Workgroup: BESS Working Group
Internet-Draft:
draft-ietf-bess-evpn-fast-df-recovery-08
Updates: 8584 (if approved)
Published: 10 July 2023
Intended Status: Standards Track
Expires: 11 January 2024
Authors: P. Brissette, Ed. A. Sajassi LA. Burdet
Cisco Cisco Cisco
J. Drake J. Rabadan
Juniper Nokia

Fast Recovery for EVPN Designated Forwarder Election

Abstract

The Ethernet Virtual Private Network (EVPN) solution provides Designated Forwarder (DF) election procedures for multihomed Ethernet Segments. These procedures have been enhanced further by applying Highest Random Weight (HRW) algorithm for Designated Forwarder election in order to avoid unnecessary DF status changes upon a failure. This document improves these procedures by providing a fast Designated Forwarder election upon recovery of the failed link or node associated with the multihomed Ethernet Segment. This document updates Section 2.1 of [RFC8584] by optionally introducing delays between some of the events therein.

The solution is independent of the number of EVPN Instances (EVIs) associated with that Ethernet Segment and it is performed via a simple signaling between the recovered node and each of the other nodes in the multihoming group.

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 11 January 2024.

Table of Contents

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 inte)rconnect (DCI) services, and in service provider (SP) applications for next generation virtual private LAN services.

[RFC7432] describes Designated Frowarder (DF) election procedures for multihomed Ethernet Segments. These procedures are enhanced further in [RFC8584] by applying the Highest Random Weight (HRW) 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 the number of EVPN Instances (EVIs) associated with that Ethernet Segment and it is performed via simple signaling between the recovered node and each of the other nodes in the multihomed group. This document updates the state machine described in Section 2.1 of [RFC8584], by optionally introducing delays between some events, as further detailed in Section 2.2. The solution is based on a simple one-way signaling mechanism.

1.1. Requirements Language

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.2. Terminology

PE: Provider Edge device.

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

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

1.3. Challenges with Existing Mechanism

In EVPN technology, multiple Provider Edge (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 a redundancy group that can result in duplications (and even loops) because of multiple DFs if the timer is too short or packets being dropped 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 the packet re-entering the Ethernet Segment 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.

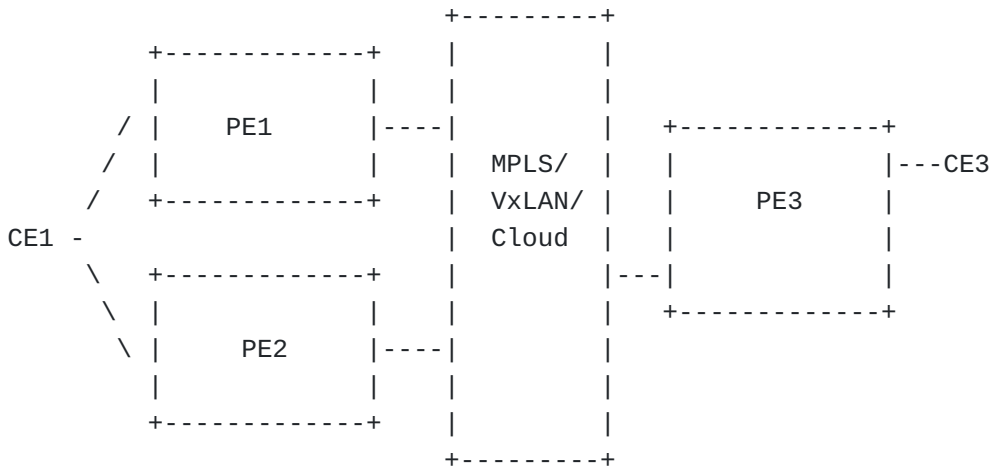


Figure 1: CE1 multihomed to PE1 and PE2.

In Figure 1, when PE2 is inserted in the Ethernet Segment or its CE1-facing interface recovered, 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

1.4. Design Principles for a Solution

The clock-synchronization solution presented in this document follows several design principles and presents multiples advantages, namely:

*Complicated handshake signamling mechanisms and state machines are avoided in favor of a simple uni-directional signaling approach.

*The solution is backwards-compatible (see Section 4), by PEs simply discarding the unrecognized new BGP Extended Community.

*Existing DF Election algorithms are supported.

*The solution is independent of any BGP delays in propagation of Ethernet Segment routes (Route Type 4)

*The solution is agnostic of the actual time synchronization mechanism used.

2. DF Election Synchronization Solution

The solution relies on the concept of common clock alignment between partner PEs participating in 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 the result at the 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). When a new PE is inserted in an Ethernet Segment or a failed PE device of the Ethernet Segment recovers, that PE communicates to peering partners the current time plus the value of the timer for partner discovery from step 2 in Section 8.5 of [RFC7432]. This constitutes an "end time" or "absolute time" as seen from the local PE. That absolute time is called the "Service Carving Time" (SCT).

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

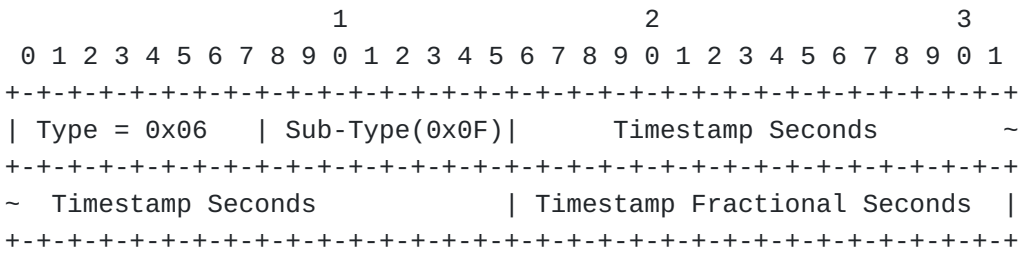
Upon receipt of that new BGP Extended Community, partner PEs can determine the service carving time of the newly insterted PE. The notion of skew is introduced to eliminate any potential duplicate traffic or loops. The receiving partner PEs add a skew (default = -10ms) to the Service Carving Time to 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 the newly added/recovered PE. The newly inserted PE initiates the SCT, and carves immediately on its local 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 is 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 the Ethernet Segment route. The expected Service Carving Time is encoded as an 8-octet value as follows:

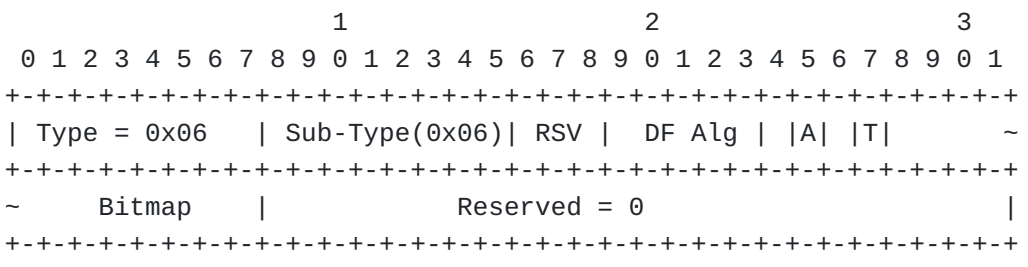


The timestamp exchanged uses the NTP epoch of January 1, 1900 [RFC5905]. As the current NTP era value is not exchanged, a local clock which is "synchronized" but to the wrong era is outside of the scope of this document.
The 64-bit timestamp of NTP consists of a 32-bit part for seconds and a 32-bit part for fractional second:

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

- *Timestamp Fractional Seconds: the high order 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.

2.2. Updates to RFC8584

This document introduces an additional delay to the events and transitions defined for the default DF election algorithm FSM in Section 2.1 of [RFC8584] without changing the FSM states or events itself.

The peering PE's FSM in DF_DONE which receives a RECV_ES transitions to DF_CALC. Because of the SCT carried in the Ethernet-Segment update, the output of the DF_CALC and transition back into DF_DONE are delayed, as are accompanying forwarding updates to DF/NDF state.

The corresponding actions when transitions are performed or states are entered/exited is modified as follows:

9. DF_CALC on CALCULATED: Mark the election result for the VLAN or Bundle.

   9.1 Where SCT timestamp is present on the RECV_ES event of Action 11, wait until the time indicated by the SCT before continuing to 9.2.

   9.2 Assume a DF/NDF for the local PE for the VLAN or VLAN Bundle, and transition to DF_DONE.

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 Section 8.5 of [RFC7432], using the value of the timer configured for all PEs on the Ethernet Segment.

Based on Section 8.5 of [RFC7432] and using the default 3 second timer in step 2:

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 timer to allow the reception of RT-4 from other PE nodes

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 being dropped over duplicate traffic. With the above procedure, traffic drops will occur as part of each PE recovery sequence since PE1 has transitioned some VLANs to Non-Designated-Forwarder (NDF) immediately upon reception.
The timer value (default = 3 seconds) has a direct effect on the duration of the packets being dropped. A shorter (especially zero) 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 a 3 second timer to allow the reception of RT-4 from other PE nodes

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 timer's 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 a good transition of DF role with contained amount of loss.

Using SCT approach, the negative effect of the timer to allow the reception of RT-4 from other PE nodes 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 this 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 rougly 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 a 3 second timer to allow the reception of RT-4 from other PE nodes

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 a 3 second timer to allow the reception of RT-4 from other PE nodes

7. PE2 cancels the running 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

In the eventuality a PE in a Ethernet Segment group recovers during the discovery window specified in Section 8.5 of [RFC7432], and does

not support or advertise the T-bit, then all PEs in the current peering sequence SHALL immediately revert to the default [RFC7432] behavior.

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.

For the new SCT Extended Community, attack vectors may be setting the value to zero, to a value in the past or to large times in the future. The procedures in this document address implicitly what occurs with a carving time in the past, as this would be a naturally occurring event with a large BGP propagation delay: the receiving PE SHALL treat the DF Election at the peer as having occurred already, and proceed without starting any carving delay timer. For timestamp values in the future, a rogue PE may be advertising a value inconsistent with its local behaviour. This is no different than a rogue PE setting all its DF Election results inconstently to its peers using (or ignoring adherence to) the procedures from [RFC7432], and the result would similarly be duplicate or dropped traffic.

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

IANA maintains the "EVPN Extended Community Sub-Types" registry set up by [RFC7153]. IANA is requested to confirm the First Come First Served assignment as follows:

Sub-Type Value	Name	Reference	Date
0x0F	Service Carving Timestamp	This document	TBD

IANA should replace the field TBD with the date of publicaton of this document as an RFC.

IANA maintains the "DF Election Capabilities" registry set up by [RFC8584]. IANA is requested to make the following assignment from this registry:

Bit	Name	Reference	Date
3	Time Synchronization	This document	TBD

IANA should replace the field TBD with the date of publicaton of this document as an RFC.

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