Workgroup: BESS WG

Published: 24 October 2021

Intended Status: Standards Track

Expires: 27 April 2022

Authors: Y. Wang Q. Niu

ZTE Corporation ZTE Corporation

Distributed Bump-in-the-wire Use Case

Abstract

The Bump-in-the-wire use-case of <u>Section 4.3</u> of [<u>RFC9136</u>] is a centerlized inter-subnet forwarding solution. The centerlized intersubnet forwarding burdens the DGWs with the L3 traffics among different subnets inside the same DC.

This draft extends the Bump-in-the-wire use-case of Section 4.3 of [RFC9136] in order to achieve a distributed inter-subnet forwarding solution.

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 27 April 2022.

Copyright Notice

Copyright (c) 2021 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 Simplified BSD License text as described in

Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

- 1. Introduction
 - 1.1. Terminology and Acronyms
- 2. Problem Statement
 - 2.1. Centerlized Inter-subnet Forwarding
 - 2.2. RT-1 Confliction among Multiple Bump-in-the-wires
- Solutions
 - 3.1. Supplementary BD for Bump-in-the-wire
 - 3.2. Constructing IP Prefix Advertisement Route
 - 3.3. ACI-specific Supplementary Overlay Index Extended Community
 - 3.4. Determining the Aliasing Pathes for RT-5E
 - 3.5. Other Considerations
- 4. IANA Considerations
- <u>5. Security Considerations</u>
- References
 - 6.1. Normative References
 - 6.2. Informative References

Authors' Addresses

1. Introduction

As shown in <u>Figure 1</u>, the Bump-in-the-wire use-case of <u>Section 4.3</u> of [RFC9136] is a centerlized inter-subnet forwarding solution. The centerlized inter-subnet forwarding burdens the DGWs with the L3 traffics among different subnets (e.g. SN1 and H3 of <u>Figure 2</u>) inside the same DC.

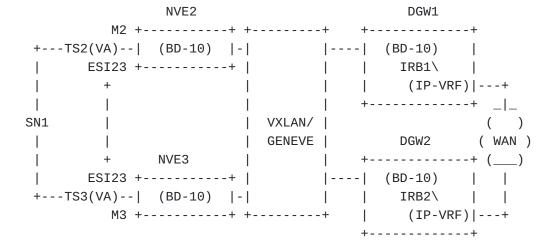


Figure 1: RFC9136's Figure 7

When a SBD is added (see <u>Figure 4</u>) for the IP-VRF instance, using this SBD and its SBD IRB, we can extend the Bump-in-the-wire use

case to form a distributed inter-subnet forwarding solution which will not burden the DGWs with the L3 traffics among different subnets inside the same DC.

But when multiple Bump-in-the-wires are integrated into the same IP-VRF (as shown in <u>Figure 3</u>), the above extension is not enough, the details are discribed in <u>Section 2.2</u>, thus some futher extensions are introduced to solve that problem.

The RT-5 route that specifies an ESI as overlay index is first defined in <u>Section 4.3</u> of [RFC9136], where the Bump-in-the-wire use case (which is called the first type RT-5E usage) is also defined there.

Note that the RT-5E routes (which are called the second type RT-5E usage) of <u>Section 4.3.2</u> of [<u>I-D.wang-bess-evpn-arp-nd-synch-without-irb</u>] and <u>Section 1.3</u> of [<u>I-D.sajassi-bess-evpn-ip-aliasing</u>] are different from these RT-5E routes of Bump-in-the-wire use case in the following factors:

- *Source MAC The ethernet header can not be absent in the first type usage even if the data plane is MPLS. The source MAC MUST be set to the MAC address of the IRB interface of BD-10 in Bump-inthe-wire usecase. But in the second type usage the ethernet header can be absent if the data plane is MPLS.
- *Recursive Resolution The recursive resolution of the first type usage are done in the context of a BD, But the recursive resolution of the second type usage are done in the context of a IP-VRF.
- *EVPN label The EVPN label of the corresponding RT-1 per EVI route of the first type usage is a MPLS label which identifies a BD, But the EVPN label of the corresponding RT-1 per EVI route of the second type usage is a MPLS label which identifies an IP-VRF.
- *ESI The ESI of the first type usage is attached to a BD, But ESIs of the second type usage are attached to IP-VRFs.

The Bump-in-the-wire use case is a special form of EVPN IRB use case, that's why its corresponding RT-1 per EVI routes are resolved in BD context.

1.1. Terminology and Acronyms

Most of the acronyms and terms used in this documents comes from $[\underline{\mathsf{RFC9136}}]$ and $[\underline{\mathsf{I-D.wang-bess-evpn-ether-tag-id-usage}}]$ except for the following:

- An Attachment Circuit (AC) that attaches a CE to an IP-VRF but is not an IRB interface.
- * VRF Interface An IRB interface or a VRF-AC or an IRC interface.

 Note that a VRF interface will be bound to the routing space of an IP-VRF.
- * L3 EVI An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN which contains VRF ACs and maybe contains IRB interfaces or IRC interfaces.
- * RT-1 per EVI Ethernet Auto-Discovery route per EVI, and the EVI here is an IP-VRF. Note that the Ethernet Tag ID of an RT-1 per EVI route may be not zero.
- * **IP-AD/ES** Ethernet Auto-Discovery route per ES, and the EVI for one of its route targets is an IP-VRF.
- * RMAC Router's MAC, which is signaled in the Router's MAC extended community.
- * ESI Overlay Index ESI as overlay index.
- * **ET-ID** Ethernet Tag ID, it is also called ETI for short in this document.
- * RT-5E An EVPN Prefix Advertisement Route with a non-reserved ESI as its overlay index (the ESI-as-Overlay-Index-style RT-5)
- * **CE-BGP** The BGP session between PE and CE. Note that CE-BGP route doesn't have a RD or Route-Target.
- * **CE-Prefix -** An IP Prefixes behind a CE is called as that CE's CE-Prefix.
- * ETI-Agnostic BD A Broadcast Domain (BD) whose data packets can be received along with any Ethernet Tag ID (ETI). Note that a broadcast domain of an L2 EVI of VLAN-aware bundle service interface is a good example of an ETI-Specific BD.
- * ETI-Specific BD A Broadcast Domain (BD) whose data packets are expected to be received along with a normalized Ethernet Tag ID (ETI). Note that a broadcast domain of an L2 EVI of VLAN-bundle or VLAN-based service interface is a good example of an ETI-Agnostic BD.
- * BDI-Specific EADR When the <ESI, BD> uses BDI-Specific Ethernet Auto-discovery mode, the only Ethernet A-D per EVI route of that <ESI, BD> is called as a BDI-Specific EADR in this draft.

* ACI-Specific EADR -

When the <ESI, BD> uses ACI-Specific Ethernet Auto-discovery mode, the Ethernet A-D per EVI routes of that <ESI, BD> are called as ACI-Specific EADRs in this draft.

2. Problem Statement

2.1. Centerlized Inter-subnet Forwarding

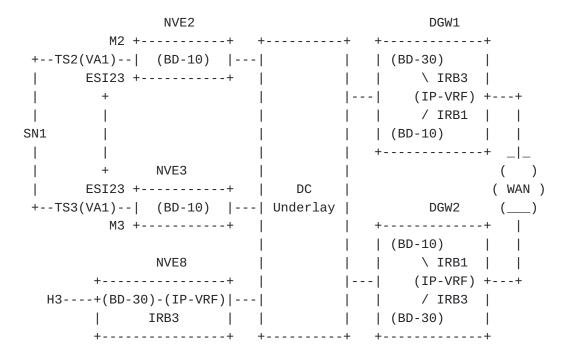


Figure 2: Centerlized Bump-in-the-wire Use Case

As shown in <u>Figure 2</u>, SN1 and H3 are both internal hosts of the same DC. But the communication between them have to pass through a DGW, that's why the DGWs will be burdened with inter-subnet forwarding of the internal hosts.

The <u>Section 4.3</u> of [<u>RFC9136</u>] defined the Bump-in-the-wire use-case, where a style (which is called as RT-5E in this draft) of RT-5 routes (whose overlay index is a non-zero ESI), is used to advertise the IP prefix of subnet SN1 (see <u>Figure 3</u>). The RT-5E routes (whose IP prefix is SN1, and ESI is ESI23) of <u>Section 4.3</u> of [<u>RFC9136</u>] is called as RT5E_SN1 in this draft. And the RT-1 routes (whose ESI is ESI23) corresponding to the RT5E_SN1 is called as RT1_ESI23 in this draft.

Note that when DGW1 or DGW2 receives RT5E_SN1, it should know (before the recursive resolution) that RT5E_SN1's ESI (ESI23) should be resolved in the context of BD-10, not in BD-30 (whether BD-30 is another Bump-in-the-wire BD or not). Because of RT5E_SN1's Route

target (which identifies BD-10), DGW1 can know that before the recursive resolution.

2.2. RT-1 Confliction among Multiple Bump-in-the-wires

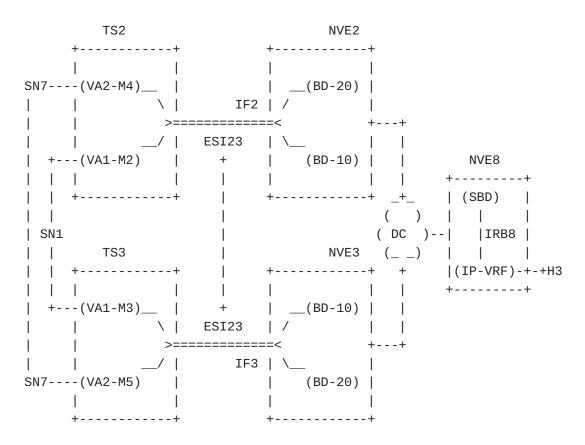


Figure 3: ET-ID Confliction of Bump-in-the-wire

This network is another view of a part of <u>Figure 4</u>, and it is similar to <u>Section 4.3</u> of [RFC9136] with a few notable exceptions as below:

The NVE2, NVE3, BD-10, ESI23, TS2, TS3 and SN1 here is the NVE2, NVE3, BD-10, ESI23, TS2, TS3 and SN1 there (Section 4.3 of [RFC9136]). The VA1 here is the Virtual Appliance (whose VA-MAC is M2/M3 on TS2/TS3) there. The NVE8 here is the DGW1 there. The IRB8 here takes the place of the IRB1 there.

But here we have another Bump-in-the-wire instance for Virtual Appliance VA2, which are attached to another Broadcast Domain BD-20. Both BD-10 and BD-20 are integrated into the same IP-VRF by DGW1. But the subnet SN1 can only be reached through BD-10, while the subnet SN7 can only be reached through BD-20.

RT5E_SN1 (whose route-target identifying BD-10) is imported into the BD-10 at first, although it can be imported into the IP-VRF following BD-10's IRB interface, RT5E_SN1 will not be imported into

the IP-VRF on other PEs which don't have an instance of BD-10. Thus such PEs are precluded from connecting to the hosts of SN1 by such rules.

Note that both BD-10 and BD-20 are L2 EVIs of VLAN-based Service Interfaces.

The solution for this problem is decribed in Section 3.5.

3. Solutions

3.1. Supplementary BD for Bump-in-the-wire

As shown in Figure 4, the SN1, BD-10, IP-VRF are the same as Figure 2, except that the TS2, TS3 and ESI23 are not shown in Figure 4, but they are still there unchanged. Then we add a SBD for the IP-VRF instance, and each SBD will be configured with an IRB interface (which is called its SBD IRB). Using this SBD and its SBD IRB, we can extend the Bump-in-the-wire use case to form a distributed inter-subnet forwarding solution which will not burden the DGWs with the L3 traffics among different subnets inside the same DC.

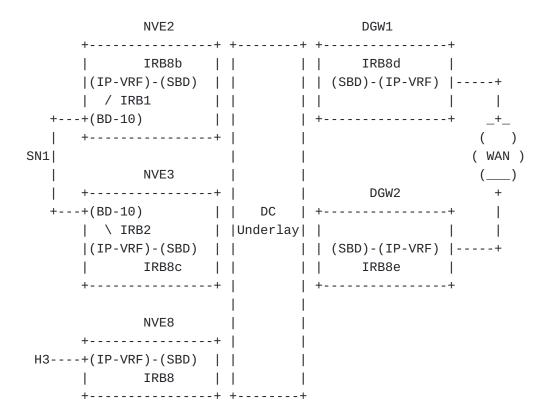


Figure 4: Distributed Bump-in-the-wire Use Case

The RT-5 route (say RT5E_SN1) advertised by NVE2/NVE3 for SN1 is the same as <u>Section 4.3</u> of [<u>RFC9136</u>] except for the following notable differences:

The route-targets of RT5E SN1 is set to the export-RT of the SBD.

* The RT-1 route of ESI23 MUST be advertised both for BD-10 and the SBD, when they are advertised for the SBD, the EVPN label of the RT-1 per EVI route should be set to the EVPN label of the BD-10, as if it is advertised for BD-10.

Note that when it is advertised for the SBD, it may use different RD than it is advertised for BD-10.

* In order to process the RT5E_SN1 properly, the DGW1 and DGW2 don't have to change its behavior of Section 4.3 of [RFC9136]. But the configurations of DGW1 and DGW2 must be changed, because that the BD-10 is removed and the SBD takes its place.

Note that to the RT5E_SN1 route, the NVE8 is actually no different from DGW1 and DGW2. NVE8 is not a DC gateway, but whether NVE8 is a DC gateway is not awared by NVE1 and NVE2.

3.2. Constructing IP Prefix Advertisement Route

The RT5E_SN1 is constructed following <u>Section 4.3</u> of [RFC9136] except for the following differences:

* Route target and RD

The route target of RT5E_SN1 MUST be set to the route-target which identifies the SBD. In other words, RT5E_SN1 is advertise for the SBD, or we can see RT5E_SN1 is advertised in the context of the SBD.

The RD of RT5E SN1 can be set to the RD of SBD too.

* ESI and ET-ID

No matter whether BD-10 is an ETI-agnostic BD or ETI-specific BD, it will be enough to configure the SBD as an ETI-agnostic BD. But the Ethernet Tag ID of the Ethernet A-D per EVI routes of the SBD may be set to non-reserved ET-IDs.

When an CE-prefix of a Bump-in-the-wire instance is advertised by a RT-5E route, The RT-5E route is advertised in the SBD's context. The RT-5E route's ESI MUST be determined by the CE-prefix's VA MAC (which will be known by policy). Take SN1 of Figure 4 for example, by policy, we can know that the VA MAC M1 is in BD-10, then we can know that VA MAC M1 is learnt over <ESI23, BD-10>, so the ESI of RT5E_SN1 should be set to ESI23.

If BD-10 is an ETI-agnostic BD (e.g. BD-10 is of VLAN-based service interface), the ET-ID of RT5E_SN1 MUST be set to 0. If

BD-10 is an ETI-specific BD (e.g. BD-10 is of VLAN-aware bundle service interface), the ET-ID of RT5E_SN1 MUST be set to the BD-ID of BD-10 (even if the SBD is ETI-agnostic).

Note that the ET-ID of RT5E_SN1 is not used to resolve (as described in $\underline{\text{Section 3.4}}$) RT5E_SN1's ESI overlay index to a proper Ethernet A-D per EVI route.

* ACI-Specific Supplementary Overlay Index

When an IP Prefix Advertisement is advertised, The ACI-Specific Supplementary Overlay Index (SOI) extended community is always recommanded to be carried along with it, if it is not clear that whether there will be conflictions among Ethernet A-D per EVI routes inside the SBD in the future.

Note that the ACI-Specific SOI here is not used to isolate IP address spaces. It is just used to resolve (as described in Section 3.4) RT5E_SN1's ESI overlay index to a proper Ethernet A-D per EVI route.

ACI-specific Overlay Index extended community should be advertised along with the RT-5E routes. Thus the ET-ID of these RT-5E routes can be set to zero if BD-10 and BD-20 are ETI-agnostic BDs.

Note that the combination of <ESI, SOI> will be used to select the corresponding RT-1 per EVI routes (in SBD) for these RT-5E routes on other PEs.

Note that in the data plane, the EVPN label that is encapsulated by NVE8 for NVE2 or NVE3 will be a label that identifies BD-10. So when BD-10 is an ETI-Specific BD, the ET-ID of RT5E_SN1 MUST be encapsulated into the ethernet header of the data packets. Otherwise such data packets won't be received by BD-10 (of NVE2 or NVE3).

3.3. ACI-specific Supplementary Overlay Index Extended Community

A new EVPN BGP Extended Community called Supplementary Overlay Index is introduced. This new extended community is a transitive extended community with the Type field of 0x06 (EVPN) and the Sub-Type of TBD. It is advertised along with EVPN MAC/IP Advertisement Route (Route Type 2) per [RFC7432] in ACI-Sepecific Ethernet Auto-Discovery mode. It may also be advertised along with EVPN Prefix Advertisement Route (Route Type 5) as per [RFC9136]. Generically speaking, the new extended community must be attached to any routes which are leant over an <ESI, EVI> of ACI-specific Ethernet Auto-Discovery.

The Supplementary Overlay Index Extended Community is encoded as an 8-octet value as follows:

Figure 5: Supplementary Overlay Index Extended Community

o F: Format Indicator, its value is always 1 in this draft. Other values are reserved.

o Type: .

* 0: VLAN-based AC-ID.

No.	Use Cases	Туре	VLAN2	VLAN1	MBZ
1	untag	type 0	Θ	Θ	0
2	<u>default</u>	type 0	0	<u>FFF</u>	0
3	dot1q	type 0	0	<u>E</u>	0
4	QinQ	type 0	<u>E</u>	Ī	0

Table 1: VLAN-based AOIs

Notes:

E: That field is the External VLAN of the AC.

I: That field is the Internal VLAN of the AC.

0 : The tag corresponding to that field is absent.

FFF: The AC is the <u>default subinterface</u> (<u>Section 3.3</u>) of the corresponding ES.

untag : An untagged subinterface should be matched by that format.

default : A default subinterface should be matched by that
 format. When the AC is a default subinterface, it will
 match all the remaining VLAN-tags (which are left over by
 other subinterfaces) on its main-interface.

dot1q : A dot1q subinterface should be matched by that format.

QinQ: A QinQ subinterface should be matched by that format.

- * **1-15**: Reserved.
- **o O Flag:** Overlay Index Flag, this extended community is used as overlay index.

When type field is 0-1: For ACI-Specific Ethernet auto-discovery mode, when it is carried along with a RT-2 route, the 0 Flag should be set to 1, For BDI-Specific Ethernet auto-discovery, when it is carried along with a RT-2 route, the 0 Flag should be set to 0.

When the O Flag is set to 1, this AC-ID is also called as AOI (ACI-Specific Overlay Index), and the <ESI, AOI> of that RT-2R or RT-5E should be used to determine ECMP pathes. At the same time, the AOI should also be used like Attachment Circuit ID Extended Community too.

Note that only the lowest 8 bits of MBZ field should be used to select RT-1 per EVI routes. <lowest 8 bits of MBZ, VLAN2, VLAN1> of a type-0 AOI forms an Ethernet Tag ID of an ACI-Specific EADR.

- **o Z Flag:** Must be zero. Reserved for future use, the receiver should ignore this extended coummunity if Z flag is not zero at now.
- **o Flags:** Reserved for future use. it is set to 0 on advertising, and ignored on receiving.

Note that although this extended community is similar to the AC-ID extended community (as per [I-D.sajassi-bess-evpn-ac-aware-bundling]), we can assume that they may be of different Sub-Types because that they have different behaviors.

3.4. Determining the Aliasing Pathes for RT-5E

No matter whether a RT-5 route is constructed following $\frac{Section \ 4.3}{Section \ 3.2}$ of this draft, the RT-1 per EVI routes corresponding to that RT-5E route will be resolved in the context of a BD, not in an IP-VRF.

When resolving corresponding RT-1 per EVI routes for a RT-5E route, the AOI (ACI-specific SOI) Extended Community of the RT-5E route can be used.

Note that when the RT-5E's AOI is Y (Y!=0), the ET-IDs of the selected Ethernet A-D per EVI routes (of that RT-5E) should be all Y.

Note that when the RT-5E's ET-ID is not 0, and an AOI is advertised along with the RT-5E, the Ethernet A-D per EVI routes of that RT-5E should be selected according to the <ESI, AOI>.

Note that when a data packet is load-balanced according to <ESI, AOI>, in Bump-in-the-wire use case, it is the RT-5E's ET-ID which

should be encapsulated into the data packet (as 802.1q Tag), not the AOT.

Note that [I-D.sajassi-bess-evpn-ac-aware-bundling] requires the Presence of Attachment Circuit ID Extended Community MUST be ignored by non multihoming PEs. It requires the remote PE (non-multihome PE, e.g. PE3) MUST process MAC route as defined in [RFC7432]. But the AOI of this case should be used to select ETI-Specific EADRs. This is non-compatible with the Attachment Circuit Extended Community, thus the new ACI-Specific Overlay Index Extended Community is defined.

3.5. Other Considerations

We can assume that maybe neither BD-10 nor BD-20 will be configured on NVE8, as illustrated in <u>Figure 4</u>. In such case, we assume that a SBD (Supplementary BD) can be provisoned on NVE8.

The SBD is similar to the combination of the SBD of Section 4.4.3 of [RFC9136] and the BD-10 of Section 4.3 of [RFC9136], except for the following factors:

The RT-1 per EVI routes advertised for SBD is originated from the BD-10. and the SBD don't have to advertise any EVPN routes (e.g. IMET route) of its own. because there are no hosts (even the IP address of SBD IRB will not be provisoned in this case) in the SBD.

Note that DGWs will advertise their own IP prefixes using their own L3 EVPN label and route-targets. They don't have to expect any data packets to be received from such SBD.

The route advertisement behavior of NVE2 and NVE3 should also be changed:

*When BD-10 advertised a RT-1 per EVI route RT1a, another RT-1 per EVI route RT1b (which is the mirroring of RT1a) should be advertised for the SBD. Although RT1b is advertised for the SBD, RT1b's EVPN label should be set to BD-10's EVPN label, not the SBD's EVPN label. RT1b's ET-ID MUST be set to the AC-ID of the AC corresponding to RT1a.

Otherwise the RT-1 per EVI routes for BD-10 and BD-20 will conflict with each other, because that both BD-10 and BD-20 are of VLAN-based Servcice Interface.

*The MAC addresses of IRB interface of each Bump-in-the-wire BD (e.g. BD-10 and BD-20) should be the same as the SBD IRB interface of the same L3 EVI, otherwise the source MAC may be not expected to be learnt by the CE-side L2 switches.

4. IANA Considerations

A new transitive extended community Type of 0x06 and Sub-Type of TBD for EVPN Supplementary Overlay Index Extended Community needs to be allocated by IANA.

5. Security Considerations

TBD.

6. References

6.1. Normative References

[I-D.sajassi-bess-evpn-ac-aware-bundling]

Sajassi, A., Brissette, P., Mishra, M., Thoria, S., Rabadan, J., and J. Drake, "AC-Aware Bundling Service Interface in EVPN", Work in Progress, Internet-Draft, draft-sajassi-bess-evpn-ac-aware-bundling-04, 11 July 2021, https://datatracker.ietf.org/doc/html/draft-sajassi-bess-evpn-ac-aware-bundling-04.

[I-D.sajassi-bess-evpn-ip-aliasing]

Sajassi, A., Badoni, G., Warade, P., Pasupula, S., Drake, J., and J. Rabadan, "EVPN Support for L3 Fast Convergence and Aliasing/Backup Path", Work in Progress, Internet-Draft, draft-sajassi-bess-evpn-ip-aliasing-02, 8 June 2021, https://datatracker.ietf.org/doc/html/draft-sajassi-bess-evpn-ip-aliasing-02.

- [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>.
- [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>.

6.2. Informative References

synch-without-irb-08, 1 September 2021, <<u>https://</u>
datatracker.ietf.org/doc/html/draft-wang-bess-evpn-arpnd-synch-without-irb-08>.

[I-D.wang-bess-evpn-ether-tag-id-usage]

Wang, Y., "Ethernet Tag ID Usage Update for Ethernet A-D per EVI Route", Work in Progress, Internet-Draft, draft-wang-bess-evpn-ether-tag-id-usage-03, 26 August 2021, https://datatracker.ietf.org/doc/html/draft-wang-bess-evpn-ether-tag-id-usage-03>.

Authors' Addresses

Yubao Wang ZTE Corporation No.68 of Zijinghua Road, Yuhuatai Distinct Nanjing China

Email: wang.yubao2@zte.com.cn

Qibo Niu ZTE Corporation No. 50 Software Ave, Yuhuatai Distinct Nanjing China

Email: niu.qibo@zte.com.cn