BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2022

P. Brissette, Ed.
A. Sajassi
LA. Burdet
Cisco
J. Drake
Juniper
J. Rabadan
Nokia

July 6, 2021

Fast Recovery for EVPN DF Election
draft-ietf-bess-evpn-fast-df-recovery-02

Abstract

Ethernet Virtual Private Network (EVPN) solution provides Designated Forwarder election procedures for multi-homing Ethernet Segments. These procedures have been enhanced further by applying Highest Random Weight (HRW) Algorithm for Designated Forwarded election in order to avoid unnecessary DF status changes upon a failure. This draft improves these procedures by providing a fast Designated Forwarder (DF) election upon recovery of the failed link or node associated with the multi-homing Ethernet Segment. The solution is independent of number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each PEs in the multi-homing group.

Requirements Language

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] and RFC 8174 [RFC8174].

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 7, 2022.

Table of Contents

1. Introduction

Ethernet Virtual Private Network (EVPN) solution [RFC7432] is becoming pervasive in data center (DC) applications for Network Virtualization Overlay (NVO) and DC interconnect (DCI) services, and

in service provider (SP) applications for next generation virtual private LAN services.

EVPN solution [RFC7432] describes DF election procedures for multi-homing Ethernet Segments. These procedures are enhanced further in

[RFC8584] by applying Highest Random Weight Algorithm for DF election in order to avoid DF status change unnecessarily upon a link or node failure associated with the multi-homing Ethernet Segment. This draft makes further improvement to DF election procedures in [RFC8584] by providing an option for a fast DF election upon recovery of the failed link or node associated with the multi-homing Ethernet Segment. This DF election is achieved independent of number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each PE in the multi-homing group. The solution is based on simple one-way signaling mechanism.

1.1. Terminology

Provider Edge (PE): A device that sits in the boundary of Provider and Customer networks and performs encap/decap of data from L2 to L3 and vice-versa.

Designated Forwarder (DF): A PE that is currently forwarding (encapsulating/decapsulating) traffic for a given VLAN in and out of a site.

2. Challenges with Existing Solution

In EVPN technology, multiple PE devices have the ability to encap and decap data belonging to the same VLAN. In certain situations, this may cause L2 duplicates and even loops if there is a momentary overlap of forwarding roles between two or more PE devices, leading to broadcast storms.

EVPN [RFC7432] currently uses timer based synchronization among PE devices in redundancy group that can result in duplications (and even loops) because of multiple DFs if the timer is too short or blackholing if the timer is too long.

Using ESI label Split Horizon filtering can prevent loops (but not

duplicates), however if there are overlapping DFs in two different sites at the same time for the same VLAN, the site identifier will be different upon re-entry of the packet and hence the split horizon check will fail, leading to L2 loops.

The current state of art [RFC8584] uses the well known HRW (Highest Random Weight) algorithm to avoid reshuffling of VLANs among PE devices in the redundancy group upon failure/recovery and thus reducing the impact of failure/recovery to VLANs not on the failed/recovered ports. This eliminates loops/duplicates in failure scenarios.

However, upon PE insertion or port bring-up, HRW cannot help as a transfer of DF role need to happen to the newly inserted device/port while the old DF is still active.

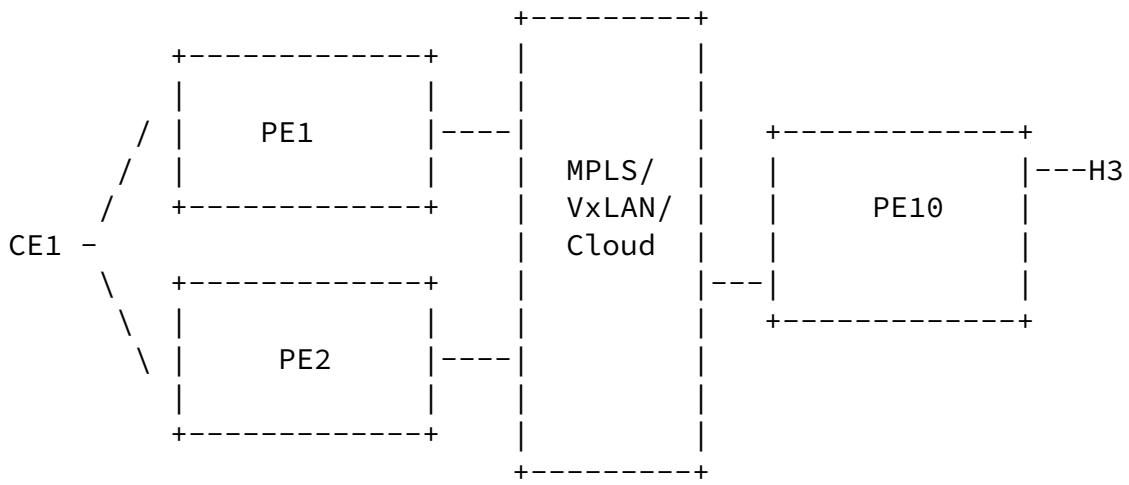


Figure 1: CE1 multi-homed to PE1 and PE2.

In the Figure 1, when PE2 is inserted or booted up, PE1 will transfer DF role of some VLANs to PE2 to achieve load balancing. However, because there is no handshake mechanism between PE1 and PE2, duplication of DF roles for a give VLAN is possible. Duplication of DF roles may eventually lead to L2 loops as well as duplication of traffic.

Current state of EVPN art relies on a blackholing timer for

transferring the DF role to the newly inserted device. This can cause the following issues:

* Loops/Duplicates if the timer value is too short

* Prolonged Traffic Blackholing if the timer value is too long

3. DF Election Synchronization Solution

The solution relies on the concept of common clock alignment between partner PEs participating to a common Ethernet-Segment. The main idea is to have them all to perform/apply their carving state, resulting from DF election, at the well-known time.

The DF Election procedure, as described in [RFC7432] and as optionally signalled in [RFC8584], is applied. All PEs attached to a given Ethernet-Segment are clock-synchronized; using a networking protocol for clock synchronization (e.g. NTP, PTP, etc.). Newly inserted device PE or during failure recovery of a PE, that PE

communicates the current time to peering partners plus the remaining peering timer time left. This constitute an "end" or "absolute" time as seen from local PE. That absolute time is called "Service Carving Time" (SCT).

A new BGP Extended Community is advertised along with Ethernet-Segment route (RT-4) to communicate to other partners the Service Carving Time.

Upon reception of that new BGP Extended Community, partner PEs know exactly its carving time. The notion of skew is introduced to eliminate any potential duplicate traffic or loops. They add a skew (default = -10ms) to the Service Carving Time to enforce this. The previously inserted PE(s) must carve first, followed shortly(skew) by the newly insterted PE.

To summarize, all peering PEs carve almost simultaneously at the time announced by newly added/recovered PE. The newly inserted PE initiates the SCT, and carves immediately on peering timer expiry. The previously inserted PE(s) receiving Ethernet-Segment route (RT-4) with a SCT BGP extended community, carve shortly before Service Carving Time.

3.1. Advantages

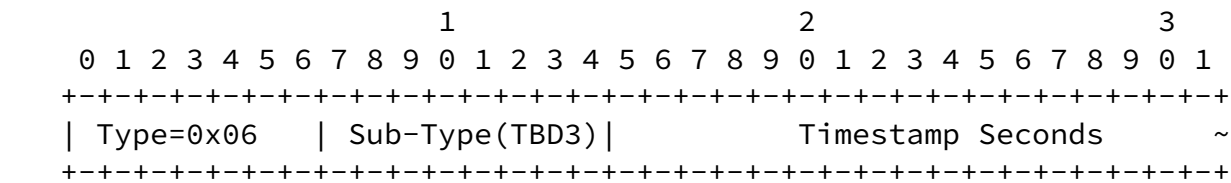
There are multiples advantages of using the approach. Here is a non-exhaustive list:

- A simple uni-directional signaling is all needed
- Backwards-compatible: PEs supporting only older [RFC7432] shall simply discard unrecognized new "Service Carving Timestamp" BGP Extended Community
- Multiple DF Election algorithms can be supported:
  * [RFC7432] default ordered list ordinal algorithm (Modulo),
  * [RFC8584] highest-random weight, etc.
- Independent of BGP transmission delay regarding Ethernet-Segment route (RT-4)
- Agnostic of the time synchronization mechanism used (e.g .NTP, PTP, etc.)

3.2. BGP Encoding

A new BGP extended community needs to be defined to communicate the Service Carving Timestamp for each Ethernet Segment.

A new transitive extended community where the Type field is 0x06, and the Sub-Type is [TBD3] is advertised along with Ethernet Segment route. Timestamp for expected Service carving is encoded as a 8-octet value as follows:



```
~  Timestamp Seconds          |  Timestamp Fractional Seconds   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

This document introduces a new flag called "T" (for Time Synchronization) to the bitmap field of the DF Election Extended Community defined in [RFC8584].

```
                      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 | |A| |T|       ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~     Bitmap    |            Reserved = 0                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

T: This flag is located in bit position 27 as shown above. When set to 1, it indicates the desire to use Time Synchronization capability with the rest of the PEs in the ES. This capability is used in conjunction with the agreed upon DF Type (DF Election Type). For example if all the PEs in the ES indicated that they have Time Synchronization capability and they want the DF type be of HRW, then HRW algorithm is used in conjunction with this capability.

3.3. Note on NTP-based synchronization

The 64-bit timestamp used by NTP protocol consists of a 32-bit part for seconds and a 32-bit part for fractional second. The timestamp exchanged uses the NTP epoch of January 1, 1900 [RFC5905]. The use of a 32-bit seconds and 16-bit fractional seconds yields adequate precision of 15 microseconds (2^-16 s).

3.4. Synchronization Scenarios

Let's take Figure 1 as an example where initially PE2 had failed and PE1 had taken over. This example shows the problem with known mechanism.

Based on [RFC7432]:

- Initial state: PE1 is in steady-state, PE2 is recovering

- PE2 recovers at (absolute) time t=99

- PE2 advertises RT-4 (sent at t=100) to partner PE1

- PE2, it starts its 3sec peering timer as per RFC7432

- PE1 carves immediately on RT-4 reception, i.e. t=100 + minimal BGP propagation delay

- PE2 carves at time t=103

[RFC7432] aims of favouring traffic black hole over duplicate traffic With above procedure, traffic black hole will occur as part of each PE recovery sequence. The peering timer value (default = 3 seconds) has a direct effect on the duration of the prolonged blackholing. A short (esp. zero) peering timer may, however, result in duplicate traffic or traffic loops.

Based on the Service Carving Time (SCT) approach:

- Initial state: PE1 is in steady-state, PE2 is recovering

- PE2 recovers at (absolute) time t=99

- PE2 advertises RT-4 (sent at t=100) with target SCT value t=103 to partner PE1

- PE2 starts its 3 second peering timer as per [RFC7432]

- Both PE1 and PE2 carves at (absolute) time t=103

In fact, PE1 should carve slightly before PE2 (skew). The previously inserted PE2 that is recovering performs both transitions DF to NDF and NDF to DF per VLANs at the peering timer expiry. Since the goal is to prevent duplicates, the original PE1, which received the SCT will apply:

- DF to NDF transition at t=SCT minus skew where both PEs are NDF for

'skew' amount of time

- NDF to DF transition at t=SCT

It is this split-behaviour which ensures good transition of DF role with contained amount of loss.

Using SCT approach, the negative effect of the peering timer is mitigated. Furthermore, the BGP Ethernet-Segment route (RT-4) transmission delay (from PE2 to PE1) becomes a no-op. The usage of SCT approach remedies to the exposed problem with the usage of peering timer. The 3 seconds timer window is shorthen to few milliseconds.

3.5. Backwards Compatibility

Per redundancy group, 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. It is, however, possible that some PEs continue to use the existing modulus based DF election and do not rely on the new SCT BGP extended community. PEs running an baseline DF election mechanism shall simply discard unrecognized new SCT BGP extended community.

A PE can indicate its willingness to support clock-synched carving by signaling the new 'T' DF Election Capability as well as including the new Service Carving Time BGP extended community along with the Ethernet-Segment Route (Type-4). In the case where one or more PEs attached to the Ethernet-Segment do not signal T=1, all PEs in the Ethernet-Segment may revert back to the RFC7432 timer approach.

4. Security Considerations

The mechanisms in this document use EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable. This document uses MPLS and IP-based tunnel technologies to support data plane transport. Security considerations described in [RFC7432] and in [RFC8365] are equally applicable.

5. IANA Considerations

This document solicits the allocation of the following sub-type in the "EVPN Extended Community Sub-Types" registry setup by [RFC7153]:

TBD3	Service Carving Timestamp	This document

This document solicits the allocation of the following values in the "DF Election Capabilities" registry setup by [RFC8584]:

Bit	Name	Reference
3	Time Synchronization	This document

6. Normative References

[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>.

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <https://www.rfc-editor.org/info/rfc5905>.

[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>.

[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>.

[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>.

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

[RFC8584] Rabadan, J., Ed., Mohanty, S., Ed., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for Ethernet VPN Designated Forwarder Election Extensibility", RFC 8584, DOI 10.17487/RFC8584, April 2019, <https://www.rfc-editor.org/info/rfc8584>.

Appendix A. Contributors

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

Gaurav Badoni
Cisco

Email: gbadoni@cisco.com

Dhananjaya Rao
Cisco

Email: dhrao@cisco.com

Appendix B. Acknowledgements

Authors would like to acknowledge helpful comments and contributions of Satya Mohanty and Bharath Vasudevan.

Authors' Addresses

Patrice Brissette (editor)
Cisco

Email: pbrisset@cisco.com

Ali Sajassi
Cisco

Email: sajassi@cisco.com

Luc Andre Burdet
Cisco

Email: lburdet@cisco.com

John Drake
Juniper

Email: jdrake@juniper.net

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com