BESS Workgroup   J. Rabadan, Ed.
Internet Draft   Nokia
                                                  S. Mohanty, Ed.
Intended status: Standards Track   A. Sajassi
                                                            Cisco
                                                        J. Drake
                                                         Juniper
                                                      K. Nagaraj
                                                    S. Sathappan
                                                            Nokia

Expires: September 6, 2018   March 5, 2018

        Framework for EVPN Designated Forwarder Election Extensibility
                draft-ietf-bess-evpn-df-election-framework-00

Abstract

   The Designated Forwarder (DF) in EVPN networks is the PE responsible
   for sending broadcast, unknown unicast and multicast (BUM) traffic to
   a multi-homed CE, 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
   addition, this document 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.

Status of this Memo

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 September 6, 2018.

Copyright Notice

Table of Contents

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
   BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all
   capitals, as shown here.

   o AC and ACS - Attachment Circuit and Attachment Circuit Status. An
     AC has an Ethernet Tag associated to it.

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

   o DF, NDF and BDF - Designated Forwarder, Non-Designated Forwarder
     and Backup Designated Forwarder

   o Ethernet A-D per ES route - refers to [RFC7432] route type 1 or

Auto-Discovery per Ethernet Segment route.

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

o ES and ESI - Ethernet Segment and Ethernet Segment Identifier.

o EVI - EVPN Instance.

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

o HRW - Highest Random Weight

o VID and CE-VID - VLAN Identifier and Customer Equipment VLAN
  Identifier.

o Ethernet Tag - used to represent a Broadcast Domain that is
  configured on a given ES for the purpose of DF election. Note that
  any of the following may be used to represent a Broadcast Domain:
  VIDs (including double Q-in-Q tags), configured IDs, VNI,
  normalized VID, I-SIDs, etc., as along the representation of the
  broadcast domains is configured consistently across the multi-homed
  PEs attached to that ES.

o 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>.

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

2. Introduction

2.1. Default Designated Forwarder (DF) Election in EVPN

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

o 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.

o 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 and example that we will be used to explain the
Designated Forwarder function.

```
                    +---------------+
                    |   IP/MPLS     |
                    |   CORE        |
       +----+ ES1 +----+         +----+
       | CE1|-----|    |-----------|    |____ES2
       +----+     | PE1|         | PE2|    \
                  |    |--------  +----+     \+----+
                +----+         |    |         | CE2|
                  |            |  +----+      /+----+
                  |            |__|   |____/    |
                  |            | PE3|   ES2 /
                  |            +----+      /
                  |              |        /
         +-------------+----+      /
                       | PE4|____/ES2
                       |    |
                       +----+
```
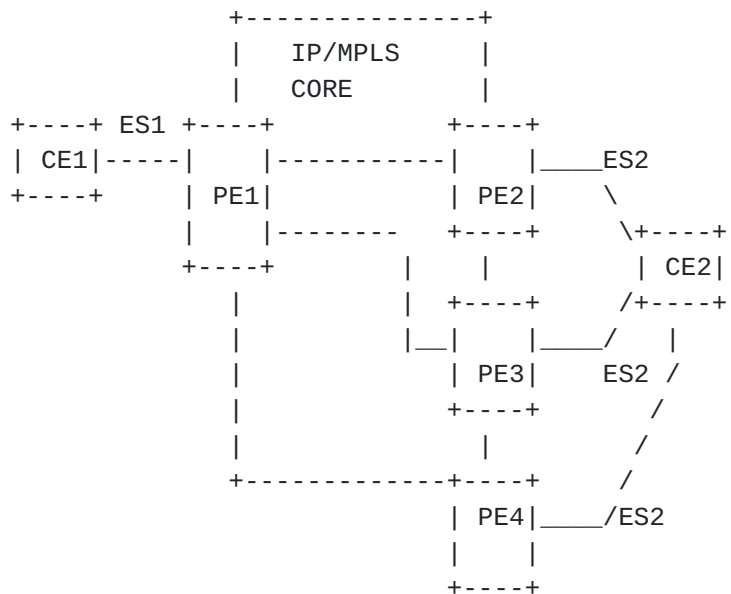
                Figure 1 Multi-homing Network of EVPN

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.

Layer-2 devices are particularly susceptible to forwarding loops
because of the broadcast nature of the Ethernet traffic. Therefore it
is very important that, in case of multi-homing, only one of the
links be used to direct traffic to/from the core.

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 unanimously in agreement.

The default procedure for DF election defined by [RFC7432] at the granularity of (ESI,EVI) is referred to as "service carving". In this document, service carving or default DF Election algorithm is used indistinctly. 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 BDspace among the redundant PE nodes evenly, in such a way that every PE is the DF for a disjoint 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 VLAN) is multi-homed from 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 service carving can be achieved.

The described default DF Election algorithm has some undesirable properties and in some cases can be somewhat disruptive and unfair. This document describes those issues and proposes a mechanism for dealing with them. These mechanisms do involve changes to the default DF Election algorithm, however they do not require any protocol changes to the EVPN Route exchange and have minimal changes to their content per se.

Note that while [RFC7432] elects a DF per <ES, EVI>, this document elects a DF per <ES, BD>. This means that unlike [RFC 7432], 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.

2.2. Problem Statement

This section describes some potential issues on the default DF Election algorithm.

2.2.1. Unfair Load-Balancing and Service Disruption

There are three fundamental problems with the current 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; all the VLANs will only pick one
   of the PEs as the DF. This is very sub-optimal. It defeats the
   purpose of service carving as the DFs are not really evenly spread
   across. In this particular case, in fact one of the PEs does not
   get elected all as the DF, 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 of the form (3x+1), where x is an integer.
   This will result in PE3 always be selected as the DF.

2- Even in the case when the Ethernet Tag distribution is uniform the
   instance of a PE being up or down results in re-computation ((v
   mod N-1) or (v mod N+1) as is the case); the resulting modulus
   value need not be uniformly distributed because it can be subject
   to the primality of N-1 or N+1 as may be the case.

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 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 mod 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 10001 respectively. So PE1, PE2 and PE3 are also 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 current 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 have a V4
peering with each other or via a Route Reflector. This need not be
the case as there can be a v6 peering and supporting the EVPN
address-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 programming the CE and PE side ports as blocking or non-blocking at potentially all PEs when the DF changes either because (i) a new PE is added or (ii) another one goes down or loses connectivity or else cannot take part in the DF election process for whatever reason. This document addresses this problem and furnishes a solution to this undesirable behavior.

2.2.2. 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 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 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.

```
                            +---+
                            |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|
            +----+                  +----+
```
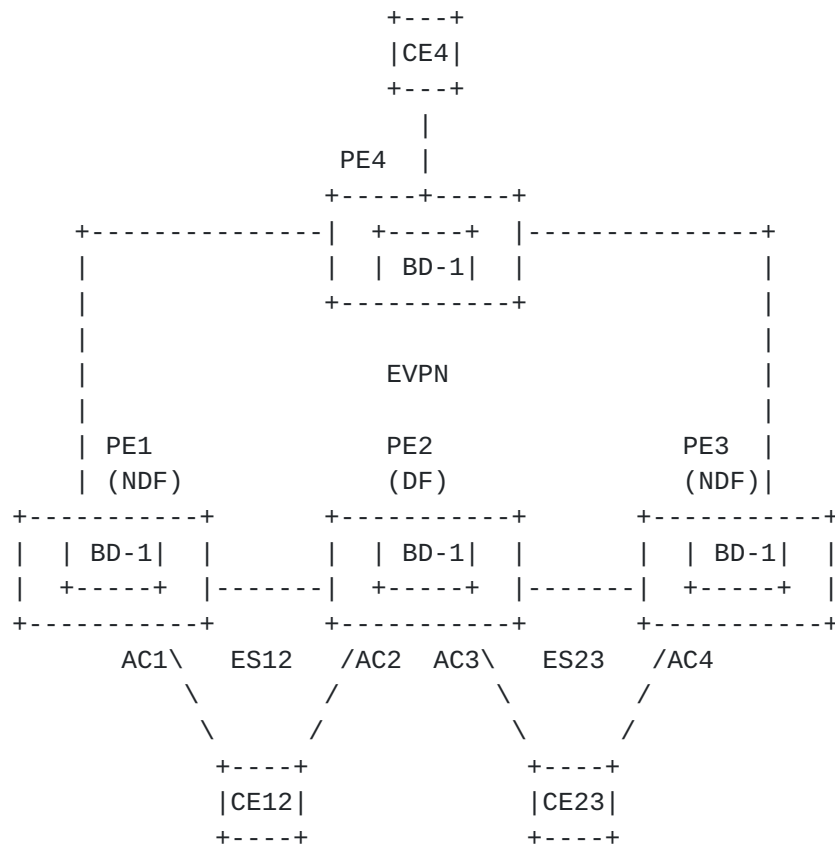
Figure 2 Default DF Election and Traffic Black-Holing

BD-1 is defined in PE1, PE2, PE3 and PE4. CE12 is a 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 running 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 modifies the default 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.

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

Section 2.2 describes some of the issues that exist in the default DF
Election procedures. In order to address those issues, this document
describes 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 4.

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

o Both, HRW and AC-DF MAY be used independently or simultaneously.
The AC-DF capability MAY be used with the default DF Election
algorithm too.

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.

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 specification in EVPN
[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 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 normative in the sense that any design or implementation
MUST behave towards external peers and as observable external
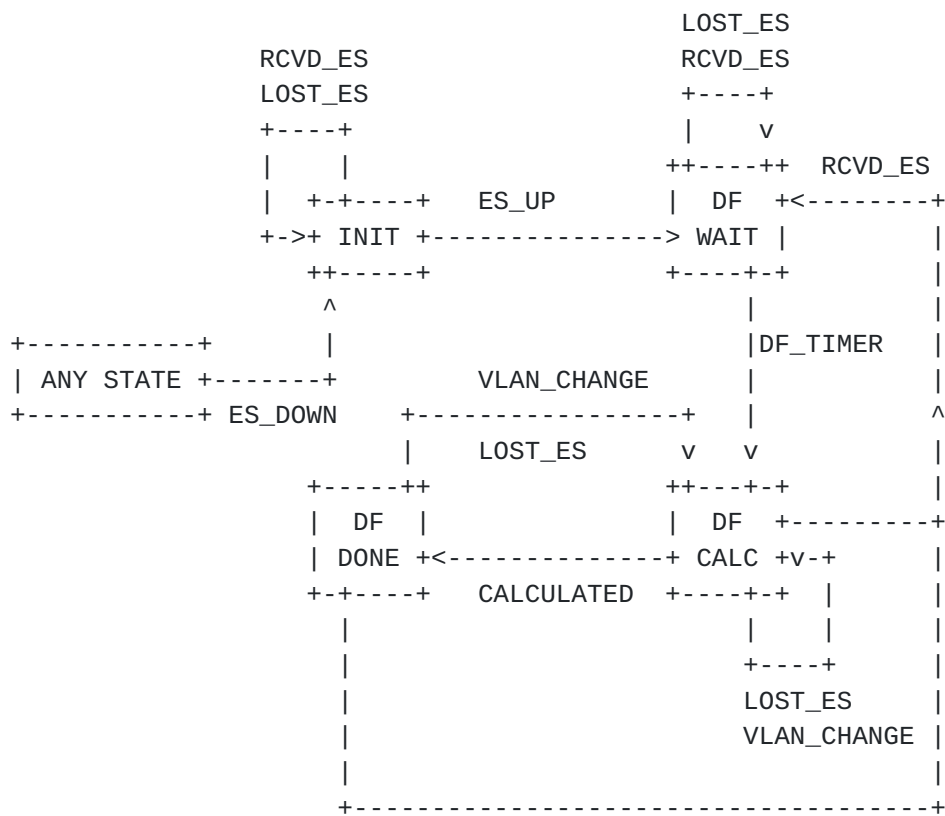behavior (DF) in a manner equivalent to this FSM.

```
                                    LOST_ES
                    RCVD_ES         RCVD_ES
                    LOST_ES         +----+
                    +----+          |    v
                    |    |          ++----++  RCVD_ES
                    |  +-+----+   ES_UP   |  DF  +<--------+
                    +->+ INIT +--------------> WAIT |         |
                      ++-----+          +----+-+         |
                        ^               |         |
       +-----------+      |                  |DF_TIMER   |
       | ANY STATE +-------+      VLAN_CHANGE    |         |
       +-----------+ ES_DOWN    +----------------+   |         ^
                      |    LOST_ES       v   v         |
                    +-----++          ++---+-+         |
                    |  DF  |          |  DF  +---------+
                    | DONE +<--------------+ CALC +v-+      |
                    +-+----+   CALCULATED  +----+-+  |      |
                      |                  |    |  |
                      |                  |    +----+  |
                      |                  LOST_ES    |
                      |                  VLAN_CHANGE |
                      |                  |      |
                      +------------------------------------+
```

                 Figure 3 DF Election Finite State Machine

   States:

   1. INIT: Initial State

   2. DF WAIT: State in which the participants 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.

   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 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 timer (ii) assume non-DF for
    local PE.

2. INIT on ES_UP: (i)do nothing.

3. INIT on RCVD_ES, LOST_ES: (i)do nothing.

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

5. DF_WAIT on RCVD_ES, LOST_ES: do nothing.

6. DF_WAIT on DF_TIMER: do nothing.

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

8. DF_CALC on LOST_ES or VLAN_CHANGE: do nothing.

9. DF_CALC on RCVD_ES: do nothing.

10. DF_CALC on CALCULATED: (i) mark election result for VLAN or
    bundle.

11. DF_DONE on exiting the state: (i)if RFC7432 election or new
    election and lost primary DF then assume non-DF for local PE for
    VLAN or VLAN-Bundle.

12. DF_DONE on VLAN_CHANGE or LOST_ES: do nothing.

3.2 The DF Election Extended Community

   For the DF election procedures to be globally convergent and
   unanimous, it is necessary that all the participating PEs agree on
   the DF Election algorithm 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 boxes, its 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.

```
    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)|   DF Type    |    Bitmap      |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                        Reserved = 0                          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

             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.

   o DF Type (1 octet) - Encodes the DF Election algorithm values
     (between 0 and 255) that the advertising PE desires to use for the
     ES. This document requests IANA to set up a registry called "DF
     Type 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 2-254: Unassigned.

     - Type 255: Reserved for Experimental Use.

o Bitmap (1 octet) - Encodes "capabilities" associated to the DF
  Election algorithm in the field "DF Type". This document requests
  IANA to create a registry for the Bitmap field, called "DF Election
  Capabilities" and solicits the following values:

  - Bit 24: Unassigned.

  - Bit 25: 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 26-31: 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 for DF Type HRW and/or AC-DF. In the
  Extended Community, the PE indicates the desired "DF Type"
  algorithm and "Bitmap" capabilities to be used for the ES. Only one
  DF Election Extended Community can be sent along with an ES route.

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

  - In general, a specific DF Type MAY determine the use of the
    reserved bits in the Extended Community. In case of DF Type HRW,
    the reserved bits will be sent as 0 and will be ignored on
    reception.

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 Type and Bitmap:

  - In the case that they do, this particular PE will follow the
    procedures for the advertised DF Type and capabilities. For
    instance, if all ES routes for a given ES indicate DF Type 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
    not received with the locally configured DF Type and capability,
    the default DF Election algorithm (modulus) algorithm MUST be
    used as in [RFC7432].

  - The absence of the DF Election Extended Community MUST be
    interpreted by a receiving PE as an indication of the default DF
    Election algorithm on the sending PE, that is, DF Type 0 and no

DF Election capabilities.

o When all the PEs in an ES advertise DF Type 255, they will rely on
  the local policy to decide how to proceed with the DF Election.

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. This document
extends the same auto-derivation procedure to ESI types 0, 4, and 5.

4. **The Highest Random Weight DF Election Type**

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 state of 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 Unix utilities rand and srand and easily
constructed XOR functions that perform considerably well. This
imparts very good properties in the load balancing context. Also each
server independently and unambiguously arrives at the primary server
selection. HRW already finds use in multicast and ECMP [RFC2991],
[RFC2992].

In the default DF Election algorithm (Section 2.1), whenever a new PE
comes up or an existing PE goes down, there is a significant interval

before the change is noticed by all peer PEs as it has to be conveyed
by the BGP update message involving the type-4 route. There is a
timer to batch all the messages before triggering the service carving
procedures.

When the timer expires, each PE will build the ordered list and
follow the procedures for DF Election. In the proposed method which
we will describe shortly this "jittered" behavior is retained.

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

The applicability of HRW to DF Election is described here. 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 server i, Es denotes the Ethernet Segment Identifier and
weight is a pseudo-random function of v and Si.

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

Since the Weight is a Pseudo-random function with domain as the
three-tuple (v, Es, S), it is an efficient deterministic algorithm
which 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 probably better
than the default algorithm. As mentioned previously, such functions
are described in the HRW paper. We take as candidate hash functions
two of the ones that are preferred in [HRW1999].

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

2. Wrand2(v, Es, Si) = (1103515245((1103515245.D(v,Es)+12345)XOR
    Si)+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). 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. Although both the
above hash functions perform similarly, we select the first hash
function (1) of choice, as the hash function has to be the same in
all the PEs participating in the DF election.

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 V4
or V6) is not really relevant The existing algorithm in [RFC7432] as
is cannot employ both V4 and V6 neighbor peering address.

HRW solves the disadvantage pointed out in Section 2.2.1 and ensures:

o with very high probability that the task of DF election for
  respective VLANs is more or less equally distributed among the PEs
  even for the 2 PE case.

o If a PE, hosting some VLANs on given ES, but is neither the DF nor
  the BDF for that VLAN, goes down or its connection to the ES goes
  down, it does not result in a DF and BDF reassignment the other
  PEs. This saves computation, especially in the case when the
  connection flaps.

o More importantly it avoids the needless disruption case of Section
  2.2.1 (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 MAY be used with any "DF Type" algorithm. It
modifies the default DF Election procedures in [RFC7432] 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, the AC-DF capability modifies the Step 3 in the default 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 connected 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, excluding all the PEs for which no Ethernet A-D per ES route has been received, or for which the route has been withdrawn. Afterwards, the DF Election algorithm is applied on a per <ES,VLAN> or <ES,VLAN-bundle>, however, the IP address for a PE will not be considered candidate for a given <ES,VLAN> or <ES,VLAN-bundle> 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.

The above paragraph differs from [RFC7432] Section 8.5, Step 3, in two aspects:

o Any DF Type algorithm can be used, and not only the modulus-based one (which is the default DF Election, or DF Type 0 in this document).

o The candidate list is pruned based on the 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,VLAN> or <ES,VLAN-bundle>, if its Ethernet A-D per EVI route for the <ES,VLAN> or <ES,VLAN-bundle> respectively, is withdrawn.

The following example illustrates the AC-DF behavior, 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 timer at the same time. Likewise, PE2 and PE3 advertise an ES route for ES23 with AC-DF=1 and start a 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. All the PEs for which no Ethernet A-D per ES route has been received, are 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 procedure described above, the following events SHALL modify the candidate PE list and trigger the DF re-election in a PE for a given <ES,VLAN> or <ES,VLAN-Bundle>:

i.   Local ES going DOWN due to a physical failure or reception of an ES route withdraw for that ES.

ii.   Local ES going UP due to its detection/configuration or reception of a new ES route update for that ES.

iii. Local AC going DOWN/UP.

iv.   Reception of a new Ethernet A-D per EVI update/withdraw for the <ES,VLAN> or <ES,VLAN-Bundle>.

v.   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 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>. 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:

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> in [RFC7432]. Now, the withdrawal
   of an Ethernet 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>.

o 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>.

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

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 as "DF Election Extended Community" in
the "EVPN Extended Community Sub-Types" registry.

o Set up a registry "DF Type" for the DF Type octet in the Extended
Community. The following values in that registry are requested:

  - Type 0: Default DF Election.
  - Type 1: HRW algorithm.
  - Type 255: Reserved for Experimental use.

o Set up a registry "DF Election Capabilities" for the Bitmap octet
in the Extended Community. The following values in that registry
are requested:

  - Bit 25: AC-DF capability.

o The registration policy for the two registries is "Specification
Required".

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>.

   [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.

   [I-D.ietf-idr-extcomm-iana] Rosen, E. and Y. Rekhter, "IANA
              Registries for BGP Extended Communities", draft-ietf-idr-extcomm-
              iana-02 (work in progress), December 2013.

   [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
              Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March
              1997, <http://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>.

   [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A
              Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271,
              January 2006, <http://www.rfc-editor.org/info/rfc4271>.

   [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>.

9.2. Informative References

   [VPLS-MH] Kothari, Henderickx et al., "BGP based Multi-homing in
              Virtual Private LAN Service", draft-ietf-bess-vpls-multihoming-
              01.txt, work in progress, January, 2016.

   [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>.

10. Acknowledgments

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.com

      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