      Framework for Ethernet VPN Designated Forwarder Election Extensibility

Abstract

   An alternative to the default Designated Forwarder (DF) selection
   algorithm in Ethernet VPNs (EVPNs) is defined. The DF is the
   Provider Edge (PE) router responsible for sending Broadcast, Unknown
   Unicast, and Multicast (BUM) traffic to a multihomed Customer Edge
   (CE) device on a given VLAN on a particular Ethernet Segment (ES).
   In addition, the ability 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 services. Therefore, it updates the EVPN
   specification (RFC 7432).

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/rfc8584.

Copyright Notice

Table of Contents

1. Introduction

The Designated Forwarder (DF) in Ethernet VPNs (EVPNs) is the
Provider Edge (PE) router responsible for sending Broadcast, Unknown
Unicast, and Multicast (BUM) traffic to a multihomed Customer Edge
(CE) device on a given VLAN on a particular Ethernet Segment (ES).
The DF is elected from the set of multihomed PEs attached to a given
ES, each of which advertises an ES route for the ES as identified by
its Ethernet Segment Identifier (ESI). By default, the EVPN uses a
DF election algorithm referred to as "service carving". The DF
election algorithm 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 document addresses inefficiencies in the default DF
election algorithm by defining a new DF election algorithm and an
ability 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 (see Section 7). It also
presents a formal definition and clarification of the DF election
Finite State Machine (FSM). Therefore, this 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 those described in [RFC7432] and [RFC8214].
Apart from the formal description of the FSM, this document does not
intend to update other procedures described in [RFC7432]; it only
aims to improve the behavior of the DF election on PEs that are
upgraded to follow the procedures described in this document.

1.1. 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.

o AC: Attachment Circuit. An AC has an Ethernet Tag associated
   with it.

o ACS: Attachment Circuit Status.

o BUM: Broadcast, unknown unicast, and multicast.

o DF: Designated Forwarder.

o NDF: Non-Designated Forwarder.

o BDF: Backup Designated Forwarder.

o Ethernet A-D per ES route: Refers to Route Type 1 as defined in
   [RFC7432] or to Auto-discovery per Ethernet Segment route.

o Ethernet A-D per EVI route: Refers to Route Type 1 as defined in
   [RFC7432] or to Auto-discovery per EVPN Instance route.

o ES: Ethernet Segment.

o ESI: Ethernet Segment Identifier.

o EVI: EVPN Instance.

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 BD
   (VLAN-based or VLAN Bundle services) or multiple BDs (VLAN-aware
   Bundle services).

o Bridge table: An instantiation of a BD on a MAC-VRF.

o HRW: Highest Random Weight.

o VID: VLAN Identifier.

o CE-VID: Customer Edge VLAN Identifier.

o Ethernet Tag: Used to represent a BD 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 BD: VIDs (including Q-in-Q tags),
   configured IDs, VNIs (Virtual Extensible Local Area Network
   (VXLAN) Network Identifiers), normalized VIDs, I-SIDs (Service
   Instance Identifiers), etc., as long as the representation of the
   BDs is configured consistently across the multihomed PEs attached
   to that ES. The Ethernet Tag value MUST be different from zero.

o 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 the definition for Ethernet Tag) when advertising
   routes for VLAN-aware Bundle services. Note that in the case of
   VLAN-based or VLAN Bundle services, the Ethernet Tag ID is zero.

o DF election procedure: Also called "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.

o DF algorithm: A component of the DF election procedure. Strictly
   refers to the selection of a PE for a given <ES, Ethernet Tag>.

o RR: Route Reflector. A network routing component for BGP
   [RFC4456]. It offers an alternative to the logical full-mesh
   requirement of the Internal Border Gateway Protocol (IBGP). The
   purpose of the RR is concentration. Multiple BGP routers can peer
   with a central point, the RR -- acting as a route reflector server
   -- rather than peer with every other router in a full mesh. This
   results in an O(N) peering as opposed to O(N^2).

o TTL: Time To Live.

This document also assumes that the reader is familiar with the
terminology provided in [RFC7432].

1.2. Default Designated Forwarder (DF) Election in EVPN Services

   [RFC7432] defines the DF as the EVPN PE responsible for:

o Flooding BUM traffic on a given Ethernet Tag on a particular ES to
   the CE. This is valid for Single-Active and All-Active EVPN
   multihoming.

o Sending unicast traffic on a given Ethernet Tag on a particular ES
   to the CE. This is valid for Single-Active multihoming.

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

```
                      +---------------+
                      |   IP/MPLS     |
                      |   Core        |
   +----+ ES1 +----+  |               |  +----+
   | CE1|-----|    |  |               |  |    |---- ES2
   +----+     | PE1|  |               |  | PE2|     \
             |    |  |               |  +----+      \+----+
             +----+  |               |    |          | CE2|
               |     |               |  +----+      /+----+
               |     |               |  |    |---- /     |
               |     |               |  | PE3|    / ES2 /
               |     |               |  +----+   /      /
               |     |               |    |     /      /
               +-----+---------------+----+    /      /
                     |               | PE4|___/ES2
                     |               |    |
                     |               +----+
```
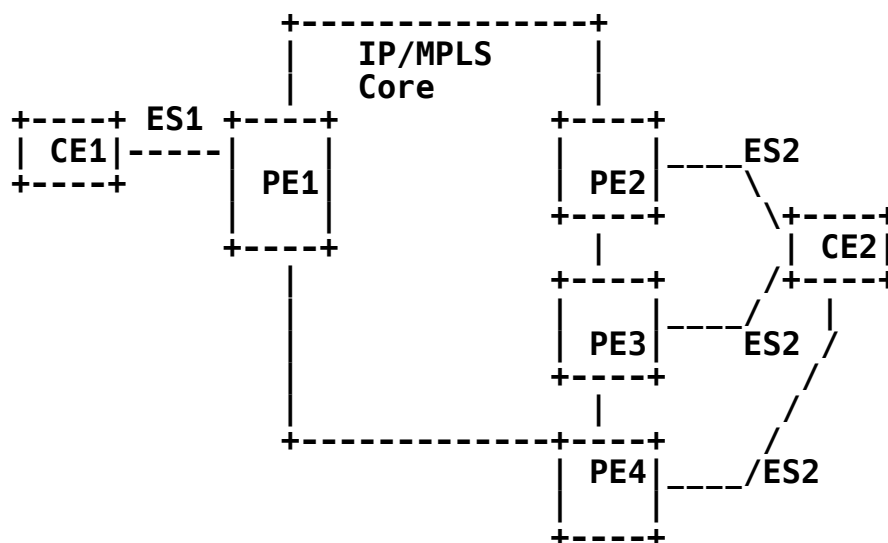
                  Figure 1: EVPN Multihoming

Figure 1 illustrates a case where there are two ESes: ES1 and ES2.
PE1 is attached to CE1 via 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 multihomed to different PEs on the same ES, 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 TTL. Therefore, it is very important that, in the case of
a multihomed CE, only one of the PEs be used to send BUM traffic
to it.

One of the prerequisites for this support is that participating PEs
must agree amongst themselves as to who would act as the 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 the default DF election algorithm
are used interchangeably. With service carving, it is possible to
elect multiple DFs per ES (one per EVI) in order to perform load

balancing of traffic destined to a given ES. 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 multihomed form an ordered
(ordinal) list in ascending order by PE IP address value. For
example, there are N PEs: PE0, PE1,... PE(N-1) ranked as per
increasing IP addresses in the ordinal list; then, for each VLAN with
Ethernet Tag V, configured on ES1, PEx is the DF for VLAN V on ES1
when x equals (V mod N). In the case of a VLAN Bundle, only the
lowest VLAN is used. In the case when the planned density is high
(meaning there are a 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 ES 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 changes in the EVPN
routes will be minimal.

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.

1.3. Problem Statement

This section describes some potential issues with the default DF election algorithm.

1.3.1. Unfair Load Balancing and Service Disruption

There are three fundamental problems with the current default DF election algorithm.

1. 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 multihomed to two PEs; one of the PEs will be elected as the DF for all of the VLANs. This is very suboptimal. It defeats the purpose of service carving, as the DFs are not really evenly spread across the PEs hosting the ES. In fact, in this particular case, one of the PEs does not get elected as the DF at all, so it does not participate in DF responsibilities at all. Consider another example where, referring to Figure 1, let's assume that (1) PE2, PE3, and PE4 are listed in ascending order by IP address and (2) 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 being 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 (Section 8.5 of [RFC7432]). N corresponds to M above. Since N, N-1, or N+1 need not satisfy the primality properties of M, as per the modulo-based DF assignment [RFC7432], whenever a PE goes down or a new PE boots up (attached to the same ES), the modulo scheme will not necessarily map BDs to PEs uniformly.

3. Disruption is another problem. Consider a case when the same ES
    is multihomed 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 ES.
    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 ES, 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 values 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 multihomed to the same ES (and interested in
the DF election by exchanging EVPN routes) use an Originating
Router's IP address [RFC7432] 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 multihomed
to the ES in question. 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 reprogramming 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.

1.3.2. Traffic Black-Holing on Individual AC Failures

   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 in 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 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
   protection against "logical" failures or human errors that may occur
   at the 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 AC defined in an ES is accidentally shut down
        or is not provisioned yet (hence, the 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 either shut down 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
   multihomed 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 tiebreaker and trigger in Virtual Private LAN Service (VPLS)
   multihoming procedures [VPLS-MH], there is no procedure defined in
   the EVPN specification [RFC7432] to trigger the DF re-election based
   on the ACS change on the DF.

Figure 2 shows an example of logical AC failure.

```
                              +---+
                              |CE4|
                              +---+
                                |
                     PE4        |
                     +-----+-----+
       +--------------|  +-----+  |--------------+
       |              |  | BD-1|  |              |
       |              +-----------+              |
       |                                         |
       |                     EVPN                |
       |                                         |
       |  PE1                PE2                PE3 |
       |  (NDF)              (DF)               (NDF)|
     +-----------+        +-----------+        +-----------+
     |  | BD-1|  |--------|  | BD-1|  |--------|  | BD-1|  |
     |  +-----+  |--------|  +-----+  |--------|  +-----+  |
     +-----------+        +-----------+        +-----------+
       AC1\    ES12   /AC2   AC3\   ES23   /AC4
          \         /            \        /
           \       /              \      /
           +----+                 +----+
           |CE12|                 |CE23|
           +----+                 +----+
```
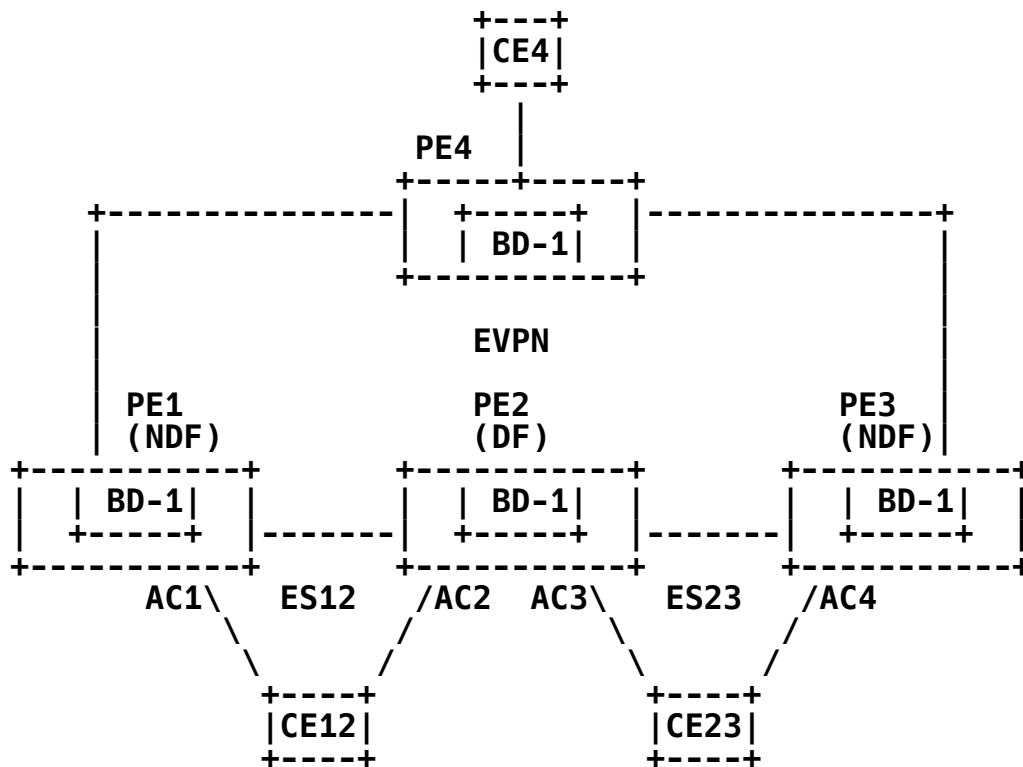
            Figure 2: Default DF Election and Traffic Black-Holing

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

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

(a)   If AC2 is accidentally shut down or is not configured yet, CE12
       traffic will be impacted. In the case of All-Active
       multihoming, the BUM traffic to CE12 will be "black-holed",
       whereas for Single-Active multihoming, all the traffic to/from
       CE12 will be discarded. This is because a logical failure in
       PE2's AC2 may not trigger an ES route withdrawal for ES12 (since
       there are still other ACs active on ES12); therefore, PE1 will
       not rerun the DF election procedures.

(b)   If the bridge table for BD-1 is administratively shut down or is
       not configured yet on PE2, CE12 and CE23 will both be impacted:
       BUM traffic to both CEs will be discarded in the case of
       All-Active multihoming, and all traffic will be discarded
       to/from the CEs in the case of Single-Active multihoming. This
       is because PE1 and PE3 will not rerun the DF election procedures
       and will keep assuming that 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; therefore, EVPN can provide protection against
logical failures.

1.4. The Need for Extending the Default DF Election in EVPN Services

Section 1.3 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:

o The new DF election algorithm is referred to as "Highest Random
   Weight" (HRW). The HRW procedures are described in Section 3.

o The new DF election capability is referred to as "AC-Influenced DF
   election" (AC-DF). The AC-DF procedures are described in
   Section 4.

o 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 the
   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 2.2 for more details.

2. 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] leaves several questions open as to the precise FSM
behavior of the DF election, Section 2.1 precisely describes the
intended behavior.

2.1. The DF Election Finite State Machine (FSM)

Per [RFC7432], the FSM shown in Figure 3 is executed per <ES, VLAN>
in the case of VLAN-based service or <ES, [VLANs in VLAN Bundle]> in
the case of a VLAN Bundle on each participating PE. Note that the
FSM is conceptual. Any design or implementation MUST comply with
behavior that is equivalent to the behavior outlined in this FSM.

```
                VLAN_CHANGE                  VLAN_CHANGE
                RCVD_ES                      RCVD_ES
                LOST_ES                      LOST_ES
                +----+                       +-------+
                |    |                        |       v
                |  +-+----+    ES_UP         ++-------++
                +->+ INIT +-------------->+ DF_WAIT |
                   ++-----+               +-------+-+
                    ^                               |
  +-----------+     |                               |DF_TIMER
  | ANY_STATE +-------+         VLAN_CHANGE          |
  +-----------+ ES_DOWN   +----------------+        |
                     |     RCVD_ES     v    v
            +--------++    LOST_ES    ++------+-+
            | DF_DONE +<-------------+ DF_CALC +<-+
            +--------+    CALCULATED  +-------+-+  |
                                       |       |
                                       +----+  |
                                       VLAN_CHANGE
                                       RCVD_ES
                                       LOST_ES
```
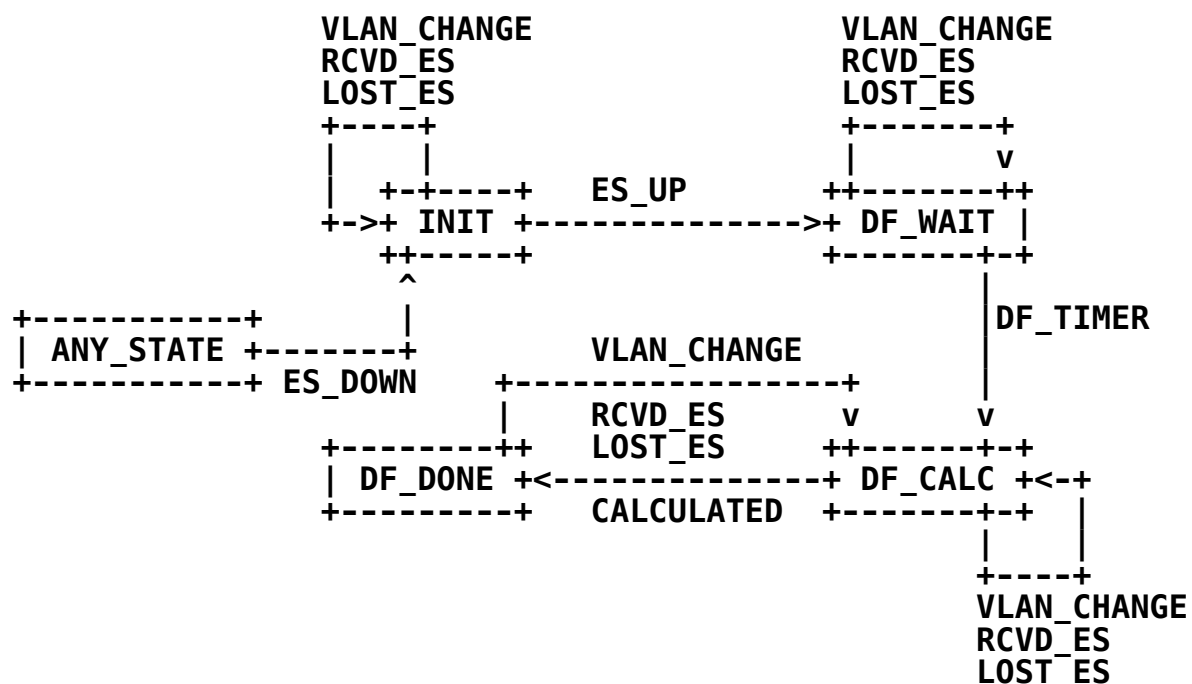
                 Figure 3: DF Election Finite State Machine

   Observe that each EVI is locally configured on each of the multihomed
   PEs attached to a given ES and that the FSM does not provide any
   protection against inconsistent configuration between these PEs.
   That is, for a given EVI, one or more of the PEs are inadvertently
   configured with a different set of VLANs for a VLAN-aware Bundle
   service or with different VLANs for a VLAN-based service.

   The states and events shown in Figure 3 are defined as follows.

   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 corresponding 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 ES has been locally configured as "UP".

2. ES_DOWN: The ES has been locally configured as "DOWN".

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

4. DF_TIMER: DF timer [RFC7432] (referred to as "Wait timer" in this
    document) has expired.

5. RCVD_ES: A new or changed ES route is received in an Update
    message with an MP_REACH_NLRI. Receiving an unchanged Update
    MUST NOT trigger this event.

6. LOST_ES: An Update message with an MP_UNREACH_NLRI for a
    previously received ES route has been received. If such a
    message is seen for a route that has not been advertised
    previously, the event MUST NOT be triggered.

7. CALCULATED: DF has been successfully calculated.

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

1. ANY_STATE on ES_DOWN:
    (i) Stop the DF Wait timer.
    (ii) Assume an NDF for the 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 the DF Wait timer if not started already or expired.
    (ii) Assume an NDF for the 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 the candidate list, perform a hash, and perform the
    election.
    (ii) Afterwards, the FSM generates a CALCULATED event against
    itself.

8.   DF_CALC on VLAN_CHANGE, RCVD_ES, or LOST_ES: Do as prescribed in
     Transition 7.

9.   DF_CALC on CALCULATED: Mark the election result for the VLAN or
     bundle, and transition to DF_DONE.

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

11. 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 4, the use of the AC-DF
capability introduces additional events and transitions.

2.2. 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 for some PEs to continue to use the default DF election
algorithm while some PEs use HRW. For brownfield deployments and for
interoperability with legacy PEs, it is important that all PEs have
the ability 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 ES route (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 ES. Figure 4 shows the
encoding of the DF Election Extended Community.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type = 0x06   | Sub-Type(0x06)| RSV |  DF Alg |    Bitmap     ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~    Bitmap     |              Reserved                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: DF Election Extended Community

Where:

o Type: 0x06, as registered with IANA (Section 7) for EVPN Extended
   Communities.

o Sub-Type: 0x06. "DF Election Extended Community", as registered
   with IANA.

o RSV/Reserved: Reserved bits for information that is specific to
   DF Alg.

o 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 creates an IANA registry called "DF Alg" (Section 7),
   which contains the following values:

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

   - Type 1: HRW Algorithm (Section 3).

   - Types 2-30: Unassigned.

   - Type 31: Reserved for Experimental Use.

o Bitmap (2 octets): Encodes "capabilities" to use with the DF
   election algorithm in the DF Alg field. This document creates an
   IANA registry (Section 7) for the Bitmap field, with values 0-15.
   This registry is called "DF Election Capabilities" and includes
   the bit values listed below.

```
                          1 1 1 1 1 1
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     | |A|                           |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

   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 Capability (AC-Influenced DF election; see
      Section 4). When set to 1, it indicates the desire to use
      AC-DF with the rest of the PEs in the ES.

   - Bits 2-15: Unassigned.

The DF Election Extended Community is used as follows:

o 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 DF 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 to signal the preferred one.

   - DF Alg values 0 and 1 can both be used with Bit 1 (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
      Alg values 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:

   - If 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, then the PEs attached to the ES will perform the DF
      election as per the HRW algorithm and following the AC-DF
      procedures.

   - Otherwise, if even a single advertisement for Route Type 4 is
      received without the locally configured DF Alg and capability,
      the default DF election algorithm MUST be used as prescribed in
      [RFC7432]. This procedure handles the case where participating
      PEs in the ES disagree about the DF algorithm and capability to
      be applied.

   - 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/with the existing DF Alg
      values must be assessed on a case-by-case basis.

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

2.2.1. Backward Compatibility

   Implementations that comply with [RFC7432] only (i.e.,
   implementations 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-DF.

   Similarly, an implementation that complies with [RFC7432] only and
   that receives a DF Election Extended Community will ignore it and
   will continue to use the default DF election algorithm.

3. The Highest Random Weight DF Election Algorithm

   The procedure discussed in this section is applicable to the DF
   election in EVPN services [RFC7432] and the EVPN Virtual Private Wire
   Service (VPWS) [RFC8214].

   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 server 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-to-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

pseudorandom 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].

3.1. 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.

3.2. HRW Algorithm for EVPN DF Election

This section describes the application of HRW to DF election. Let
DF(V) denote the DF and BDF(V) denote the BDF for the Ethernet Tag V;
Si is the IP address of PE i; Es is the ESI; and Weight is a function
of V, Si, and Es.

Note that while the DF election algorithm provided in [RFC7432] uses
a PE address and VLAN as inputs, this document uses an Ethernet Tag,
PE address, and ESI as inputs. This is because if the same set of
PEs are multihomed 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 BDs on each ES; this could have adverse
side effects on both load balancing and redundancy. Including an 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 BDs 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 the 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 the case of a tie, choose the PE whose IP address is
    numerically the least. Note that 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 the case of a tie,
    choose the PE whose IP address is numerically the least.

Where:

o DF(V) is defined to be the address Si (index i) for which
   Weight(V, Es, Si) is the highest; 0 <= i < N-1.

o 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 and k are selected values.

Since the Weight is a pseudorandom function with the 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 pseudorandom
function is an important consideration for this algorithm to perform
better than the default algorithm. As mentioned previously, such
functions are described in [HRW1999]. We take as a candidate hash
function the first one out of the two that are listed as 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
most significant bit (MSB), as noted in [HRW1999]) of the 14-octet
stream (the 4-octet Ethernet Tag V followed by the 10-octet ESI). It
is mandated that the 14-octet stream be formed by the concatenation
of the Ethernet Tag and the ESI in network byte order. The CRC
should proceed as if the stream is in network byte order
(big-endian). Si is the 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, the ES, 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 defined 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 1.3.1 of this
document and ensures that:

o With very high probability, the task of DF election for the VLANs
   configured on an ES is more or less equally distributed among the
   PEs, even in the case of two PEs (see the first fundamental
   problem listed in Section 1.3.1).

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 third fundamental problem listed
   in Section 1.3.1 (needless disruption) 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.

4. The AC-Influenced DF Election Capability

   The procedure discussed in this section is applicable to the DF
   election in EVPN services [RFC7432] and EVPN VPWS [RFC8214].

   The AC-DF capability is expected to be generally applicable to 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 4.1 describes the procedures for VLAN-aware Bundle service
   interfaces.

   In particular, when used with the default DF algorithm, the AC-DF
   capability modifies 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 ES
   (including itself), in increasing numeric value. The candidate
   list is based on the Originating 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 to be a 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 to be a candidate for a
   given BD.

   If the default DF algorithm 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 ES, 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 a VLAN (-aware) Bundle service,
then the numerically lowest VLAN value in that bundle on that ES
MUST be used in the modulo function as the Ethernet Tag.

It should be noted that using the Originating Router's IP Address
field [RFC7432] in the ES route to get the PE IP address needed
for the ordered list allows for a CE to be multihomed across
different Autonomous Systems (ASes) if such a need ever arises.

The modified Step 3, above, differs from [RFC7432], Section 8.5,
Step 3 in two ways:

o Any DF Alg can be used -- 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 to be a candidate DF for an
   <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 and PE2 advertise an Ethernet A-D per ES route for ES12.
     PE2 and PE3 advertise an Ethernet A-D per ES route for ES23.

(c) In addition, PE1, PE2, and 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 with 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 attached to the ES
         (including itself) as explained in the modified Step 3 above.
         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
         to be a 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 to be a
         candidate for a given BD. For example, PE1 will not consider
         PE2 as a 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 2.1, 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
   shown in Figure 3, the events below MUST trigger a transition from
   DF_DONE to DF_CALC:

   1.   Local AC going DOWN/UP.

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

   3.   Reception of a new Ethernet A-D per ES route update/withdrawal
        for the ES.

4.1. AC-Influenced DF Election Capability for VLAN-Aware Bundle
      Services

   The procedure described in Section 4 works for VLAN-based and VLAN
   Bundle service interfaces because, for those service types, a PE
   advertises only one Ethernet A-D per EVI route per <ES, VLAN> or
   <ES, VLAN Bundle>. In Section 4, an Ethernet Tag represents a given
   VLAN or VLAN Bundle for the purpose of DF election. The withdrawal

of such a 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 election.

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 and not
necessarily all the ACs in the <ES, VLAN Bundle>.

This document modifies the DF election for VLAN-aware Bundle services
in the following ways:

o 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> as described 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 election
   in the <ES, VLAN>.

o The PEs will now follow the procedures in Section 4.

For example, assuming three bridge tables in PE1 for the same MAC-VRF
(each one associated with 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 to be a
valid candidate PE for DF election in <ES12, VLAN-1>, <ES12, VLAN-2>,
and <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.

5. Solution Benefits

The solution described in this document provides the following
benefits:

(a) It extends the DF election as defined in [RFC7432] to address
     the unfair load balancing and potential black-holing issues with
     the default DF election algorithm. The solution is applicable
     to the DF election in EVPN services [RFC7432] and EVPN VPWS
     [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 allows the
        advertising PE to indicate its support for the capabilities
        defined in this document as well as any subsequently defined DF
        election algorithms or capabilities.

   (c) It 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 DF election as defined in
        [RFC7432].

6. 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 ES to exchange additional information and agree on the DF
   election type and capabilities to be used.

   By following the procedures in this document, the operator will
   minimize such undesirable situations as unfair load balancing,
   service disruption, and traffic black-holing. Because such
   situations could be purposely created by a malicious user with access
   to the configuration of one PE, this document also enhances the
   security of the network. Note that the network will not benefit from
   the new procedures if the DF election algorithm is not consistently
   configured on all the PEs in the ES (if there is no unanimity among
   all the PEs, the DF election algorithm falls back to the default DF
   election as provided in [RFC7432]). This behavior could be exploited
   by an attacker that manages to modify the configuration of one PE in
   the ES so that the DF election algorithm and capabilities in all the
   PEs in the ES 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 mentioned earlier.

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

7. IANA Considerations

   IANA has:

   o Allocated 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 exist:

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
      2-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 exists:

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

8. References

8.1. 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, <https://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>.

8.2. Informative References

   [VPLS-MH] Kothari, B., Kompella, K., Henderickx, W., Balus, F., and
              J. Uttaro, "BGP based Multi-homing in Virtual Private LAN
              Service", Work in Progress,
              draft-ietf-bess-vpls-multihoming-03, March 2019.

   [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, DOI 10.1145/258533.258660,
              May 1997.

   [CLRS2009] Cormen, T., Leiserson, C., Rivest, R., and C. Stein,
              "Introduction to Algorithms (3rd Edition)", MIT
              Press, ISBN 0-262-03384-8, 2009.

   [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and
              Multicast Next-Hop Selection", RFC 2991,
              DOI 10.17487/RFC2991, November 2000,
              <https://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,
              <https://www.rfc-editor.org/info/rfc2992>.

   [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route
              Reflection: An Alternative to Full Mesh Internal BGP
              (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006,
              <https://www.rfc-editor.org/info/rfc4456>.

   [HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings
              to Increase Hit Rates", IEEE/ACM Transactions on
              Networking, Volume 6, No. 1, February 1998,
              <https://www.microsoft.com/en-us/research/wp-content/
              uploads/2017/02/HRW98.pdf>.

   [Knuth]   Knuth, D., "The Art of Computer Programming: Volume 3:
              Sorting and Searching", 2nd Edition, Addison-Wesley,
              Page 516, 1998.

Contributors

   The following people have contributed substantially to this document
   and should be considered coauthors:

   Antoni Przygienda
   Juniper Networks, Inc.
   1194 N. Mathilda Ave.
   Sunnyvale, CA 94089
   United States of America

   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 (editor)
     Nokia
     777 E. Middlefield Road
     Mountain View, CA 94043
     United States of America

     Email: jorge.rabadan@nokia.com

     Satya Mohanty (editor)
     Cisco Systems, Inc.
     225 West Tasman Drive
     San Jose, CA 95134
     United States of America

     Email: satyamoh@cisco.com

     Ali Sajassi
     Cisco Systems, Inc.
     225 West Tasman Drive
     San Jose, CA 95134
     United States of America

     Email: sajassi@cisco.com

John Drake
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
United States of America

Email: jdrake@juniper.net

Kiran Nagaraj
Nokia
701 E. Middlefield Road
Mountain View, CA 94043
United States of America

Email: kiran.nagaraj@nokia.com

Senthil Sathappan
Nokia
701 E. Middlefield Road
Mountain View, CA 94043
United States of America

Email: senthil.sathappan@nokia.com