BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2022

J. Rabadan, Ed.
S. Sathappan
Nokia

October 25, 2021

# Domain Path (D-PATH) for Ethernet VPN (EVPN) Interconnect Networks
## draft-sr-bess-evpn-dpath-00

Abstract

The BGP Domain PATH (D-PATH) attribute is defined for Inter-Subnet Forwarding (ISF) BGP Sub-Address Families that advertise IP prefixes. When used along with EVPN IP Prefix routes or IP-VPN routes, it identifies the domain(s) through which the routes have passed and that information can be used by the receiver BGP speakers to detect routing loops or influence the BGP best path selection. This document extends the use of D-PATH so that it can also be used along with EVPN MAC/IP Advertisement routes in EVPN Broadcast Domains (BD) and Auto-Discovery per EVPN Instance routes in Virtual Private Wire Services (VPWS).

Table of Contents



## 1. Introduction and Problem Statement

The BGP Domain PATH (D-PATH) attribute [I-D.ietf-bess-evpn-ipvpn-interworking] is defined for Inter-Subnet Forwarding (ISF) BGP Sub-Address Families that advertise IP prefixes. When used along with EVPN IP Prefix routes or IP-VPN routes, it identifies the domain(s) through which the routes have passed and that information can be used by the receiver BGP speakers to detect routing loops or influence the BGP best path selection. This document extends the use of D-PATH so that it can also be used along with EVPN MAC/IP Advertisement routes in EVPN Broadcast Domains (BD) [I-D.ietf-bess-rfc7432bis] and Auto-Discovery per EVPN Instance routes in Virtual Private Wire Services (VPWS) [RFC8214].

## 2. Conventions used in this document

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.

## 3. Terminology

This section summarizes the terminology that is used throughout the rest of the document.

- MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [I-D.ietf-bess-rfc7432bis]. It is also the instantiation of an EVI (EVPN Instance) in a PE.

- BD and BT: a Broadcast Domain and Bridge Table, as defined in [I-D.ietf-bess-rfc7432bis]. A BD is a group of PEs attached to the same EVPN layer-2 multipoint service. A BT is the instantiation of a Broadcast Domain in a PE. When there is a single Broadcast Domain in a given EVI, the MAC-VRF in each PE will contain a single BT. When there are multiple BTs within the same MAC-VRF, each BT is associated to a different Ethernet Tag. The EVPN routes specific to a BT, will indicate which Ethernet Tag the route corresponds to.

- AC: Attachment Circuit or logical interface associated to a given BT. To determine the AC on which a packet arrived, the PE will examine the combination of a physical port and VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).

- RT-2: Route Type 2 or MAC/IP Advertisement route, as per [I-D.ietf-bess-rfc7432bis].

- RT-1: Route Type 1 or Ethernet Auto-Discovery route, as per [I-D.ietf-bess-rfc7432bis].

- ES and ESI: Ethernet Segment and Ethernet Segment Identifier, as defined in [I-D.ietf-bess-rfc7432bis].

- TS: Tenant System.

- EVPN Layer2-Domain: two PEs are in the same Layer2-Domain if they are attached to the same tenant and the packets between them do not require a data path MAC lookup (in the BT of a MAC-VRF) in any intermediate router. A Layer2-Domain Gateway PE is always configured with multiple Layer2-Domain identifiers (Layer2-Domain-ID) in the MAC-VRF that connects those Layer2-Domains, each Layer2-Domain-ID representing a Layer2-Domain.

  Example: Figure 1 illustrates an example where PE1 and PE2 belong to different Layer2-Domains since packets between them (for flows

between TS1 and TS2) require a MAC lookup in two of the gateways that are connecting the three EVPN Layer2-Domains. E.g., if frames from TS1 to TS2 go through PE1, GW1, GW3 and PE2, a MAC lookup is performed at GW1 and GW3.

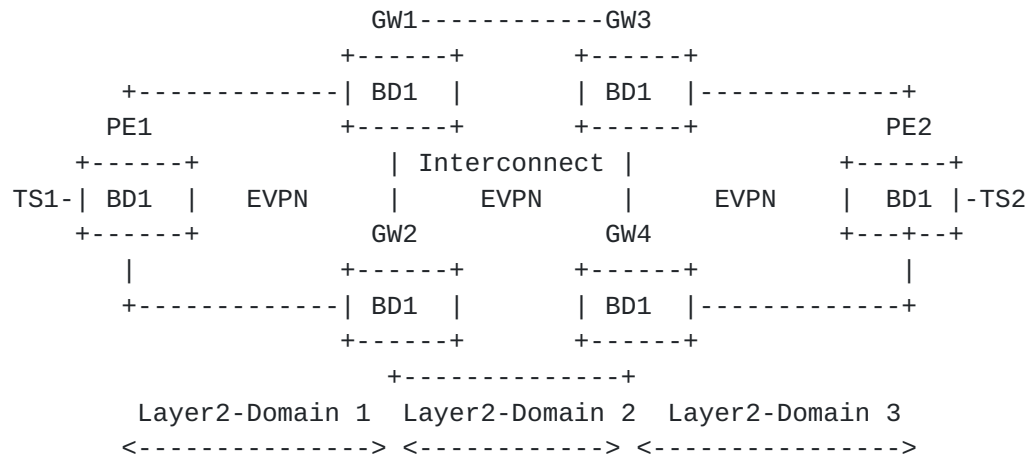


Figure 1: EVPN Sub-Domain example

- Layer2-Domain Gateway PE: a PE that is attached to two or more EVPN Layer2-Domains. An example of Layer2-Domain Gateway PE is a PE following the procedures in section 4.4 or section 4.6 of [RFC9014]. In the example in Figure 1, GW1 and GW2 connect Layer2-Domains 1 and 2, whereas GW3 and GW4 connect Layer2-Domains 2 and 3. GW1 and GW2 import the MAC/IP Advertisement route for TS1 coming from the Layer2-Domain 1 into the MAC-VRF for BD1, and re-advertise it into Layer2-Domain 2. Likewise, GW3 and GW4 import the route into their MAC-VRF and re-advertise it into Layer2-Domain 3.

## 4. Use of Domain Path Attribute (D-PATH) with EVPN MAC/IP Advertisement routes

This document extends the use of the D-PATH attribute specified in [I-D.ietf-bess-evpn-ipvpn-interworking] so that D-PATH can be advertised and processed along with the following EVPN route types:

- EVPN MAC/IP Advertisement routes that are not used for Inter-Subnet Forwarding (ISF) [RFC9135]. Note: if the EVPN MAC/IP Advertisement route is used for Inter-Subnet Forwarding, the procedures for the D-PATH advertisement and processing are described in [I-D.ietf-bess-evpn-ipvpn-interworking].

- EVPN A-D per EVI routes that are used for EVPN-VPWS [RFC8214]. The use of D-PATH in A-D per EVI routes not used for EVPN-VPWS is out of scope of this document.

The use of D-PATH along with other EVPN route types will be described in future versions of this document.

When used along with non-ISF MAC/IP Advertisement routes or A-D per EVI routes, the D-PATH attribute is characterized as follows:

1. D-PATH is composed of a sequence of Domain segments following the format specified in [I-D.ietf-bess-evpn-ipvpn-interworking] where each Domain is represented as <DOMAIN-ID:ISF_SAFI_TYPE>. In this specification, DOMAIN-ID is a Layer2-Domain identifier configured in a MAC-VRF and ISF_SAFI_TYPE is set to either 70 (EVPN) or 0 (local route). To simplify the explanation, this document represents the domains for EVPN RT-1s and RT-2s as <Layer2-Domain-ID:TYPE>.

2. D-PATH identifies the sequence of Layer2-Domains the route has gone through, with the last <Layer2-Domain-ID:TYPE> entry (rightmost) identifying the first PE that added the D-PATH attribute.

3. D-PATH SHOULD be added/modified by a Layer2-Domain Gateway PE that re-advertises the route and MAY be added by a PE that originates the route, as follows:

   A. A Layer2-Domain Gateway PE that connects two Layer2-Domains "X" and "Y", and receives a route on a Layer2-Domain identified by a Layer2-Domain-ID "X" SHOULD append a domain <X:EVPN> to the existing (or newly added) D-PATH attribute when re-advertising the route to Layer2-Domain "Y". The route is re-advertised if it is first imported in a MAC-VRF (or VPWS instance), the MAC (or Ethernet Tag) is active, and policy allows the re-export of the route to a BGP neighbor.

   B. A Layer2-Domain Gateway PE MAY also add the D-PATH attribute for locally learned MACs or MAC/IP pairs. In this case, the domain added would be <A:0>, where A is the Layer2-Domain-ID configured on the Gateway PE's MAC-VRF that is specific to local routes (MAC/IP learned via local AC), and "0" is the TYPE of the Layer2-Domain and indicates that the route is locally originated and not re-advertised after receiving it from a BGP-EVPN neighbor. The Layer2-Domain-ID for local routes MAY be shared by all the redundant Layer2-Domain Gateway PEs for local routes, or each Layer2-Domain Gateway PE of the redundancy group can use its own Layer2-Domain-ID.

   C. A PE that is connected to a single Layer2-Domain (therefore the PE is not a Layer2-Domain Gateway) MAY add the D-PATH with a domain <B:0>, where B is the Layer2-Domain-ID

configured on the PE's MAC-VRF (or VPWS) for locally learned MAC/IPs (or Ethernet Tag IDs for VPWS). "0" is the TYPE that indicates the route is not re-advertised, but originated in the PE.

4. On received EVPN routes, D-PATH is processed and used for loop detection (Section 4.1) as well as to influence the best path selection (Section 4.2 and Section 4.3).

## 4.1. Loop Detection

An EVPN route received by a PE with a D-PATH attribute that contains one or more of its locally associated Layer2-Domain-IDs for the MAC-VRF or VPWS instance is considered to be a looped route. A looped route MUST NOT be installed, but kept in the BGP RIB, flagged as "looped".

For instance, in the example of Figure 1, assuming PE1 advertises TS1's MAC/IP and does not add the D-PATH attribute, the Layer2-Domain Gateway GW1 receives two MAC/IP Advertisement routes for TS1's MAC/IP:

- A RT-2 with next-hop PE1 and no D-PATH.

- A RT-2 with next-hop GW2 and D-PATH={length=1,<6500:1:EVPN>}, assuming that the Layer2-Domain-ID for Layer2-Domain 1 is 6500:1.

In this case, Layer2-Domain Gateway GW1 flags the RT-2 with D-PATH as "looped", and does not install the MAC in the BT of the MAC-VRF, since the route includes one of GW1's Layer2-Domain-IDs.

## 4.2. D-PATH and Best Path Selection for MAC/IP Advertisement routes

When two (or more) MAC/IP Advertisement routes with the same route key (and same or different RDs) are received, a best path selection algorithm is used. This section summarizes the best path selection for MAC/IP Advertisement routes, which extends the rules in [I-D.ietf-bess-rfc7432bis] by including D-PATH in the tie-breaking algorithm. While the algorithm may be implemented in different ways, the selection result SHOULD be the same as the result of the rules that follow.

The tie-breaking algorithm begins by considering all EVPN MAC/IP Advertisement routes equally preferable routes to the same destination, and then selects routes to be removed from consideration. The process terminates as soon as only one route remains in consideration.

1. When the Default Gateway extended community is present in some of the routes, the MAC/IP Advertisement routes without the Default Gateway indication are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis].

2. Then the routes with the Static bit set in the MAC Mobility extended community are preferred, and the routes without the Static bit set are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis]. Note that this rule does not apply to routes with the Default Gateway extended community, since these routes SHALL NOT convey the MAC Mobility extended community [I-D.ietf-bess-rfc7432bis].

3. Then the routes with the highest MAC Mobility Sequence number are preferred, hence the routes that are not tied for having the highest Sequence number are removed from consideration, as defined in [I-D.ietf-bess-rfc7432bis].

4. Then routes with the highest Local Preference are preferred, hence routes that are not tied for having the highest Local Preference are removed from consideration, as defined in [RFC4271].

5. Then routes with the shortest D-PATH are preferred, hence routes not tied for the shortest D-PATH are removed. Routes without D-PATH are considered zero-length D-PATH.

6. Then routes with the numerically lowest left-most Sub-Domain-ID are preferred, hence routes not tied for the numerically lowest left-most Sub-Domain-ID are removed from consideration.

7. If the steps above do not produce a single route, the rest of the rules after the highest Local Preference in [RFC4271] apply after step 6.

The above selection criteria is followed irrespective of the ESI value in the routes. EVPN Multi-Homing procedures for Aliasing or Backup paths in [I-D.ietf-bess-rfc7432bis] are applied to the selected MAC/IP Advertisement route.

### 4.3. D-PATH and Best Path Selection for Ethernet A-D per EVI routes

When two (or more) EVPN A-D per EVI routes with the same route key (and same or different RDs) are received for a VPWS, a best path selection algorithm is used. The selection algorithm follows the same steps as in Section 4.2 except for steps 1-3 which do not apply since the Default Gateway and MAC Mobility extended community are irrelevant to the EVPN A-D per EVI routes.

The above selection is followed for A-D per EVI routes with ESI=0. For non-zero ESI routes, the EVPN Multi-Homing procedures in [RFC8214] for Aliasing and Backup path are followed to select the routes (P and B flags are considered for the selection of the routes when sending traffic to a remote Ethernet Segment). If the mentioned Multi-Homing procedures do not produce a single route to each of the remote PEs attached to the same ES, steps 4-7 in Section 4.2 are followed.

## 4.4. Error Handling

The error handling for the D-PATH attribute is described in [I-D.ietf-bess-evpn-ipvpn-interworking]. This document extends the use of D-PATH to non-ISF EVPN routes.

# 5. Use-Case Examples

This section illustrates the use of D-PATH in EVPN routes with examples.

Figure 2 and Figure 3 illustrate an integrated interconnect solution for an EVPN BD, as described in section 4.4 and section 4.6 of [RFC9014]. GW1 and GW2 are Layer2-Domain Gateway PEs connecting two L2-Domains identified by D-PATH domain {1:1:EVPN} and {1:2:EVPN}, respectively. Received Ethernet A-D routes, ES routes, and Inclusive Multicast routes from the routers in one Layer2-Domain are consumed and processed by GW1 and GW2, but not re-advertised to the other Layer2-Domain. However, MAC/IP Advertisement routes received by GW1 and GW2 in one Layer2-Domain are processed and, if installed, re-advertised into the other Layer2-Domain.

Consider the example of Figure 2, where PE1 advertises a MAC/IP Advertisement route for M1/IP1. The route is processed and installed by GW1 and GW2 in BD1, and both will re-advertise the routes into the Layer2-Domain-2. By using D-PATH in GW1 and GW2, when the route is received on PE2, PE2 has the visibility of the Layer2-Domains through which the route has gone, and can also use the D-PATH for best path selection in case PE2 receives a MAC/IP Advertisement route for M1/IP1 by some other means. In addition, GW1 and GW2 can compare the D-PATH of the incoming routes with their local list of Layer2-Domain-IDs, and detect a loop if any of the local Layer2-Domain-IDs matches a domain in the received D-PATH. This procedure prevents the re-advertisement of the route back into Layer2-Domain-1.

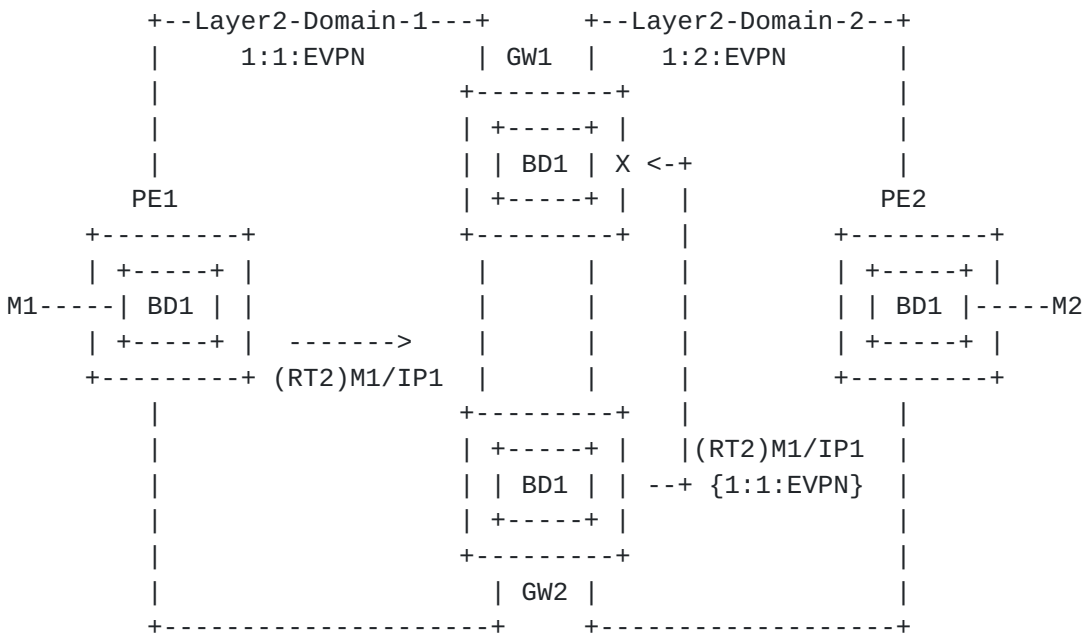


Figure 2: EVPN Interconnect

The example of Figure 3 illustrates how GW1 and GW2 can also have local ACs in BD1 and learn local MAC (or MAC/IP) addresses from devices connected to the ACs. Assuming GW2 learns M3/IP3 via local AC, GW2 advertises a MAC/IP Advertisement route for M3/IP3 into each of the Layer2-Domains that GW2 is connected to. As described in Section 4, GW2 can advertise these two MAC/IP Advertisement routes with a configured Layer2-Domain-ID for local MAC/IPs routes that can be shared with GW1. Consider this Layer2-Domain-ID is 1:3 and it is configured on both, GW1 and GW2. When GW2 advertises the route into each Layer2-Domain, it adds the D-PATH attribute with a domain {1:3:0}. These routes will be flagged by GW1 as "looped" since 1:3 is configured as a local Layer2-Domain-ID in GW1. In addition, PE1 and PE2 will receive the routes with the D-PATH and they will have the visibility of the origin of the routes, in this case local Layer2-Domain Gateway routes. This information can be used to influence the best path selection in case of multiple routes for M3/IP3 are received on PE1 or PE2 for BD1.

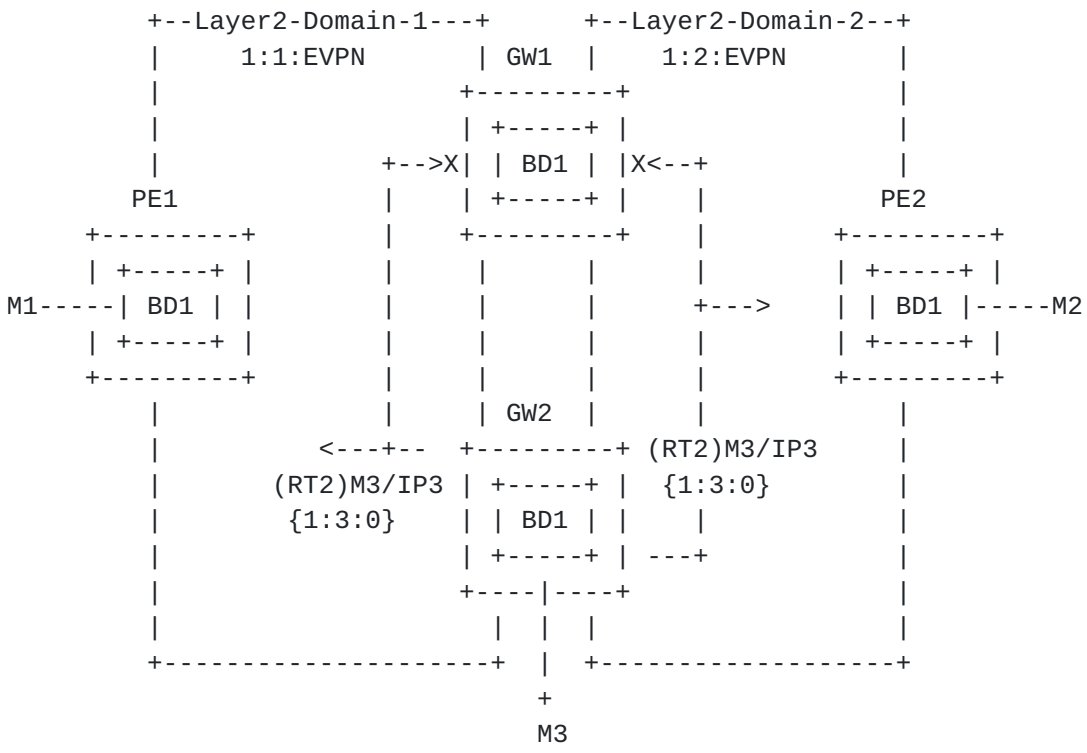


Figure 3: EVPN Interconnect local AC

As an alternative solution to configuring the same Layer2-Domain-ID for local routes on both Layer2-Domain Gateways, GW2 can be configured with Layer2-Domain-ID 1:3 for local routes, and GW1 can use a different Layer2-Domain-ID, e.g., 1:4. In this case, GW2 advertises the route for M3/IP3 into each Layer2-Domain as before, but now GW1 will not flag the route as "looped" since 1:3 is not on the list of GW1's local Layer2-Domain-IDs. GW1 receives the routes from both Layer2-Domains, and GW1 selects the route from e.g., Layer2-Domain-1. GW1 then installs the route in its BT and re-advertises the route into Layer2-Domain-2 with D-PATH {1:1:EVPN, 1:3:0}. When PE2 receives two routes for M3/IP3, one from GW2 with D-PATH {1:3:0} and another from GW1 with D-PATH {1:1:EVPN, 1:3:0}, PE2 will use best path selection and choose to send its traffic to GW2. Also GW2 will receive the route for M3/IP3 from GW1 and mark it as "looped" since that route conveys its own Layer2-Domain-IDs 1:1 and 1:3.

In a nutshell, the use of D-PATH in MAC/IP Advertisement routes helps prevent loops and influences the best path selection so that PEs choose the shortest paths to the destination PEs.

## 6. Security Considerations

To be added.

## 7. IANA Considerations

None.

## 8. Acknowledgments

## 9. Contributors

## 10. References

### 10.1. Normative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[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>.

[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>.

[I-D.ietf-bess-evpn-ipvpn-interworking] Rabadan, J., Sajassi, A., Rosen, E., Drake, J., Lin, W., Uttaro, J., and A. Simpson, "EVPN Interworking with IPVPN", draft-ietf-bess-evpn-ipvpn-interworking-06 (work in progress), September 2021.

[I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", draft-ietf-bess-rfc7432bis-01 (work in progress), July 2021.

[RFC9014] Rabadan, J., Ed., Sathappan, S., Henderickx, W., Sajassi, A., and J. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, DOI 10.17487/RFC9014, May 2021, <https://www.rfc-editor.org/info/rfc9014>.

[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>.

## 10.2. Informative References

[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>.

## Authors' Addresses

J. Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

S. Sathappan
Nokia
701 Middlefield Road
Mountain View, CA 94043
USA

Email: senthil.sathappan@nokia.com