

Workgroup: BESS Working Group
Internet-Draft:
draft-ietf-bess-evpn-fast-df-recovery-06
Published: 24 August 2022
Intended Status: Standards Track
Expires: 25 February 2023

A	P. Brissette, Ed.	A. Sajassi	LA. Burdet
	uCisco	Cisco	Cisco
	t		
	h		
	o		
	r		
	s		
	:		
	J. Drake	J. Rabadan	
	Juniper	Nokia	

Fast Recovery for EVPN Designated Forwarder Election

Abstract

The Ethernet Virtual Private Network (EVPN) solution provides Designated Forwarder election procedures for multihomed 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 document improves these procedures by providing a fast Designated Forwarder (DF) election upon recovery of the failed link or node associated with the multihomed Ethernet Segment. The solution is independent of the number of EVIs associated with that Ethernet Segment and it is performed via a simple signaling between the recovered PE and each of the other PEs in the multihoming 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 25 February 2023.

Copyright Notice

Copyright (c) 2022 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

Table of Contents

- [1. Introduction](#)
 - [1.1. Terminology](#)
 - [1.2. Challenges with Existing Solution](#)
 - [1.3. Advantages to Proposed Solution](#)
- [2. DF Election Synchronization Solution](#)
 - [2.1. BGP Encoding](#)
- [3. Synchronization Scenarios](#)
 - [3.1. Concurrent Recoveries](#)
- [4. Backwards Compatibility](#)
- [5. Security Considerations](#)
- [6. IANA Considerations](#)
- [7. Normative References](#)
- [Appendix A. Contributors](#)
- [Appendix B. Acknowledgements](#)
- [Authors' Addresses](#)

1. Introduction

The 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.

[RFC7432] describes DF election procedures for multihomed Ethernet Segments. These procedures are enhanced further in [RFC8584] by applying Highest Random Weight Algorithm for DF election in order to avoid unnecessary DF status changes upon a link or node failure associated with the multihomed Ethernet Segment. This document makes further improvements to the 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 multihomed 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 of the other PEs in the multihomed group. The solution is based on simple one-way signaling mechanism.

1.1. Terminology

Designated Forwarder (DF):

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

1.2. Challenges with Existing Solution

In EVPN technology, multiple PE devices have the ability to encapsulate and decapsulate 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 split-horizon filtering (Section 8.3 of [RFC7432]) 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 updated DF procedures in [RFC8584] use the well known Highest Random Weight (HRW) algorithm to avoid reshuffling of VLANs among PE devices in the redundancy group upon failure/recovery. This reduces the impact to VLANs not assigned to the failed/recovered ports and eliminates loops or duplicates at failure/recovery events.

However, upon PE insertion or a port being newly added to a multihomed Ethernet Segment, HRW also cannot help as a transfer of DF role to the new port must occur while the old DF is still active.

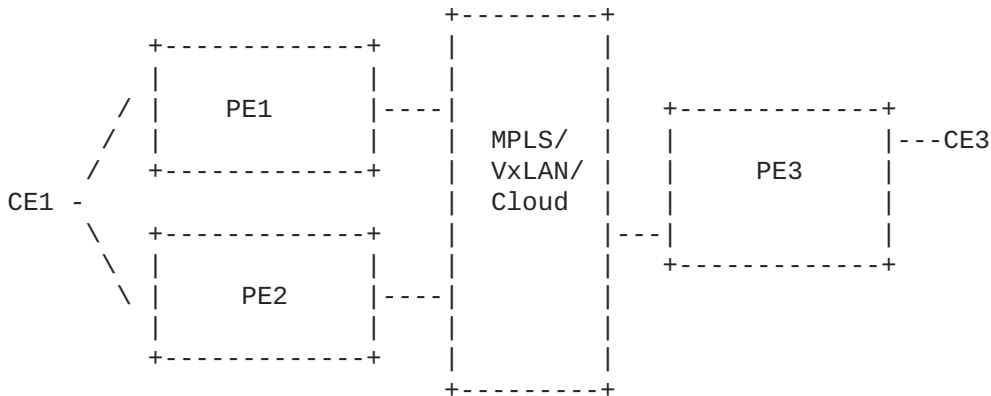


Figure 1: CE1 multihomed to PE1 and PE2.

In Figure 1, when PE2 is inserted or booted up, PE1 will transfer the 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 given VLAN is possible. Duplication of DF roles may eventually lead to duplication of traffic as well as L2 loops.

Current EVPN specifications [[RFC7432](#)] and [[RFC8584](#)] rely on a timer-based approach 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

1.3. Advantages to Proposed Solution

There are multiples advantages of using the proposed clock-synchronization approach, namely:

- *A simple uni-directional signaling is all that is needed, no complicated handshake or state machine.

- *Solution is backwards-compatible: PEs supporting only older [[RFC7432](#)] shall simply discard the unrecognized new "Service Carving Timestamp" BGP Extended Community

- *Many of the existing DF Election algorithms can be supported:

 - [RFC7432](#) default ordered list ordinal algorithm (Modulo),

 - [RFC8584](#) highest-random weight, etc.

- *Solution is independent of any BGP propagation delay of Ethernet Segment route (Route Type 4)

- *Solution is agnostic of the actual time synchronization mechanism used (e.g. NTP, PTP, etc.), while normalizing the exchange format in an NTP-based encoding.

2. DF Election Synchronization Solution

The solution relies on the concept of common clock alignment between partner PEs participating to a common Ethernet Segment i.e. PE1 and PE2 in [Figure 1](#). The main idea is to have all peering PEs of that Ethernet Segment perform DF election, and apply their resulting carving state, at a same pre-announced 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.). When a new PE is inserted or an existing PE device, that PE communicates the current time to peering partners plus the remaining peering timer time left. This constitutes an "end time" or "absolute time" as seen from local PE. That absolute time is called "Service Carving Time" (SCT).

A new BGP Extended Community, the Service Carving Timestamp 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 can determine exactly the anticipated carving time. The notion of skew is introduced to eliminate any potential duplicate traffic or loops.

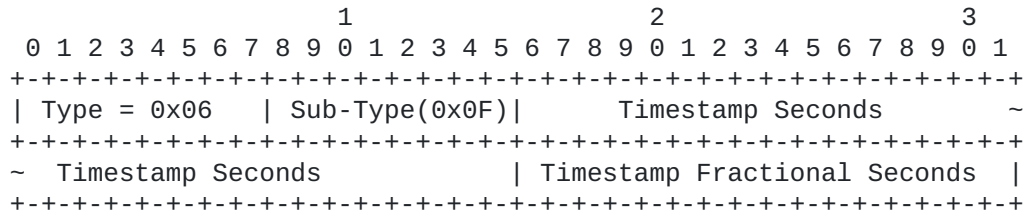
The receiving partner PEs add a skew (default = -10ms) to the Service Carving Time to enforce this. The previously inserted PE(s) must carve first, followed shortly (skew) by the newly insterted PE.

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.

2.1. 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 0x0F is advertised along with Ethernet Segment route. The expected Service Carving Time is encoded as a 8-octet value as follows:

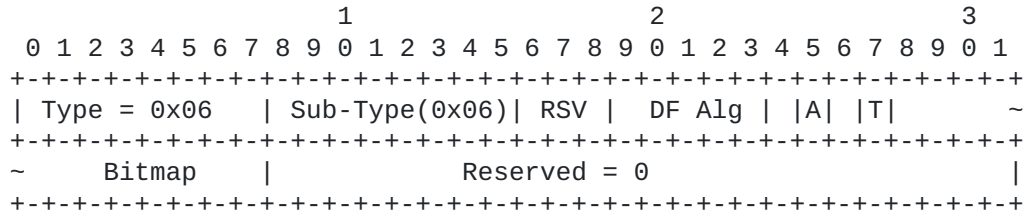


The timestamp exchanged uses the NTP epoch of January 1, 1900 [RFC5905]. The 64-bit timestamp of the NTP protocol consists of a 32-bit part for seconds and a 32-bit part for fractional second:

- *Timestamp Seconds: 32-bit NTP seconds are encoded in this field.

- *Timestamp Fractional Seconds: 16 bits of the NTP fractional seconds are encoded in this field. The use of a 16-bit fractional seconds yields adequate precision of 15 microseconds (2⁻¹⁶ s).

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



- *Bit 3: Time Synchronization (corresponds to Bit 27 of the DF Election Extended Community). When set to 1, it indicates the desire to use Time Synchronization capability with the rest of the PEs in the Ethernet Segment.

This capability is used in conjunction with the agreed upon DF Type (DF Election Type). For example if all the PEs in the Ethernet Segment indicate having Time Synchronization capability and are requesting the DF type to be HRW, then the HRW algorithm is used in conjunction with this capability.

3. 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 the DF-Election mechanism in [\[RFC7432\]](#).

Based on [Section 8.5](#) of [\[RFC7432\]](#), using the default 3 second peering timer:

1. Initial state: PE1 is in steady-state, PE2 is recovering
2. PE2 recovers at (absolute) time $t=99$
3. PE2 advertises RT-4 (sent at $t=100$) to partner PE1
4. PE2 starts a 3 second peering timer
5. PE1 carves immediately on RT-4 reception, i.e. $t=100$ + minimal BGP propagation delay
6. PE2 carves at time $t=103$

[\[RFC7432\]](#) aims of favouring traffic black hole over duplicate traffic. With above procedure, traffic black holing will occur as part of each PE recovery sequence since PE1 has transitioned some VLANs to Non-Designated-Forwarder (NDF) immediately upon reception. The peering timer value (default = 3 seconds) has a direct effect on the duration of the blackholing. A shorter (esp. zero) peering timer may, however, result in duplicate traffic or traffic loops.

Based on the Service Carving Time (SCT) approach:

1. Initial state: PE1 is in steady-state, PE2 is recovering
2. PE2 recovers at (absolute) time $t=99$
3. PE2 advertises RT-4 (sent at $t=100$) with target SCT value $t=103$ to partner PE1
4. PE2 starts 3 second peering timer
5. PE1 starts service carving timer, with remaining time until $t=103$
6. Both PE1 and PE2 carve at (absolute) time $t=103$

In fact, PE1 should carve slightly before PE2 (skew) to maintain the preference of minimal loss over duplicate traffic. 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 $t = \text{SCT} - \text{skew}$, where both PEs are NDF for 'skew' amount of time

*NDF to DF transition at $t = \text{SCT}$

It is this split-behaviour which ensures a 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 non-issue. The use of SCT approach remedies the problem associated with the peering timer: the 3 second timer window is shortened to the order of milliseconds.

3.1. Concurrent Recoveries

In the eventuality 2 or more PEs in a peering Ethernet Segment group are recovering concurrently or roughly the same time, each will advertise a Service Carving Timestamp. This SCT value would correspond to what each recovering PE considers the "end time" for DF Election. A similar situation arises in staggered recovering PEs, when a second PE recovers at roughly a first PE's advertised SCT expiry, and with its own new SCT-2 outside of the initial SCT window.

In the case of multiple outstanding DF elections, one requested by each of the recovering PEs, the SCTs must simply be time-ordered and all PEs execute only a single DF Election at the service carving time corresponding to the largest received timestamp value. The DF Election will involve all the active PEs in a single DF Election update.

Example:

1. Initial state: PE1 is in steady-state, all services elected at PE1.
2. PE2 recovers at time $t=100$, advertises RT-4 with target SCT value $t=103$ to partners (PE1)
3. PE2 starts 3 second peering timer
4. PE1 starts service carving timer, with remaining time until $t=103$
5. PE3 recovers at time $t=102$, advertises RT-4 with target SCT value $t=105$ to partners (PE1, PE2)
6. PE3 starts 3 second peering timer
7. PE2 cancels peering timer, starts service carving timer with remaining time until $t=105$
8. PE1 updates service carving timer, with remaining time until $t=105$

9. PE1, PE2 and PE3 carve at (absolute) time t=105

4. 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 modulo-based DF election and do not rely on the new SCT BGP extended community. PEs running a baseline DF election mechanism will simply discard the new SCT BGP extended community as unrecognized.

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 SHALL revert back to the [RFC7432] timer approach. This is especially important in the context of the VLAN shuffling with more than 2 PEs.

5. 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.

6. IANA Considerations

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

0x0F	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

7. 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. Also thank you to Anoop Ghanwani for his thorough review with valuable comments and corrections.

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