BESS WG
Internet-Draft
Intended status: Standards Track
Expires: 18 April 2022

Y. Wang
ZTE Corporation

15 October 2021

Distributed Bump-in-the-wire Use Case
draft-wang-bess-evpn-distributed-bump-in-the-wire-00

Abstract

The Bump-in-the-wire use-case of Section 4.3 of [RFC9136] is a centerlized inter-subnet forwarding solution. The centerlized inter-subnet 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.

Table of Contents

1. Introduction

As shown in Figure 1, the Bump-in-the-wire use-case of Section 4.3 of [RFC9136] is a centerlized inter-subnet forwarding solution. The centerlized inter-subnet forwarding burdens the DGWs with the L3 traffics among different subnets inside the same DC.

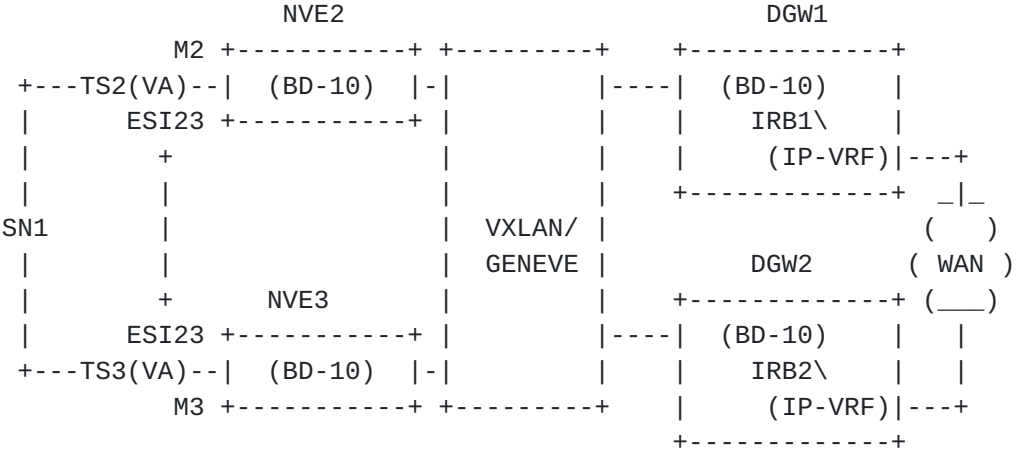

Figure 1: Centerlized Bump-in-the-wire Use Case

As shown in Figure 2, the SN1, BD-10, IP-VRF are the same as Figure 1, except that the TS2, TS3 and ESI23 are not shown in Figure 2, 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.

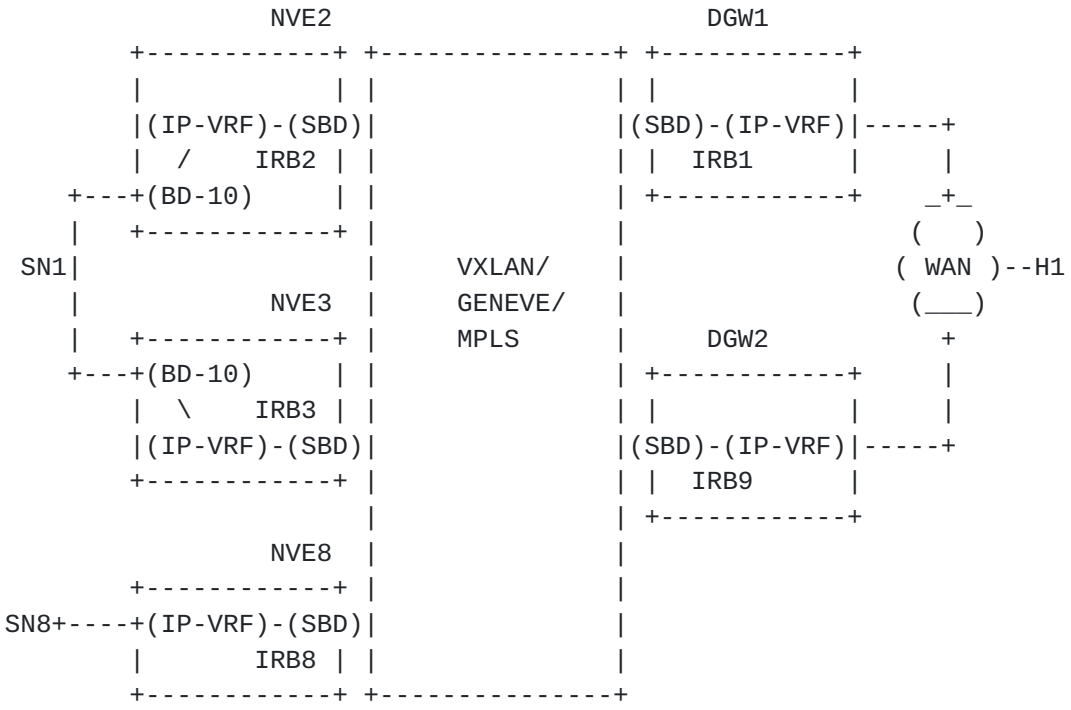

Figure 2: Distributed Bump-in-the-wire Use Case

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

* 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.

But when multiple Bump-in-the-wire are integrated into the same IP-VRF, the above extension is not enough, the details are discribed in Section 2.1, thus some extensions are introduced to solve that problem.

1.1. Terminology and Acronyms

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

* VRF AC - 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.

* IP-AD/EVI - Ethernet Auto-Discovery route per EVI, and the EVI here is an IP-VRF. Note that the Ethernet Tag ID of an IP-AD/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. Problem with Bump-in-the-wire Use-Case

The Section 4.3 of [RFC9136] 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 Figure 3). The RT-5E routes (whose IP prefix is SN1, and ESI is ESI23) of that draft 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.

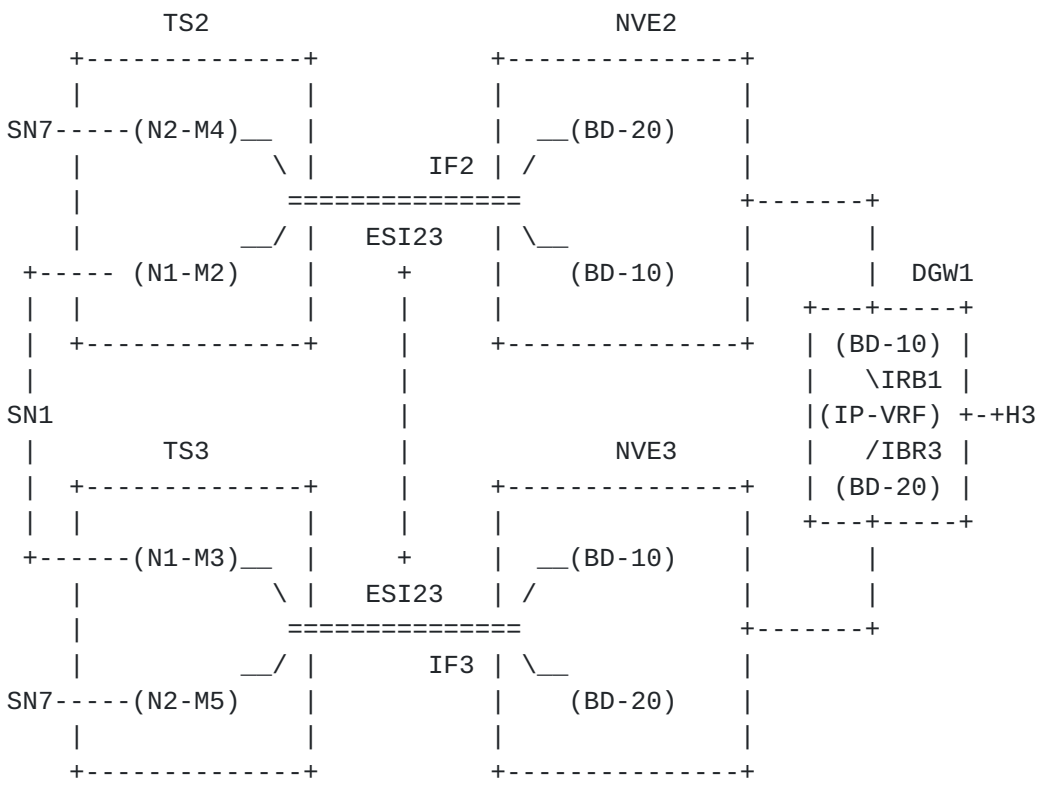

Figure 3: RT-1 Confliction of Bump-in-the-wire

This network is similar to Figure 7 of Section 4.3 of [RFC9136] with a few notable exceptions as below.

The NVE2,NVE3,DGW1,IRB1,BD-10,ESI23,TS2,TS3 and SN1 here is the NVE2,NVE3,DGW1,IRB1,BD-10,ESI23,TS2,TS3 and SN1 there. The N1 here is the Virtual Appliance (whose VA-MAC is M2/M3 on TS2/TS3) there.

But here we have another Virtual Appliance N2, 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 3.4.

3. Solutions

3.1. Determining the Aliasing Pathes for RT-5E

In this case, the RT-1 per EVI routes corresponding to that RT-5E route are selected in the context of a BD.

When selecting corresponding IP-AD/EVI routes for a RT-5E route, the AOI Extended Community (if it exists) of the RT-5E route is prefered than the ET-ID of the RT-5E route.

* Using ET-ID to select BDI-Specific EADRs - There may be multiple Ethernet A-D per EVI routes which all can match the RT-5E's ESI. In such case, The Ethernet A-D per EVI routes with the same ET-ID as the RT-5E should be selected.

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

  Note that the RT-5E's ET-ID not only just be used to select Ethernet A-D per EVI routes, but also be encapsulated into data packets in order to keep compatible with ETI-specific Bump-in-the-wire use case.

* Using AOI to select ETI-Specific EADRs - There may be multiple Ethernet A-D per EVI routes which all can match the RT-5E's ESI. In such case, The Ethernet A-D per EVI routes whose ET-ID are the same as the RT-5E's AOI should be selected.

  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 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, not the AOI.

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.

3.2. ACI-specific 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:

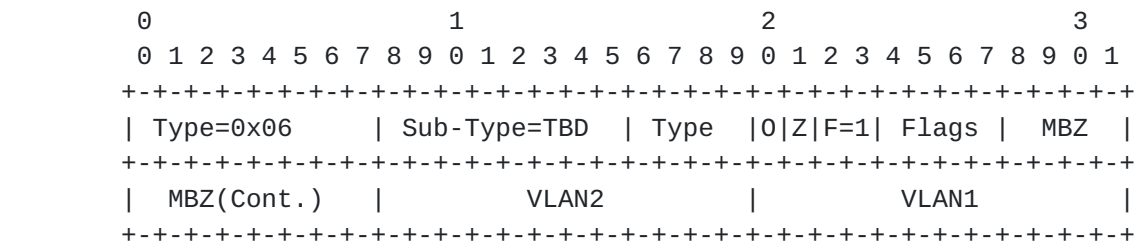

Figure 4: Supplementary Overlay Index Extended Community

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	Type	VLAN2	VLAN1	MBZ
1	untag	type 0	0	0	0
2	default	type 0	0	FFF	0
3	dot1q	type 0	0	E	0
4	QinQ	type 0	E	I	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 default subinterface (Section 3.2) 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 O Flag should be set to 1, For BDI-Specific Ethernet auto-discovery, when it is carried along with a RT-2 route, the O 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.

- 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.

3.3. Constructing IP Prefix Advertisement Route

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

Note that the ACI-Specific SOI here is not used to isolate IP address spaces. It is just used to resolve its ESI overlay index to a proper IP A-D per EVI route.

The AC-ID extended community can't be considered as a substitute of the ET-ID. Because that the AC-ID is not the key of IP A-D per EVI routes, but the ET-ID is.

3.4. Bump-in-the-wire Specific Procedures

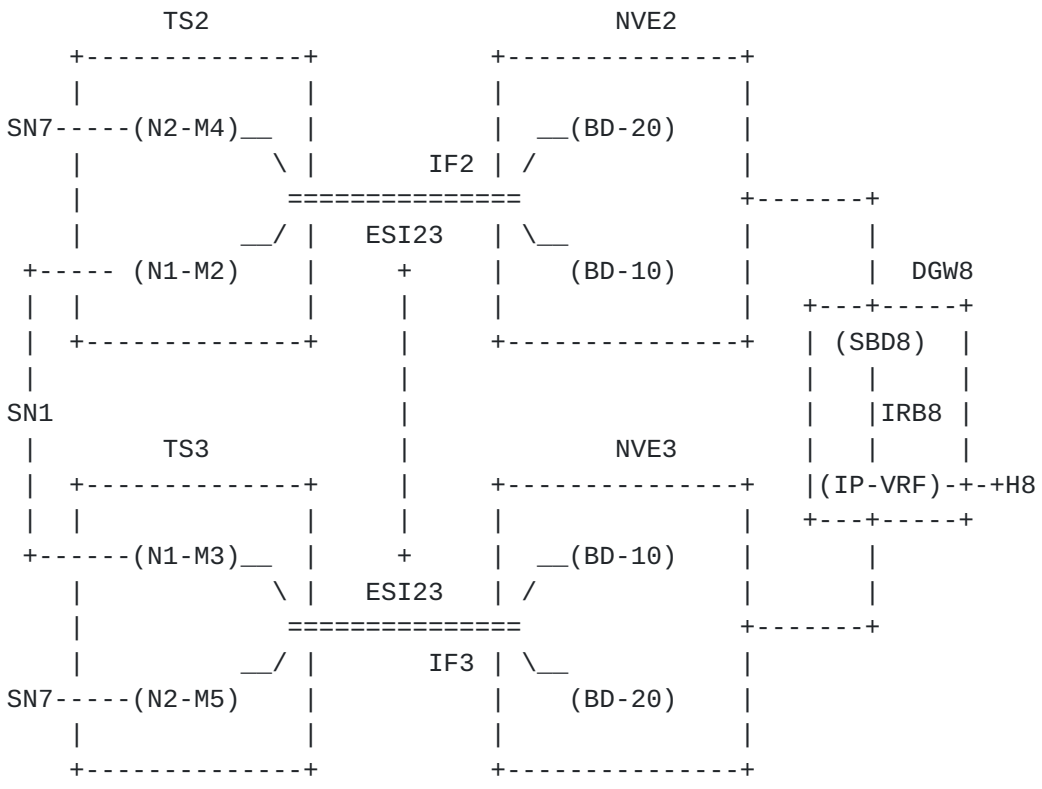

Figure 5: SBD of Bump-in-the-wire

We can assume that maybe neither BD-10 nor BD-20 will be configured on DGW8, as illustrated in Figure 5. In such case, we assume that a SBD (Supplementary BD) can be provisoned on DGW8.

The SBD8 is similar to the SBD of Section 4.4.3 of [RFC9136], except for the following factors:

- The SBD8 will import all the RT-5E routes and the RT-1 routes which are advertised for BD-10 and BD-20 (and other such BDs of Bump-in-the-wire use cases) by the NVEs of that DC, But the SBD8 don't import other EVPN routes. and it don't have to advertise any EVPN routes (e.g. IMET route) because there are no hosts (even the IP address of IRB8 will not be provisoned in this case) in the SBD.

Note that DGW3 will advertise the IP prefixes of the IP-VRF using its own EVPN label and route-targets. It 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:

* ACI-specific ethernet auto-discovery mode should be used when NVEs advertise the RT-1 routes of Bump-in-the-wire use case. Otherwise the RT-1 routes from BD-10 and BD-20 will conflict with each other, because that both BD-10 and BD-20 are of VLAN-based Servcice Interface.

* 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 because that BD-10 and BD-20 are ETI-agnostic BDs.

  Note that the combination of <ESI, AOI> will be used to select the corresponding RT-1 per EVI routes for these RT-5E routes by SBDs of other PEs.

  The RT-5E routes (for the CE-prefixes behind each BD) should be advertised follows Section 3.3. Note that the IP-ETI, EADR-ETI and IP-ACI should be determined by the outgoing AC per each CE-prefix's VA MAC. The IP-ETI is set to the BD-ID of that outgoing AC's BD. Given that <AC, BD> is ACI-Specific EAD mode, the IP-ACI is the AOI extended comunity for that <AC, BD>, and the EADR-ETI is the same value as the IP-ACI.

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

  Note that although the NVEs of the original Bump-in-the-wire don't have an IP-VRF instance, it will be no difficult for the RT5E_SN1 and RT1_ESI23 are used when there is an IP-VRF (where SN9 is learnt by a CE-BGP session) and IRBs on each of the NVEs. In such case, maybe it will be better for the NVEs to advertise its IRB (e.g. BD-10's IRB) MAC along with RT1_ESI23 (which is for BD-10) in order to indicate DGW8 to encapsulate that MAC as source MAC. By doing so, there will be no need for SBDs, thus the RT-5E routes and RT-1 per EVI routes can be advertised and imported using the IP-VRF's route-targets. The RT-1 per EVI routes (whose MPLS label identifies that BD) can be advertised along with both BD's RT and IP-VRF's RT. thus the amount of RT-1 per EVI routes will not be increased.

Note that in Bump-in-the-wire use cases, the EVPN label that is encapsulated by DGW1/DGW3 for NVE2 or NVE2 will be a label that identifies a L2 EVI. So when the BD is an ETI-Specific BD, the IP-ETI MUST be encapsulated into the ethernet header of the data packets. Otherwise such data packets won't be received by that BD.

Note that in Bump-in-the-wire use cases, even if the BD is a MPLS EVPN BD, PE3 should send data packets to NVE2/NVE3 along with the overlay ethernet header (whose SA is the SBD IRB's MAC address), because the Bump-in-the-wire use case is actually a special EVPN IRB use case. Otherwise NVE2/NVE3 can't decapsulate the data packets properly.

Note that in such case, the RT-5E routes technically can be directly imported into the IP-VRF identified by its route-target. only if there would be no more than one SBD in a specified IP-VRF. In original Bump-in-the-wire use-case, if there would be no more than one BD in a specified IP-VRF, the the RT-5E routes technically can be directly imported into the IP-VRF identified by its route-target (which is determined by policy, because a RT-5E route itself will be learnt by policy too, according to [RFC9136]). For backward comptibility with such implementations, the ET-ID (if it is not zero) of these RT-5E routes may be used to selected RT-1 per EVI routes from the only BD of that IP-VRF, and in such case it should be encapsulated into the ethernet header (in VXLAN EVPN) of the corresponding data packets, because that the RT-1 per EVI routes of Bump-in-the-wire use case are advertised for BD-10, and BD-10 may use VLAN-aware bundle service interface according to [RFC9136].

The ESIs of these RT-5E routes are used to selecte RT-1 per EVI routes in the context of a BD, as described in [RFC9136]. When the ESIs of RT-5E routes are used to selected RT-1 per EVI routes in the context of an IP-VRF, it seems that this is not explicitly described in [RFC9136].

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>.

[RFC9136] Rabadan, J., Ed., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in Ethernet VPN (EVPN)", RFC 9136, DOI 10.17487/RFC9136, October 2021, <https://www.rfc-editor.org/info/rfc9136>.

6.2. Informative References

[I-D.wang-bess-evpn-arp-nd-synch-without-irb]
Wang, Y. and Z. Zhang, "ARP/ND Synching And IP Aliasing without IRB", Work in Progress, Internet-Draft, draft-wang-bess-evpn-arp-nd-synch-without-irb-08, 1 September 2021, <https://datatracker.ietf.org/doc/html/draft-wang-bess-evpn-arp-nd-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>.

[I-D.wz-bess-evpn-vpws-as-vrf-ac]
Wang, Y. and Z. Zhang, "EVPN VPWS as VRF Attachment Circuit", Work in Progress, Internet-Draft, draft-wz-bess-evpn-vpws-as-vrf-ac-02, 28 August 2021, <https://datatracker.ietf.org/doc/html/draft-wz-bess-evpn-vpws-as-vrf-ac-02>.

Author's Address

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

Email: wang.yubao2@zte.com.cn