF Alg field in the

Extended Community. New registrations will be made through the "RFC Required" procedure defined in [RFC8126]. Value 31 is for Experimental use and does not require any other RFC than this document. The following initial values in that registry are requested:

Alg	Name	Reference
0	Default DF Election	This document
1	HRW algorithm	This document
2-30	Unassigned	
31	Reserved for Experimental use	This document

o Set up a registry called "DF Election Capabilities" for the two-octet Bitmap field in the Extended Community. New registrations will be made through the "RFC Required" procedure defined in [RFC8126]. The following initial value in that registry is requested:

Bit	Name	Reference
0	Unassigned	
1	AC-DF capability	This document
2-15	Unassigned	

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>.

[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>.

[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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <http://www.rfc-editor.org/info/rfc4360>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <https://www.rfc-editor.org/info/rfc7153>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

9.2. Informative References

[VPLS-MH] Kothari, Henderickx et al., "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-bess-vpls-multihoming-02.txt, work in progress, September, 2018.

[CHASH] Karger, D., Lehman, E., Leighton, T., Panigrahy, R., Levine, M., and D. Lewin, "Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web", ACM Symposium on Theory of Computing ACM Press New York, May 1997.

[CLRS2009] Cormen, T., Leiserson, C., Rivest, R., and C. Stein, "Introduction to Algorithms (3rd ed.)", MIT Press and McGraw-Hill ISBN 0-262-03384-4., February 2009.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, DOI 10.17487/RFC2991, November 2000, <http://www.rfc-editor.org/info/rfc2991>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998, <https://www.microsoft.com/en-us/research/wp-content/uploads/2017/02/HRW98.pdf>.

[Knuth] Art of Computer Programming - Sorting and Searching,Vol 3 Pg. 516, Addison Wesley

10. Acknowledgments

The authors want to thank Sriram Venkateswaran, Laxmi Padakanti,

Ranganathan Boovaraghavan, Tamas Mondal, Sami Boutros, Jakob Heitz, Mrinmoy Ghosh, Leo Mermelstein, Mankamana Mishra, Anoop Ghanwani and Samir Thoria for their review and contributions. Special thanks to Stephane Litkowski for his thorough review and detailed contributions.

11. Contributors

In addition to the authors listed on the front page, the following coauthors have also contributed to this document:

Antoni Przygienda
Juniper Networks, Inc.
1194 N. Mathilda Drive
Sunnyvale, CA 95134
USA
Email: prz@juniper.net

Vinod Prabhu
Nokia
Email: vinod.prabhu@nokia.com

Wim Henderickx
Nokia
Email: wim.henderickx@nokia.com

Wen Lin
Juniper Networks, Inc.
Email: wlin@juniper.net

Patrice Brissette
Cisco Systems
Email: pbrisset@cisco.com

Keyur Patel
Arrcus, Inc
Email: keyur@arrcus.com

Autumn Liu
Ciena
Email: hliu@ciena.com

Authors' Addresses

Jorge Rabadan
Nokia

777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@nokia.com

Satya Mohanty
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA
Email: satyamoh@cisco.com

Ali Sajassi
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA
Email: sajassi@cisco.com

John Drake
Juniper Networks, Inc.
1194 N. Mathilda Drive
Sunnyvale, CA 95134
USA
Email: jdrake@juniper.net

Kiran Nagaraj
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: kiran.nagaraj@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043 USA
Email: senthil.sathappan@nokia.com