BESS Workgroup
Internet Draft

Intended status: Informational

J. Rabadan
K. Nagaraj
S. Sathappan
V. Prabhu
W. Henderickx
Nokia

A. Liu
Ericsson

W. Lin
Juniper Networks

Expires: April 14, 2018

October 11, 2017

AC-Influenced Designated Forwarder Election for EVPN
draft-ietf-bess-evpn-ac-df-02

Abstract

The Designated Forwarder (DF) in EVPN networks is the PE responsible for sending multicast, broadcast and unknown unicast traffic to a multi-homed CE, on a given Ethernet Tag on a particular Ethernet Segment (ES). The DF is selected based on the list of PEs that advertise the Ethernet Segment Identifier (ESI) to the EVPN network. While PE node or link failures trigger the DF re-election for a given <ESI, EVI>, individual Attachment Circuit (AC) or MAC-VRF failures do not trigger such DF re-election and the traffic may therefore be permanently impacted, even though there is an alternative path. This document improves the DF election algorithm so that the AC status can influence the result of the election and this type of "logical" failures can be protected too.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months

and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

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

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

This Internet-Draft will expire on April 14, 2018.

Copyright Notice

Table of Contents

1. Problem Statement

[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.

The default DF election algorithm defined by [RFC7432] is called service-carving and, for a given ES, is based on a (V mod N)= i function that provides a local DF election of a PEi at <ESI, EVI> level. V is the Ethernet Tag associated to the EVI (the numerically lowest Ethernet Tag value in case of multiple Ethernet Tags), whereas N is the number of PEs for which ES routes have been successfully imported. In other words, EVPN's service-carving takes into account only two variables in the DF election for a given ESI: 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 withdrawn will make the Non-DF (NDF) PEs re-elect the DF for that <ESI, EVI> and the service will be recovered.

However the current 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 an ES defined - 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 AC 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-MH], there is no procedure defined in [RFC7432] to trigger the DF re-election based on the ACS change on the DF.

This document improves the [RFC7432] service-carving 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. Solution Description

The ACS for a given Ethernet Tag on an ES is implicitly conveyed in the corresponding EVPN A-D per EVI route for that given <ESI, Ethernet Tag>. This section describes how to use the A-D per EVI routes to improve the DF election algorithm.

Figure 1 illustrates an example EVPN network that will be used to describe the proposed solution.

EVI-1 is defined in PE-1, PE-2, PE-3 and PE-4. CE12 is a multi-homed CE connected to ESI12 in PE-1 and PE-2. Similarly CE23 is multi-homed to PE-2 and PE-3 using ESI23. Both, CE12 and CE23, are connected to EVI-1 through VLAN-based service interfaces: CE12-VID 1 (VLAN ID 1 on CE12) is associated to AC1 and AC2 in EVI-1, whereas CE23-VID 1 is associated to AC3 and AC4 in EVI-1. Note that there are other ACs defined on these ES mapped to different EVIs.

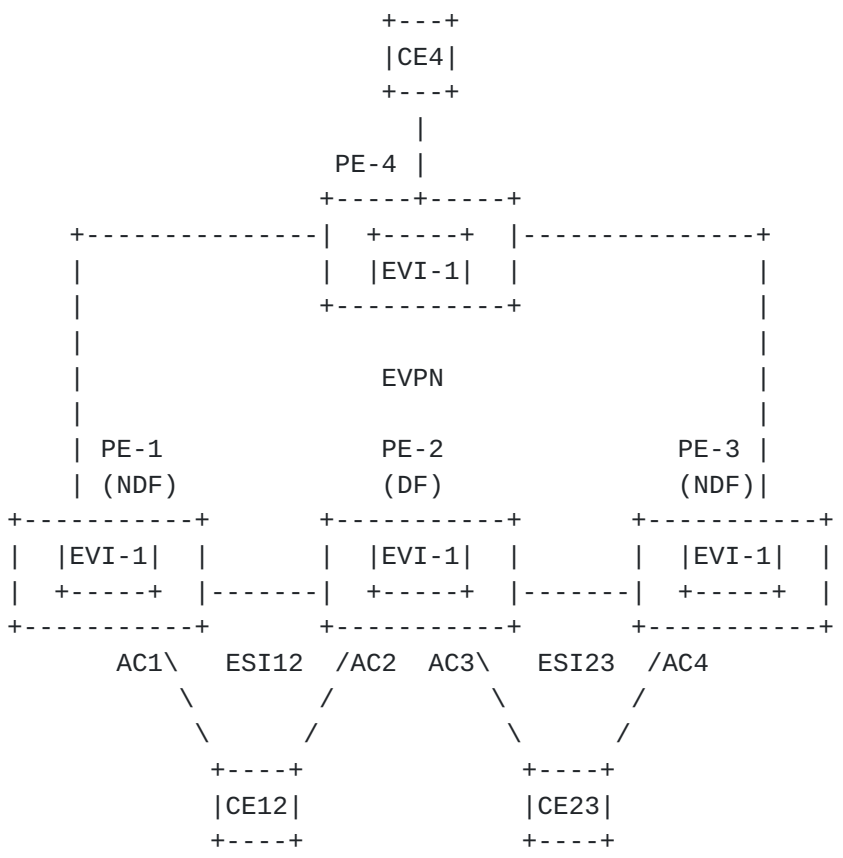

Figure 1 EVPN network example

2.1. Current DF Election Procedure And AC Failures

After running the service-carving DF election algorithm, PE-2 turns out to be the DF for ESI12 and ESI23 in EVI-1. The following two examples illustrate the issues with the existing defined procedure in [RFC7432]:

a) If AC2 is accidentally shutdown or even not configured, CE12 traffic will be impacted. In case of all-active multi-homing, only the BUM traffic to CE12 will be impacted, 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 PE-2 AC2 may not trigger an ES route withdrawn for ESI12 (since there are still other ACs active on ESI12) and therefore PE-1 will not re-run the DF election procedures.

b) If EVI-1 is administratively shutdown or even not configured yet on PE-2, 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 PE-1 and PE-3 will not

re-run the DF election procedures and will keep assuming PE-2 is the DF.

According to [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.

2.2. The Attachment Circuit (AC) Influenced DF Election

Modifying the service-carving DF election procedure in the following way solves the issue:

1. When PE-1 and PE-2 discover ESI12, they advertise an ES route for ESI12 with the associated ES-import extended community, starting a timer at the same time. Likewise, PE-2 and PE-3 advertise an ES route for ESI23 and start a timer.

2. Similarly, PE-1 and PE-2 advertise an Ethernet A-D per ES route for ESI12, and PE-2/PE-3 advertise an Ethernet A-D per ES route for ESI23.

3. In addition, PE-1/PE-2/PE-3 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).

4. 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 order. 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.

5. When electing the DF for a given EVI, 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 ESI for a given PE must be UP so that the PE is considered as candidate for a given EVI. For example, PE-1 will not consider PE-2 as candidate for DF election for <ESI12, EVI-1> until an Ethernet A-D per EVI route is received from PE-2 for <ESI12, EVI-1>.

6. Once the PEs with ACS = DOWN for a given EVI have been eliminated

from the candidate list, the (V mod N) = i function can be applied for the remaining N candidates, as per [RFC7432].

Note that this procedure does not modify the existing EVPN control plane whatsoever. It only modifies the candidate list of PEs taken into account for the DF election algorithm defined in [RFC7432].

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 <ESI,EVI>:

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

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

c) Local AC going DOWN/UP.

d) Reception of a new Ethernet A-D per EVI update/withdraw for the <ESI, EVI>.

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

This procedure is backwards compatible with the DF election procedures described in [RFC7432] since it does not add any new extension in the control plane, however, a PE not supporting the procedures in this document SHOULD NOT share a multi-homed ES with a PE following this solution since both PEs may end up with an inconsistent view on who the DF is. The AC influenced DF election procedures SHOULD be enabled by an administrative option and only used when all the PEs in the ES support it.

2.3. AC-Influenced DF Election For VLAN-Aware Bundle Services

The procedure described section 2.2 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 <ESI,EVI>. The withdrawal of such route means that the PE cannot forward traffic on that particular <ESI,EVI>.

In VLAN-aware bundle services, the PE advertises multiple Ethernet A-D per EVI routes per <ESI,EVI> (one route per Ethernet Tag). The withdrawal of an individual route only indicates the unavailability of a specific AC but not necessarily all the ACs in the <ESI,EVI>.

For the specific case of VLAN-aware bundle services, the DF election

will be influenced by the update/withdraw of any of the Ethernet A-D per EVI routes in the <ESI,EVI>.

For example, assuming three bridge tables in PE-1 for the same MAC-VRF (each one associated to a different Ethernet Tag), PE-1 will advertise three Ethernet A-D per EVI routes for <ESI12,EVI1>. Each of the three routes will indicate the status of each AC in <ESI12,EVI1>. PE-1 will be considered as a valid candidate PE for DF election as long as the three routes are active. If PE-1 withdraws one or more of the Ethernet A-D per EVI routes for <ESI12,EVI1>, the PEs in ESI12 will not consider PE-1 as a suitable DF candidate for <ESI12,EVI1>.

3. Solution benefits

The solution described in this document provides the following benefits:

a) Improves the DF election procedures for EVPN so that failures due to human errors, logical failures or even delay in provisioning of Attachment Circuits can be protected by multi-homing.

b) It does not modify or add any BGP new attributes or NLRI changes.

c) It is backwards compatible with the procedures defined in RFC7432.

4. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

5. Security Considerations

The same Security Considerations described in [RFC7432] are valid for this document.

6. IANA Considerations

There are no new IANA considerations in this document.

7. References

7.1. Normative References

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

[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, <http://www.rfc-editor.org/info/rfc7432>.

7.2. 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.

8. Acknowledgments

Will be added.

Authors' Addresses

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

Kiran Nagaraj
Nokia
Email: kiran.nagaraj@nokia.com

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

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

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

Autumn Liu
Ericsson
Email: autumn.liu@ericsson.com

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