BESS Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan, Ed.
S. Sathappan
Nokia

S. Boutros
VMware

T. Przygienda
W. Lin
J. Drake
Juniper Networks

A. Sajassi
S. Mohanty
Cisco Systems

Expires: December 23, 2017

June 21, 2017

Preference-based EVPN DF Election
draft-ietf-bess-evpn-pref-df-00

Abstract

RFC7432 defines the Designated Forwarder (DF) in (PBB-)EVPN networks as the PE responsible for sending broadcast, multicast and unknown unicast traffic (BUM) to a multi-homed device/network in the case of an all-active multi-homing ES, or BUM and unicast in the case of single-active multi-homing.

The DF is selected out of a candidate list of PEs that advertise the Ethernet Segment Identifier (ESI) to the EVPN network, according to the 'service-carving' algorithm.

While 'service-carving' provides an efficient and automated way of selecting the DF across different EVIs or ISIDs in the ES, there are some use-cases where a more 'deterministic' and user-controlled method is required. At the same time, Service Providers require an easy way to force an on-demand DF switchover in order to carry out some maintenance tasks on the existing DF or control whether a new active PE can preempt the existing DF PE.

This document proposes an extension to the current RFC7432 DF election procedures so that the above requirements can be met.

Status of this Memo

This Internet-Draft is submitted in full conformance with the

Copyright Notice

Table of Contents

1. Problem Statement

RFC7432 defines the Designated Forwarder (DF) in (PBB-)EVPN networks as the PE responsible for sending broadcast, multicast and unknown unicast traffic (BUM) to a multi-homed device/network in the case of an all-active multi-homing ES or BUM and unicast traffic to a multi-homed device or network in case of single-active multi-homing.

The DF is selected out of a candidate list of PEs that advertise the Ethernet Segment Identifier (ESI) to the EVPN network and according to the 'service-carving' algorithm.

While 'service-carving' provides an efficient and automated way of selecting the DF across different EVIs or ISIDs in the ES, there are some use-cases where a more 'deterministic' and user-controlled method is required. At the same time, Service Providers require an easy way to force an on-demand DF switchover in order to carry out some maintenance tasks on the existing DF or control whether a new active PE can preempt the existing DF PE.

This document proposes an extension to the current RFC7432 DF election procedures so that the above requirements can be met.

2. Solution requirements

This document proposes an extension of the RFC7432 'service-carving' DF election algorithm motivated by the following requirements:

a) The solution MUST provide an administrative preference option so that the user can control in what order the candidate PEs may become DF, assuming they are all operationally ready to take over.

b) This extension MUST work for RFC7432 Ethernet Segments (ES) and virtual ES, as defined in [vES].

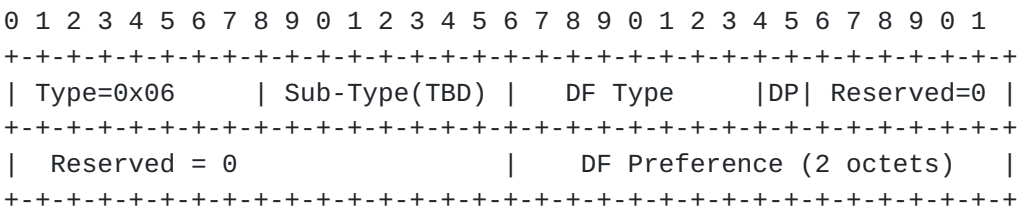
c) The user MUST be able to force a PE to preempt the existing DF for a given EVI/ISID without re-configuring all the PEs in the ES.

d) The solution SHOULD allow an option to NOT preempt the current DF, even if the former DF PE comes back up after a failure. This is also known as "non-revertive" behavior, as opposed to the RFC7432 DF election procedures that are always revertive.

e) The solution MUST work for single-active and all-active multi-homing Ethernet Segments.

3. EVPN BGP Attributes for Deterministic DF Election

This solution reuses and extends the DF Election Extended Community defined in [EVPN-HRW-DF] that is advertised along with the ES route:



Where the following fields are re-defined as follows:

o DF Type can have the following values:

- Type 0 - Default, mod based DF election as per RFC7432.
- Type 1 - HRW algorithm as per [EVPN-HRW-DF]
- Type 2 - Preference algorithm (this document)

o DP or 'Don't Preempt' bit, determines if the PE advertising the ES route requests the remote PEs in the ES not to preempt it as DF. The default value is DP=0, which is compatible with the current 'preempt' or 'revertive' behavior in RFC7432. The DP bit SHOULD be ignored if the DF Type is different than 2.

o DF Preference defines a 2-octet value that indicates the PE preference to become the DF in the ES. The allowed values are within the range 0-65535, and default value MUST be 32767. This value is the midpoint in the allowed Preference range of values, which gives the operator the flexibility of choosing a significant number of values, above or below the default Preference.

4. Solution description

Figure 1 illustrates an example that will be used in the description of the solution.

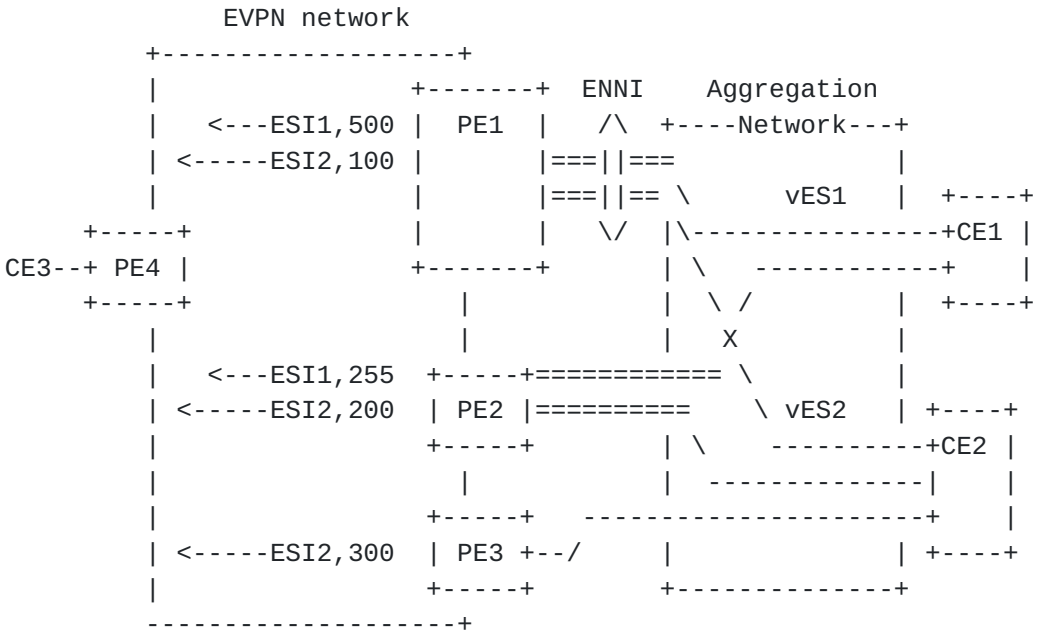


Figure 1 ES and Deterministic DF Election

Figure 1 shows three PEs that are connecting EVCs coming from the Aggregation Network to their EVIs in the EVPN network. CE1 is connected to vES1 - that spans PE1 and PE2 - and CE2 is connected to vES2, that is defined in PE1, PE2 and PE3.

If the algorithm chosen for vES1 and vES2 is type 2, i.e. Preference-based, the PEs may become DF irrespective of their IP address and based on an administrative Preference value. The following sections provide some examples of the new defined procedures and how they are applied in the use-case in Figure 1.

4.1 Use of the Preference algorithm

Assuming the operator wants to control - in a flexible way - what PE becomes the DF for a given vES and the order in which the PEs become DF in case of multiple failures, the following procedure may be used:

a) vES1 and vES2 are now configurable with three optional parameters that are signaled in the DF Election extended community. These parameters are the Preference, Preemption option (or "Don't

Preempt Me" option) and DF algorithm type. We will represent these parameters as [Pref,DP,type]. Let's assume vES1 is configured as [500,0,Pref] in PE1, and [255,0,Pref] in PE2. vES2 is configured as [100,0,Pref], [200,0,Pref] and [300,0,Pref] in PE1, PE2 and PE3 respectively.

b) The PEs will advertise an ES route for each vES, including the 3 parameters in the DF Election Extended Community.

c) According to RFC7432, each PE will wait for the DF timer to expire before running the DF election algorithm. After the timer expires, each PE runs the Preference-based DF election algorithm as follows:

- o The PE will check the DF type in each ES route, and assuming all the ES routes are consistent in this DF type and the value is 2 (Preference-based), the PE will run the new extended procedure. Otherwise, the procedure will fall back to RFC7432 'service-carving'.

- o In this extended procedure, each PE builds a list of candidate PEs, ordered based on the Preference. E.g. PE1 will build a list of candidate PEs for vES1 ordered by the Preference, from high to low: PE1>PE2. Hence PE1 will become the DF for vES1. In the same way, PE3 becomes the DF for vES2.

d) Note that, by default, the Highest-Preference is chosen for each ES or vES, however the ES configuration can be changed to the Lowest-Preference algorithm as long as this option is consistent in all the PEs in the ES. E.g. vES1 could have been explicitly configured as type Preference-based with Lowest-Preference, in which case, PE2 would have been the DF.

e) Assuming some maintenance tasks had to be executed on PE3, the operator could set vES2's preference to e.g. 50 so that PE2 is forced to take over as DF for vES2. Once the maintenance on PE3 is over, the operator could decide to leave the existing preference or configure the old preference back.

f) In case of equal Preference in two or more PEs in the ES, the tie-breakers will be the DP bit and the lowest IP PE in that order. For instance:

- o If vES1 parameters were [500,0,Pref] in PE1 and [500,1,Pref] in PE2, PE2 would be elected due to the DP bit.

- o If vES1 parameters were [500,0,Pref] in PE1 and [500,0,Pref] in PE2, PE1 would be elected, assuming PE1's IP address is lower

than PE2's.

g) The Preference is an administrative option that MUST be configured on a per-ES basis from the management plane, but MAY also be dynamically changed based on the use of local policies. For instance, on PE1, ES1's Preference can be lowered from 500 to 100 in case the bandwidth on the ENNI port is decreased a 50% (that could happen if e.g. the 2-port LAG between PE1 and the Aggregation Network loses one port). Policies MAY also trigger dynamic Preference changes based on the PE's bandwidth availability in the core, of specific ports going operationally down, etc. The definition of the actual local policies is out of scope of this document. The default Preference value is 32767.

4.2 Use of the Preference algorithm in RFC7432 Ethernet-Segments

While the Preference-based DF type described in section 4.1 is typically used in virtual ES scenarios where there is normally an individual EVI per vES, the existing RFC7432 definition of ES allows potentially up to thousands of EVIs on the same ES. If this is the case, and the operator still wants to control who the DF is for a given EVI, the use of the Preference-based DF type can also provide the desired level of load balancing.

In this type of scenarios, the ES is configured with an administrative Preference value, but then a range of EVI/ISIDs can be defined to use the Highest-Preference or the Lowest-Preference depending on the desired behavior. With this option, the PE will build a list of candidate PEs ordered by the Preference, however the DF for a given EVI/ISID will be determined by the local configuration.

For instance:

- Assuming ES3 is defined in PE1 and PE2, PE1 may be configured as [500,0,Preference] for ES3 and PE2 as [100,0,Preference].

- In addition, assuming vlan-based service interfaces, the PEs will be configured with (vlan/ISID-range,high_or_low), e.g. (1-2000,high) and (2001-4000, low).

- This will result in PE1 being DF for EVI/ISIDs 1-2000 and PE2 being DF for EVI/ISIDs 2001-4000.

4.3 The Non-Revertive option

As discussed in section 2(d), an option to NOT preempt the existing

DF for a given EVI/ISID is required and therefore added to the DF Election extended community. This option will allow a non-revertive behavior in the DF election.

Note that, when a given PE in an ES is taken down for maintenance operations, before bringing it back, the Preference may be changed in order to provide a non-revertive behavior. The DP bit and the mechanism explained in this section will be used for those cases when a former DF comes back up without any controlled maintenance operation, and the non-revertive option is desired in order to avoid service impact.

In Figure 1, we assume that based on the Highest-Pref, PE3 is the DF for ESI2.

If PE3 has a link, EVC or node failure, PE2 would take over as DF. If/when PE3 comes back up again, PE3 will take over, causing some unnecessary packet loss in the ES.

The following procedure avoids preemption upon failure recovery (please refer to Figure 1):

1) A new "Don't Preempt Me" parameter is defined on a per-PE per-ES basis, as described in section 3. If "Don't Preempt Me" is disabled (default behavior) the advertised DP bit will be 0. If "Don't Preempt Me" is enabled, the ES route will be advertised with DP=1 ("Don't Preempt Me").

2) Assuming we want to avoid 'preemption', the three PEs are configured with the "Don't Preempt Me" option. Note that each PE individually MAY be configured with different preemption value. In this example, we assume ESI2 is configured as 'DP=enabled' in the three PEs.

3) Assuming EVI1 uses Highest-Pref in vES2 and EVI2 uses Lowest-Pref, when vES2 is enabled in the three PEs, the PEs will exchange the ES routes and select PE3 as DF for EVI1 (due to the Highest-Pref type), and PE1 as DF for EVI2 (due to the Lowest-Pref).

4) If PE3's vES2 goes down (due to EVC failure - detected by OAM, or port failure or node failure), PE2 will become the DF for EVI1. No changes will occur for EVI2.

5) When PE3's vES2 comes back up, PE3 will start a boot-timer (if booting up) or hold-timer (if the port or EVC recovers). That timer will allow some time for PE3 to receive the ES routes from PE1 and PE2. PE3 will then:

- Select two "reference-PEs" among the ES routes in the vES, the "Highest-PE" and the "Lowest-PE":
  - The Highest-PE is the PE with higher Preference, using the DP bit first (with DP=1 being better) and, after that, the lower PE-IP address as tie-breakers. PE3 will select PE2 as Highest-PE over PE1, since, when comparing [Pref,DP,PE-IP], [200,1,PE2-IP] wins over [100,1,PE1-IP].
  - The Lowest-PE is the PE with lower Preference, using the DP bit first (with DP=1 being better) and, after that, the lower PE-IP address as tie-breakers. PE3 will select PE1 as Lowest-PE over PE2, since [100,1,PE1-IP] wins over [200,1,PE2-IP].
  - Note that if there were only one remote PE in the ES, Lowest and Highest PE would be the same PE.
- Check its own administrative Pref and compares it with the one of the Highest-PE and Lowest-PE that have DP=1 in their ES routes. Depending on this comparison PE3 will send the ES route with a [Pref,DP] that may be different from its administrative [Pref,DP]:
  - If PE3's Pref value is higher than the Highest-PE's, PE3 will send the ES route with an 'in-use' operational Pref equal to the Highest-PE's and DP=0.
  - If PE3's Pref value is lower than the Lowest-PE's, PE3 will send the ES route with an 'in-use' operational Preference equal to the Lowest-PE's and DP=0.
  - If PE3's Pref value is neither higher nor lower than the Highest-PE's or the Lowest-PE's respectively, PE3 will send the ES route with its administrative [Pref,DP]=[300,1].
  - In this example, PE3's administrative Pref=300 is higher than the Highest-PE with DP=1, that is, PE2 (Pref=200). Hence PE3 will inherit PE2's preference and send the ES route with an operational 'in-use' [Pref,DP]=[200,0].

Note that, a PE will always send DP=0 as long as the advertised Pref is the 'in-use' operational Pref (as opposed to the 'administrative' Pref).

This ES route update sent by PE3 (with [200,0,PE3-IP]) will not cause any DF switchover for any EVI/ISID. PE2 will continue being DF for EVI1. This is because the DP bit will be used as a tie-

breaker in the DF election. That is, if a PE has two candidate PEs with the same Pref, it will pick up the one with DP=1. There are no DF changes for EVI2 either.

6) Subsequently, if PE2 fails, upon receiving PE2's ES route withdrawal, PE3 and PE1 will go through the process described in (5) to select new Highest and Lowest-PEs (considering their own active ES route) and then they will run the DF Election.

   o If a PE selects itself as new Highest or Lowest-PE and it was not before, the PE will then compare its operational 'in-use' Pref with its administrative Pref. If different, the PE will send an ES route update with its administrative Pref and DP values. In the example, PE3 will be the new Highest-PE, therefore it will send an ES route update with [Pref,DP]=[300,1].

   o After running the DF Election, PE3 will become the new DF for EVI1. No changes will occur for EVI2.

Note that, irrespective of the DP bit, when a PE or ES comes back and the PE advertises a DF Election type different than 2 (Preference algorithm), the rest of the PEs in the ES MUST fall back to the default RFC7432 service-carving modulo-based DF Election.

5. Conclusions

Service Providers are seeking for options where the DF election can be controlled by the user in a deterministic way and with a non-revertive behavior. This document defines the use of a Preference algorithm that can be configured and used in a flexible manner to achieve those objectives.

11. 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.

12. Security Considerations

This section will be added in future versions.

13. IANA Considerations

This document solicits the allocation of DF type = 2 in the registry created by [EVPN-HRW-DF] for the DF type field.

15. References

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

15.2 Informative References

[vES] Sajassi et al. "EVPN Virtual Ethernet Segment", draft-sajassi-bess-evpn-virtual-eth-segment-01, work-in-progress, July 6, 2015.

[EVPN-HRW-DF] Mohanty S. et al. "A new Designated Forwarder Election for the EVPN", draft-mohanty-bess-evpn-df-election-02, work-in-progress, October 19, 2015.

16. Acknowledgments

The authors would like to thank Kishore Tiruveedhula for his review and comments.

17. Contributors

In addition to the authors listed, the following individuals also contributed to this document:

Kiran Nagaraj, Nokia
Vinod Prabhu, Nokia
Selvakumar Sivaraj, Juniper

17. Authors' Addresses

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

Senthil Sathappan
Alcatel-Lucent
Email: senthil.sathappan@nokia.com

Tony Przygienda
Juniper Networks, Inc.
Email: prz@juniper.net

John Drake
Juniper Networks, Inc.
Email: jdrake@juniper.net

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

Ali Sajassi
Cisco Systems, Inc.
Email: sajassi@cisco.com

Satya Ranjan Mohanty
Cisco Systems, Inc.
Email: satyamoh@cisco.com

Sami Boutros
VMware, Inc.
Email: sboutros@vmware.com