Workgroup: BESS Working Group
Internet-Draft:
draft-burdet-bess-evpn-fast-reroute-00
Published: 25 October 2021
Intended Status: Standards Track
Expires: 28 April 2022
Authors: LA.B. Burdet, Ed. P.B. Brissette T.M. Miyasaka
Cisco Cisco KDDI Corporation

EVPN Fast Reroute

Abstract

This document summarises EVPN convergence mechanisms and specifies procedures for EVPN networks to achieve sub-second and scale-independant convergence.

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 28 April 2022.

Copyright Notice

Table of Contents

1. Introduction

EVPN convergence and failure recovery methods from different types of network failures is described in [RFC7432] Section 17. Similarly for EVPN-VPWS, [RFC8214] briefly evokes an egress link protection mechanism at the end of Section 5.

The fundamentals of EVPN convergence rely on a mass-withdraw technique of the Ethernet A-D per ES route to unresolve all the associated forwarding paths ([RFC7432] Section 9.2.2 'Route Resolution'). The mass-withdraw grouping approach results in suitable EVPN convergence at lower scale, but is not sufficent to meet stricter sub-second requirements. Other control-plane enhancements such as route-prioritisation ([I-D.ietf-bess-rfc7432bis]) help further but still provide no guarantees.

EVPN convergence using only control-plane approaches is constrained by BGP route propagation delays, routes processing times in software and hardware programming. These are additionally often performed sequentially and linearly given the potential large scale of EVPN routes present in control plane.

This document presents a mechanism for fast reroute to minimise packet loss in the case of a link failure using EVPN redirect labels (ERLs) with special forwarding attributes. Multiple-failures where loops may occur are addressed, as are cascading failures. A mechanism for distributing redirect labels (ERLs) alongside EVPN service labels (ESLs) is shown.

The main objective is to achieve sub-second convergence in EVPN networks without relying on control plane actions. The procedures in this document apply equally to EVPN services (EVPN [RFC7432], EVPN-VPWS [RFC8214] and EVPN-IRB [RFC9135]), and all Ethernet-Segment load-balancing modes.

2. Specification of Requirements

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 [RFC2119].

3. Terminology

Some of the terminology in this document is borrowed from [RFC8679] for consistency across fast reroute frameworks.

CE: Customer Edge device, e.g., a host, router, or switch.

PE: Provider Edge device.

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

Egress link: Specific Ethernet link connecting a given PE-CE, which forms part of an Ethernet Segment.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

DF-Election:
Designated Forwarder election, as in [RFC7432] and [RFC8584].

DF: Designated Forwarder.

Backup-DF (BDF): Backup-Designated Forwarder.

Non-DF (NDF): Non-Designated Forwarder.

AC: Attachment Circuit.

ERL: Special-use EVPN redirect label, described in this document.

ESL: EVPN service label, as in [RFC7432], [RFC8214] and [RFC9135].

4. Requirements

1. EVPN multihoming is often described as 2 peering PEs. The solution MUST be generic enough to apply multiple peering PE and no artificial limit imposed on the number of peering PEs.

2. The solution MUST apply to all EVPN load-balancing modes.

3. The solution MUST be robust enough to tolerate failures of the same ES at multiple PEs. Simultaneous as well as cascading failures on the same ES must be addressed.

4. The solution MUST support EVPN [RFC7432], EVPN-VPWS [RFC8214] and EVPN-IRB [RFC9135] services.

5. The solution MUST meet stringent sub-second and often 50 millisecond requirements for traffic loss of EVPN services.

6. The solution MUST allow redirected-traffic to bypass port blocking states resulting from DF-Election (BDF or NDF).

7. The solution MUST be scale-independant and agnostic of EVPN route types, scale or choice of underlay.

8. The solution MUST address egress link (PE-CE link) failures.

9. The solution MUST be loop-free, and once-redirected traffic MUST never be repeatedly redirected.

10. The solution MUST not rely on pushing an additional label onto the label stack.

11. The solution SHOULD address Broadcast, unknown unicast and multicast (BUM) traffic.

5. Solution

Sub-second convergence in EVPN networks is achieved using a combined approach to minimising traffic loss:

- *Local failure detection and restoration of traffic flows in minimal time using a pre-computed redirect path ;

- *Restoration of optimal traffic paths, and reconvergence of EVPN control plane with EVPN mass withdraw.

The solution presented in this document addresses the local failure detection and restoration, without impeding on or impacting existing EVPN control plane convergence mechanisms.

Consider the following EVPN topology where PE1 and PE2 are multihoming PEs on a shared ES, ESI1. EVPN (known unicast) or EVPN-VPWS traffic from CE1 to CE2 is sent to PE1 and PE2 using EVPN service labels ESL1 and/or ESL2 (depending on load-balancing mode of the ESI1 interfaces).

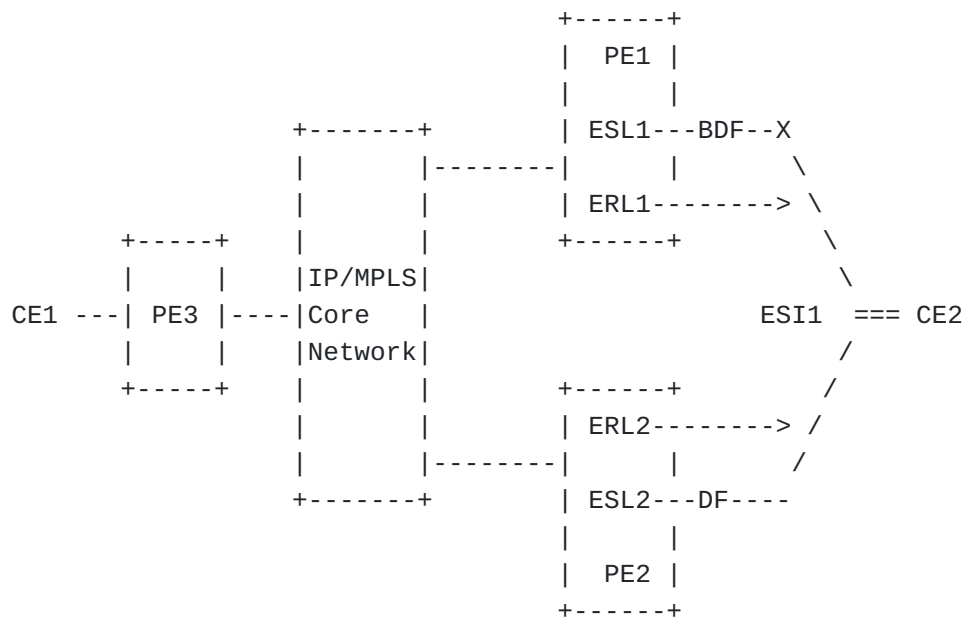

Figure 1

EVPN Multihoming with service and redirect labels

Alongside the service labels ESL1 and ESL2, two redirect labels ERL1 and ERL2 are allocated with special forwarding attributes, as detailed in Section 6. Fast-reroute and use of the ERLs is shown in Section 5.2

5.1. Pre-selection of Backup Path

EVPN DF-Election lends itself well to the selection of a pre-computed path amongst any given number of peering PEs by providing a DF-Elected and BDF-Elected node at the <EVI, ESI> granularity ([RFC8584] and [I-D.ietf-bess-rfc7432bis]).

In All-active mode, all PEs in the Ethernet Segment are actively forwarding known unicast traffic to the CE. In Single-active mode, only a single PE in the Ethernet Segment is actively forwarding known unicast traffic to the CE: the DF-Elected PE. The BDF-Elected PE is next to be elected in the redundancy group and is already known.

For consistency across PEs and load-balancing modes, the backup path selected should be in order of {DF, BDF, NDF1, NDF2, ...}. The DF-Elected PE selects the next-best BDF-Elected as backup and all BDF- and NDF-Elected nodes select the best DF-Elected for the protection of their egress links.

*PE1 (DF) -> ERL(PE2),

*PE2 (BDF) -> ERL(PE1),

*PE..n (NDF) -> ERL(PE1),

The number of peering PEs is not limited by existing DF-Election algorithms. A solution based on DF-Election supports subsequent redirection upon multiple cascading failures, once a new DF-Election has occurred. Pre-selection of a backup path is supported by all current DF-Election algorithms, and more generally by all algorithms supporting BDF-Election, as recommended in ([I-D.ietf-bess-rfc7432bis]).

5.2. Failure Detection and Traffic Restoration

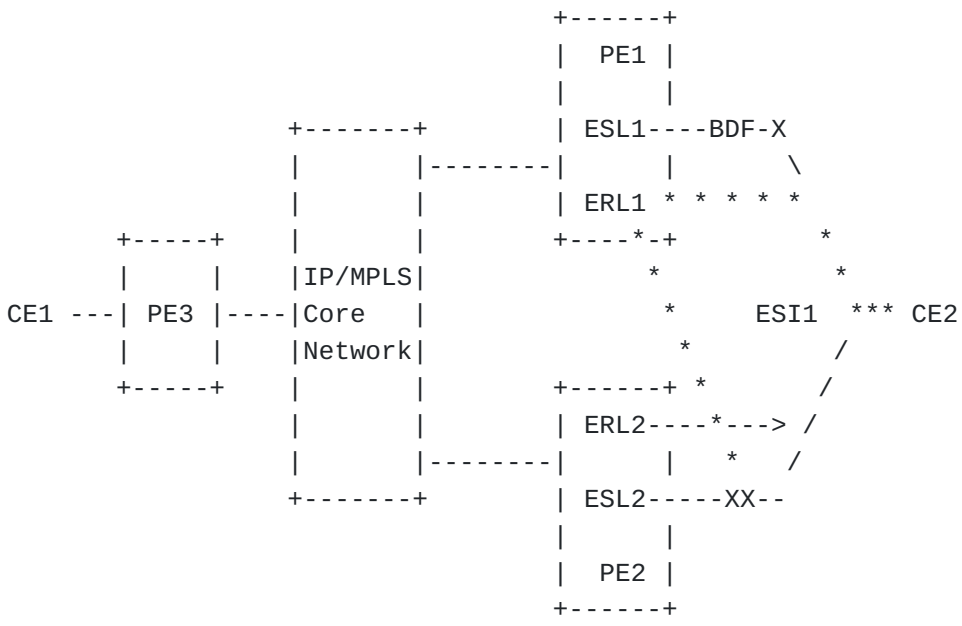

Figure 2

EVPN Multihoming failure scenario

The procedures for forwarding known unicast packets received from a remote PE on the local redirect label largely follow [RFC7432] Section 13.2.2.

Consider the EVPN multihoming topology in Figure 1, and a traffic flow from CE1 to CE2 which is currently using EVPN service label ESL2 and forwarded through the core arriving at PE2. When the local AC representing the <EVI,ESI> pair is protected using the fast-reroute solution, the pre-computed backup path's redirect label (i.e. ERL1 from BDF-Elected PE1) is installed against the AC.

Under normal conditions, PE2 disposition using ESL2 will result in forwarding the packet to the CE by selecting the local AC associated with the EVPN service label (EVPN-VPWS) or MAC address lookup (EVPN). When this local AC is in failed state, the fast-reroute solution at PE2 will begin rerouting packets using the BDF-Elected peer's nexthop and ERL1. ERL1 is chosen for redirection and not ESL1 for the redirected traffic to prevent loops and overcome DF-Election timing as described in Sections 6.2 and 6.1 respectively.

5.2.1. Simultaneous Failures in ES

In EVPN multihoming where the CE connects to peering PEs through link aggregation (LAG), a single LAG failure at the CE may manifest as multiple ES failures at all peering PEs simultaneously. As all peering PEs would enable simultaneously the fast-reroute mechanism,

redirection would be permanent causing a traffic storm or until TTL expires.

Once-redirected traffic may not be redirected again, according to the terminal nature of ERLs described in Section 6.2

5.2.2. Successive and Cascading Failures in ES

Trying to support cascading failures by redirecting once-redirected traffic is substantially equivalent to simultaneous failures above.

Once-redirected traffic may not be redirected again, according to the terminal nature of ERLs described in Section 6.2 and loss is to be expected until EVPN control plane reconverges for double-failure scenarios.

In a scenario with 3 peering PEs (PE1-DF, PE2-BDF, PE3-NDF) where PE1 fails, followed by a PE2 failure before control-plane reconvergence, there is no reroute of traffic towards PE3 because the reroute-label is terminal.

In such rapid-succession failures, it is expected that control plane must first correct for the initial failure and DF-Elect PE2 as new-DF and PE3 as the new-BDF. PE2 to PE3 redirection would then begin, unless control-plane is rapid enough to correct directly, and elect PE3 new-DF.

6. Redirect Labels: Forwarding Attributes

The EVPN redirect labels MUST be downstream assigned, and it is directly associated with the <EVI,ESI> AC being egress protected. The special forwarding characteristics and use of an EVPN redirect label (ERL) described below, are a matter of local significance only to the advertising PE (which is also the disposition PE).

Special-attributes to the ERLs do not affect any other PEs or transit P nodes. There are no extra labels appended to the label stack in the IP/MPLS network and the ERL appears to label-switching transit nodes as would any other EVPN service label.

* Traffic redirection and use of reroute labels may create routing loops upon multiple failures. Such loops are detrimental to the network and may cause congestion between protected PEs.

* Local restoration and redirection is meant to occur much faster than control-plane operations, meaning redirected packets may arrive at the BDF PE long before a DF-Election operation unblocks the egress link.

Two special forwarding characteristics of EVPN redirect labels are described below to mitigate these issues.

6.1. Bypassing DF-Election Attribute

Local detection and restoration at PE2 will begin rapidly redirecting traffic onto the backup path. Redirected packets will arrive at the Backup-DF port much faster than control plane DF-Election at the Backup-DF peer is capable of unblocking its local egress link for the shared ES (ESI1). All redirected traffic would drop at Backup-DF and no net reduction in traffic loss achieved.

Traffic restoration remains dependant upon ES route or Ethernet A-D per ES routes withdrawal for a DF-Election operation and for PE1 to assume the traffic forwarding role. This is especially important in single-active load-balancing mode where known unicast traffic is blocked.

To mitigate this, the redirect labels allocated must carry a special attribute in the local forwarding and decapsulation chain: for traffic received on the ERL when the AC is up, an override to the DF-Election is applied and traffic from the ERL will bypass the local Backup-DF blocking state. Once EVPN control plane reconverges, traffic from the ERL will cease and the optimal forwarding path based on ESLs will resume.

The EVPN redirect label MUST carry a context locally, such that from disposition to egress redirected packets are allowed to bypass the BDF blocking state that would otherwise drop. Similarly, this may open the gate to the traffic in the reverse direction.

6.2. Terminal Disposition Attribute

The reroute scheme is susceptible to loops and persistant redirects between peering PEs which have setup FRR redirection. Consider the scenario where both CE-facing interfaces fail simultaneously, fast reroute will be activated at both PE1 and PE2 effectively bouncing a redirected packet between the two PEs indefinitely (or until the TTL expires) causing a traffic storm.

To prevent this, a distinction is made between 'regular' EVPN service labels for disposition (i.e. known unicast EVI label or EVPN-VPWS label) and reroute labels with terminal disposition.

At the redirecting PE2, we consider the case of ESL2 vs. ERL2 , where both are locally allocated and provided in EVPN routes (downstream allocation) to BGP peers:

1. EVPN Service label, ESL2:

    *Regular MAC-lookup or traffic forwarding occurs towards the access AC.

    *If the AC is up, traffic will exit the interface, subject to local blocking state on the AC from DF-Election.

    *If the AC is down and fast-reroute procedures are enabled, traffic may be re-encapsulated using BDF peer's redirect label ERL1 (if received).

2. EVPN Reroute label, ERL2:

    *Regular MAC-lookup or traffic forwarding occurs towards the access AC.

    *If the AC is up, traffic will apply an override to DF-Election and bypass the local blocking state on the AC.

    *If the AC is down, traffic is dropped. No reroute must occur of once-rerouted traffic. Redirecting towards peer's redirect label ERL1 is explicitly prevented.

The ERL acts like a local cross-connect by providing a direct channel from disposition to the AC. ERLs are terminal-disposition and prevents once-redirected packets from being redirected again. With this forwarding attribute on ERLs, known only locally to the downstream-allocating PE, redirection is achieved without growing the label stack with another special purpose label.

6.3. Broadcast, Unknown Unicast and Multicast

BUM traffic is treated using EVPN defaults. There is no further extension to exiting procedure as of now, this work is left for future study.

7. Controlled Recovery Sequence

Fast reroute mechanisms such as the one described in this document generally provide a way to preserve traffic flows at failure time. Use of fast reroute in EVPN, however, permits setting up a controlled recovery sequence to shorten the period of loss between an interface coming up and the EVPN DF-Election procedures and default timers for peer discovery.

The benefit of a controlled recovery sequence is amplified when used in conjunction with [I-D.ietf-bess-evpn-fast-df-recovery] (synchronised DF-Election)>

8. Transport Underlay

The solution is agnostic to transport underlays, for instance similar behaviour is carried forward for VXLAN and SRv6

9. BGP Extensions

There are no new BGP extensions required to advertise the redirect label(s) used for EVPN egress link protection. The ESI Label Extended Community defined in [RFC7432] Section 7.5 may be advertised along with Ethernet A-D routes:

- When advertised with an Ethernet A-D per ES route, it enables split-horizon procedures for multihomed sites as described in [RFC7432] Section 8.3 ;

- When advertised with an Ethernet A-D per EVI route, it enables link protection and fast-reroute procedures for multihomed sites as described in this document. The label value represents the per-<EVI,ESI> EVPN redirect label (ERL). The Flags field SHOULD NOT be set and MUST be ignored.

Remote PEs SHALL NOT use the ERLs as a substitution for ESLs in route resolution, and is especially not to be confused with the aliasing and backup path ESL as described and used in [RFC7432] Section 8.4.

10. Security Considerations

The mechanisms in this document use the EVPN control plane as defined in [RFC7432] and [RFC8214], and the security considerations described therein are equally applicable. Reroute labels redistributed in EVPN control plane are meant for consumption by the peering PE in a same ES. It is, however, visible in the EVPN control plane to remote peers. Care shall be taken when installing reroute labels, since their use may result in bypassing DF-Election procedures and lead to duplicate traffic at CEs if incorrectly installed.

11. IANA Considerations

This document makes no specific requests to IANA.

12. References

12.1. 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>.

[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>.

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J. Rabadan, "Virtual Private Wire Service Support in Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August 2017, <https://www.rfc-editor.org/info/rfc8214>.

[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>.

12.2. Informative References

[I-D.ietf-bess-evpn-fast-df-recovery] Brissette, P., Sajassi, A., Burdet, L., Drake, J., and J. Rabadan, "Fast Recovery for EVPN DF Election", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-fast-df-recovery-02, 6 July 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-fast-df-recovery-02.txt>.

[I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", Work in Progress, Internet-Draft, draft-ietf-bess-rfc7432bis-01, 12 July 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-rfc7432bis-01.txt>.

[RFC8679] Shen, Y., Jeganathan, M., Decraene, B., Gredler, H., Michel, C., and H. Chen, "MPLS Egress Protection Framework", RFC 8679, DOI 10.17487/RFC8679, December 2019, <https://www.rfc-editor.org/info/rfc8679>.

[RFC9135] Sajassi, A., Salam, S., Thoria, S., Drake, J., and J. Rabadan, "Integrated Routing and Bridging in Ethernet VPN (EVPN)", RFC 9135, DOI 10.17487/RFC9135, October 2021, <https://www.rfc-editor.org/info/rfc9135>.

Appendix A. Acknowledgments

Appendix B. Contributors

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

Authors' Addresses

Luc Andre Burdet (editor)
Cisco

Email: lburdet@cisco.com

Patrice Brissette
Cisco

Email: pbrisset@cisco.com

Takuya
KDDI Corporation

Email: ta-miyasaka@kddi.com