Workgroup: Network Working Group
Internet-Draft:
draft-sajassi-bess-evpn-ip-aliasing-06
Published: 10 January 2023
Intended Status: Standards Track
Expires: 14 July 2023
Authors: A. Sajassi, Ed. G. Badoni P. Warade
Cisco Systems Cisco Systems Cisco Systems
S. Pasupula L. Krattiger J. Drake, Ed.
Cisco Systems Cisco Systems Juniper
J. Rabadan, Ed.
Nokia

EVPN Support for L3 Fast Convergence and Aliasing/Backup Path

Abstract

This document proposes an EVPN extension to allow several of its multihoming functions, fast convergence and aliasing/backup path, to be used in conjunction with inter-subnet forwarding. The extension is limited to All-Active and Single-Active redundancy modes.

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 14 July 2023.

Copyright Notice

Table of Contents

1. Introduction

This document proposes an EVPN extension to allow several of its multihoming functions, fast convergence and aliasing/backup path, to be used in conjunction with inter-subnet forwarding. The extension is limited to All-Active and Single-Active redundancy modes. It re-

uses the existing EVPN routes, the Ethernet A-D per ES and the Ethernet A-D per EVI routes, which are used for these multihoming functions. In particular, there are three use-cases that could benefit from the use of these multihoming functions:

a. Inter-subnet forwarding for host routes in symmetric IRB [RFC9135].

b. Inter-subnet forwarding for prefix routes in the interface-less IP-VRF-to-IP-VRF model [RFC9136].

c. Inter-subnet forwarding for prefix routes when the ESI is used exclusively as an L3 construct [RFC9136].

1.1. Ethernet Segments for Host Routes in Symmetric IRB

Consider a pair of multi-homing PEs, PE1 and PE2, as illustrated in Figure 1. Let there be a host H1 attached to them. Consider PE3 and a host H3 attached to it.

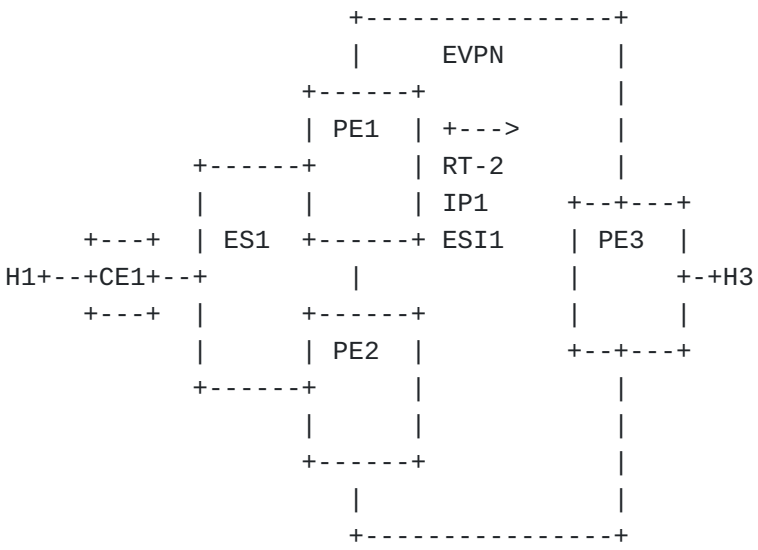


Figure 1: Inter-subnet traffic between Multihoming PEs and Remote PE

With Asymmetric IRB [RFC9135], if H3 sends inter-subnet traffic to H1, routing will happen at PE3. PE3 will be attached to the destination IRB interface and will trigger ARP/ND requests if it does not have an ARP/ND adjacency to H1. A subsequent routing lookup will resolve the destination MAC to H1's MAC address. Furthermore, H1's MAC will point to an ECMP EVPN destination on PE1 and PE2, either due to host route advertisement from both PE1 and PE2, or due to Ethernet Segment MAC Aliasing as detailed in [RFC7432].

With Symmetric IRB [RFC9135], if H3 sends inter-subnet traffic to H1, a routing lookup will happen at PE3's IP-VRF and this routing lookup will not yield the destination IRB interface and therefore MAC Aliasing is not possible. In order to have per-flow load balancing for H3's routed traffic to H1, an IP ECMP list (to PE1/PE2) needs to be associated to H1's host route in the IP-VRF route-table. If H1 is locally learned only at one of the multi-homing PEs, PE1 or PE2, due to LAG hashing, PE3 will not be able to build an IP ECMP list for the H1 host route.

With the extension described in this document, PE3's IP-VRF becomes Ethernet-Segment-aware and builds an IP ECMP list for H1 based on the advertisement of ES1 along with H1 in a MAC/IP route and the availability of ES1 on PE1 and PE2.

1.2. Inter-subnet Forwarding for Prefix Routes in the Interface-less IP-VRF-to-IP-VRF Model

In the Interface-less IP-VRF-to-IP-VRF model described in [RFC9136] there is no Overlay Index and hence no recursive resolution of the IP Prefix route to either a MAC/IP Advertisement or an Ethernet A-D per ES/EVI route, which means that the fast convergence and aliasing/backup path functions are disabled. The recursive resolution of an IP Prefix route to an Ethernet A-D per ES/EVI route is already described in [RFC9136].

The scenario illustrated in Figure 2 will be used to explain the procedures.

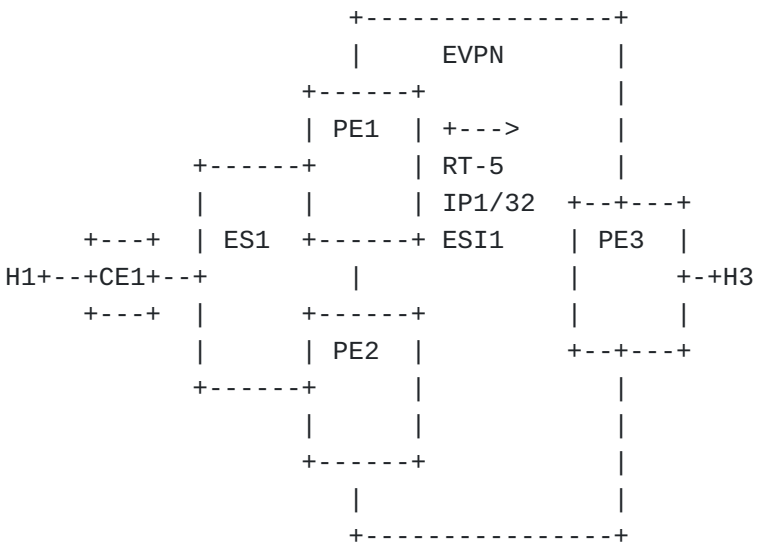


Figure 2: Inter-subnet example with IP Prefix routes

Consider PE1 and PE2 are multi-homed to CE1 (in an All-Active Ethernet Segment ES1), and PE1, PE2 and PE3 are attached to an IP-VRF of the same tenant. Suppose H1's host route is learned (via ARP or ND snooping) on PE1 only, and PE1 advertises an EVPN IP Prefix route for H1's host route. If H3 sends inter-subnet traffic to H1, a routing lookup on PE3 would normally yield a single next-hop, i.e., PE1.

This document proposes the use of the ESI in the IP Prefix route and the recursive resolution to A-D per ES/EVI routes advertised from PE1 and PE2, so that H1's host route in PE3 can be associated to an IP ECMP list (to PE1/PE2) for aliasing purposes.

1.3. Ethernet Segments for Prefix routes in IP-VRF-to-IP-VRF use-cases

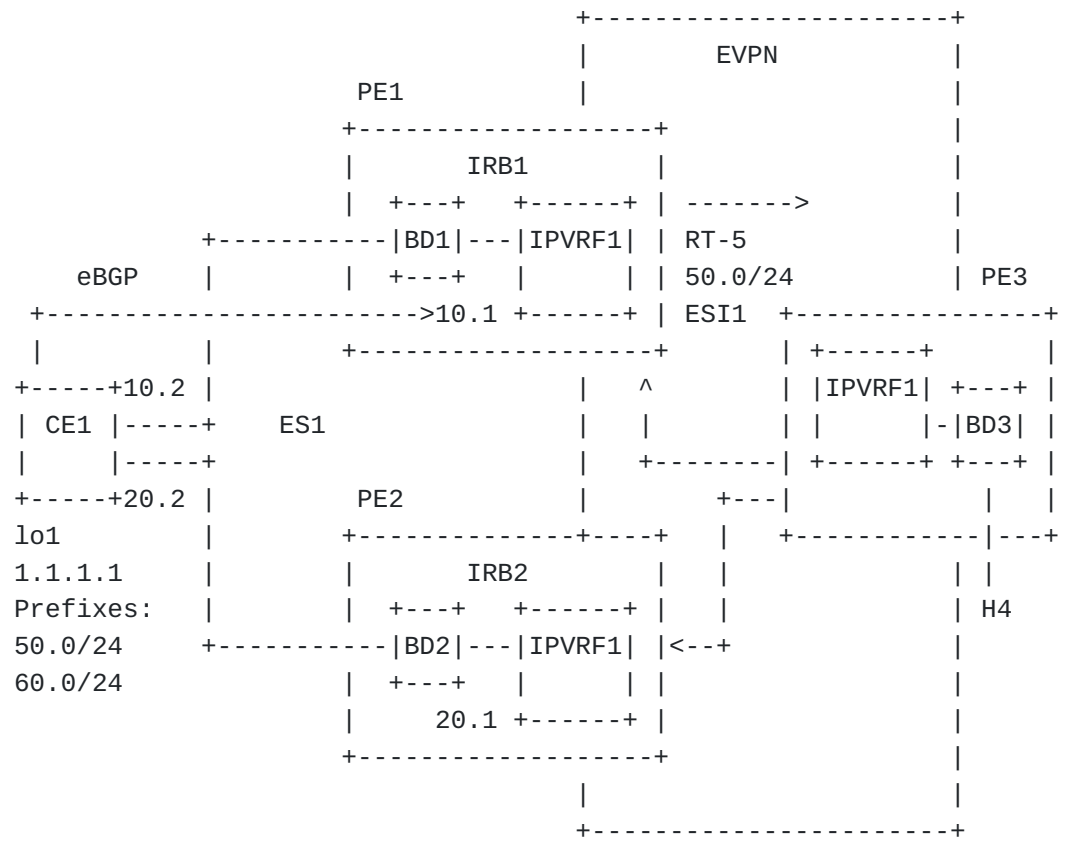
This document also enables fast convergence and aliasing/backup path to be used even when the ESI is used exclusively as an L3 construct, in an Interface-less IP-VRF-to-IP-VRF scenario [RFC9136]. There are two use cases analyzed and supported by this document:

- *IP Aliasing for EVPN IP Prefix routes

- *Centralized Routing Model

1.3.1. IP Aliasing for EVPN IP Prefix routes

As an example, consider the scenario in Figure 3 in which PE1 and PE2 are multi-homed to CE1. However, and contrary to CE1 in Figure 2, in this case the links between CE1 and PE1/PE2 are used exclusively for L3 protocols and L3 forwarding in different BDs, and a BGP session established between CE1's loopback address and PE1's IRB address.



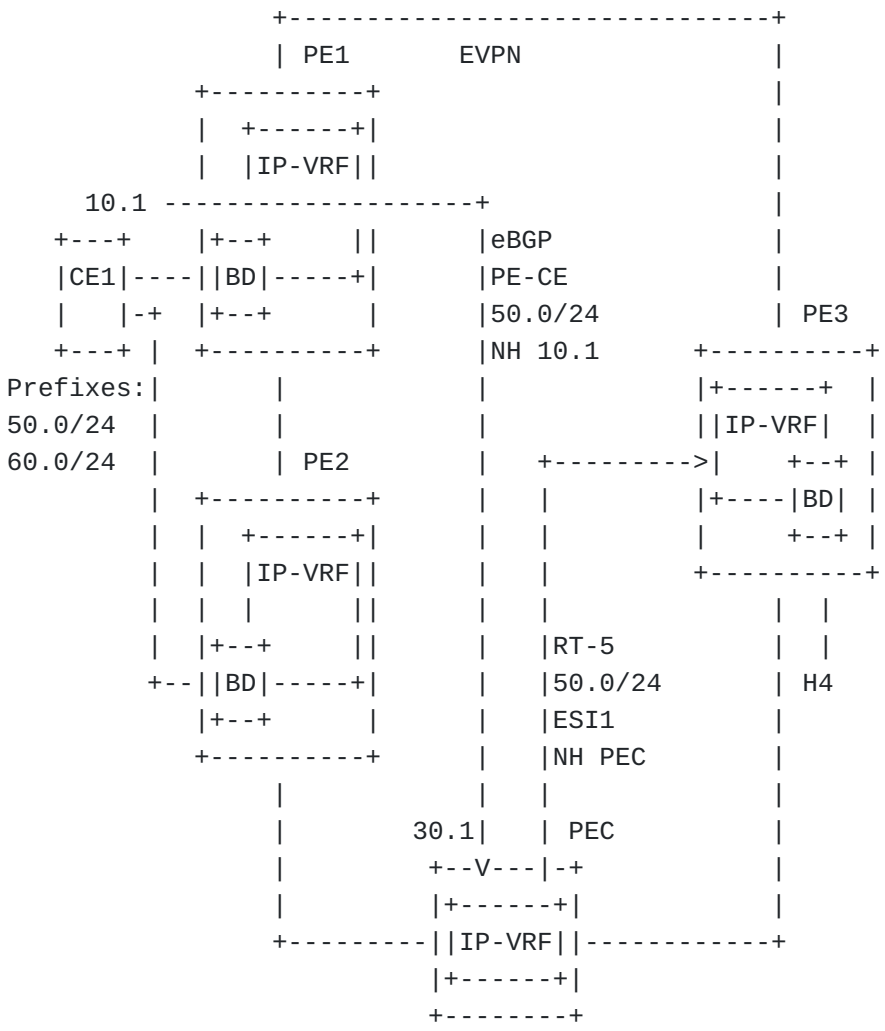
Note:
IP addresses expanded by adding 0s
E.g., 50.0 expands to 50.0.0.0

Figure 3: Layer-3 Multihoming PEs

In these use-cases, sometimes the CE supports a single BGP session to one of the PEs (through which it advertises a number of IP Prefixes seating behind itself) and yet, it is desired that remote PEs can build an IP ECMP list or backup IP list including all the PEs multi-homed to the same CE. For example, in Figure 3, CE1 has a single eBGP neighbor, i.e., PE1. Load-balancing for traffic from CE1 to H4 can be accomplished by a default route with next-hops PE1 and PE2, however, load-balancing from H4 to any of the prefixes attached to CE1 would not be possible since only PE1 would advertise EVPN IP Prefix routes for CE1's prefixes. This document provides a solution so that PE3 considers PE2 as a next-hop in the IP ECMP list for CE1's prefixes, even if PE2 did not advertise the IP Prefix routes for those prefixes in the first place.

1.3.2. Centralized Routing Model

Figure 4 illustrates a model in which multiple CEs establish an eBGP PE-CE session with a Centralized PE.



Note:
IP addresses expanded by adding 0s
E.g., 50.0 expands to 50.0.0.0

Figure 4: Centralized Routing Model

The CEs in this case are usually VNFs (Virtual Network Function entities) or CNFs (Containerized Network Function entities) and by provisioning the same network parameters on all of them, the operation gets significantly simplified. The configuration on the PEs also gets simplified, since the PE-CE eBGP sessions to the CEs are only configured on a centralized PE. In the diagram, CE1 is one of these VNF/CNFs that sets up a multi-hop eBGP session to the centralized PEC. As an example, CE1 advertises prefix 50.0.0.0/24 with Next Hop 10.0.0.1 (to PEC) via the multi-hop eBGP session. PEC then exports the prefix into a RT-5 route, following the Interface-less IP-VRF-to-IP-VRF model [RFC9136], with Next Hop PEC. When H4 sends traffic to an IP address of the subnet 50.0.0.0/24, the traffic will be forwarded to PEC first, and PEC will then forward to PE1 (or PE2). In other words, this model simplifies the configuration and operation of the CEs, however, it introduces an

inefficiency since traffic needs to go through the Centralized PE (PEC) instead of going directly to the PE(s) attached to the destination CE. The IP Aliasing solution specified in this document overcomes this inefficiency and allows traffic from PE3 to be forwarded directly to PE1 or PE2, without going through PEC.

1.4. Terminology and Conventions

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.

*IRB: Integrated Routing and Bridging

*IRB Interface: Integrated Bridging and Routing Interface. A virtual interface that connects the Bridge Table and the IP-VRF on an NVE.

*BD: Broadcast Domain. An EVI may be comprised of one BD (VLAN-based or VLAN Bundle services) or multiple BDs (VLAN-aware Bundle services).

*Bridge Table: An instantiation of a broadcast domain on a MAC-VRF.

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

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

*MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

*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'.

*IP-VRF: A VPN Routing and Forwarding table for IP routes on an NVE/PE. The IP routes could be populated by any routing protocol, E.g., EVPN, IP-VPN and BGP PE-CE IP address families. An IP-VRF is also an instantiation of a layer 3 VPN in an NVE/PE.

*EVPN IP route: An EVPN IP Prefix route or an EVPN MAC/IP Advertisement route.

*LACP: Link Aggregation Control Protocol.

*PE: Provider Edge device.

*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.

*RT-2: EVPN MAC/IP Advertisement route, as specified in [RFC7432].

*RT-4: EVPN Ethernet Segment route, as specified in [RFC7432].

*RT-5: EVPN IP Prefix route, as specified in [RFC9136].

2. Ethernet Segments for L3 Aliasing/Backup Path and Fast Convergence

The first two use cases described in Section 1 do not require any extensions to the Ethernet Segment definition and both cases support Ethernet Segments as a set of Ethernet links and specified in [RFC7432], or virtual Ethernet Segments as a set of logical links specified in [I-D.ietf-bess-evpn-virtual-eth-segment].

The third use case in Section 1 requires an extension to the way Ethernet Segments are defined and associated. In this case, the Ethernet Segment is a Layer-3 construct characterized as follows:

1. The ES is defined as a set of Layer-3 links to the multi-homed CE and its state MUST be linked to the layer-3 reachability from each multi-homed PE to the CE's loopback address via a non-EVPN route in the PE's IP-VRF.

2. The ESI SHOULD be of type 4 [RFC7432] and set to the router ID of the multi-homed CE.

3. All-active or single-active multi-homing redundancy modes are supported, however, the redundancy mode only affects the procedures in Section 3.

4. PEs attached to the same Layer-3 ES discover each other through the exchange of RT-4 routes (Ethernet Segment routes). DF Election procedures [RFC8584] MAY be used for single-active multi-homing mode.

5. The routes advertised from the multi-homed CE's and installed in the PE's IP-VRF table with the CE's loopback as the next-hop MUST be re-advertised by the PE in EVPN IP Prefix routes with the ESI of the CE. The rest of the EVPN IP Prefix routes fields are set as per the Interface-less model in [RFC9136]. Note that the BGP PE-CE routes advertised by the multi-homed CE are installed in the IP-VRF normally irrespective of the Next Hop being resolved to an EVPN or a non-EVPN route, and they are exported as a RT-5 with the ESI.

In the example depicted in Figure 3, ES1 is defined as the set of layer-3 links that connects PE1 and PE2 to CE1. Its ESI, e.g., ESI-1, is derived as a type 4 ESI using the CE's router ID. ES-1 will be operationally active in the PE as long as CE1's loopback route is installed in the PE's IP-VRF and learned via any routing protocol except for an EVPN route. E.g., an active static route to 1.1.1.1 via next-hop 10.0.0.2 would make the ES operationally active in PE1, and the eBGP routes received from CE1 with next-hop 1.1.1.1 will be re-advertised as RT-5 routes with ESI-1.

In the example illustrated in Figure 4, ES1 is a set of layer-3 links connecting PE1, PE2 and PEC to CE1. ESI-1 is derived as a type 4 ESI using the CE's router ID, as in the previous example. CE1's loopback route (which is associated to ES1) is installed in PE1 and PE2 via non-EVPN route, hence ES1 is operationally active in PE1 and PE2. On PE-C though, CE1's loopback is installed via EVPN IP Prefix route, therefore, as per point 1 in the current section, ES1 is operationally inactive in PEC. As per point 5, this does not prevent PEC from exporting CE1's prefixes into RT-5 routes with ESI-1. However, since ES-1 is operationally inactive in PEC, no IP A-D per EVI routes (Section 3) and no IP A-D per ES routes Section 4 for ESI-1 will be advertised from PEC, preventing PEC from attracting traffic destined to CE1.

3. IP Aliasing and Backup Path

In order to address the use-cases described in Section 1, above, this document proposes that:

1. A PE that is attached to a given ES will advertise a set of one or more Ethernet A-D per ES routes for that ES. Each is termed an 'IP A-D per ES' route and is tagged with the route targets (RTs) for one or more of the IP-VRFs defined on it for that ES; the complete set of IP A-D per ES routes contains the RTs for all of the IP-VRFs defined on it for that ES.

   A remote PE imports an IP A-D per ES route into the IP-VRFs corresponding to the RTs with which the route is tagged. When the complete set of IP A-D per ES routes has been processed, a

remote PE will have imported an IP A-D per ES route into each of the IP-VRFs defined on it for that ES; this enables fast convergence for each of these IP-VRFs.

2. A PE advertises, for this ES, an Ethernet A-D per EVI route for each of the IP-VRFs defined on it. Each is termed an 'IP A-D per EVI' route and is tagged with the RT for a given IP-VRF, and conveys a label that identifies that IP-VRF. A label in this context refers to an MPLS label, a VNI (VXLAN Network Identifier) or a Segment Routing IPv6 SID, depending on the transport being used.

   A remote PE imports an IP A-D per EVI route into the IP-VRF corresponding to the RT with which the route is tagged. The label contained in the route enables aliasing/backup path for the routes in that IP-VRF.

To address the third use-case described in Section 1, where the links between a CE and its multihomed PEs are used exclusively for L3 protocols and L3 forwarding, a PE uses the procedures described in 1) and 2), above.

The processing of the IP A-D per ES and the IP A-D per EVI routes is as defined in [RFC7432] and [RFC8365] except that the fast convergence and aliasing/backup path functions apply to the routes contained in an IP-VRF. In particular, a remote PE that receives an EVPN MAC/IP Advertisement route or an IP Prefix route with a non-reserved ESI and the RT of a particular IP-VRF SHOULD consider it reachable by every PE that has advertised an IP A-D per ES and IP A-D per EVI route for that ESI and IP-VRF.

3.1. Constructing the IP A-D per EVI Route

The construction of the IP A-D per EVI route is the same as that of the Ethernet A-D per EVI route, as described in [RFC7432], with the following exceptions:

- The Route-Distinguisher is for the corresponding IP-VRF.

- The Ethernet Tag should be set to 0.

- The route SHOULD carry the Route Target of the corresponding IP-VRF.

- The route MUST carry the MPLS label, VNI (VXLAN or Virtual Network Identifier as in [RFC8365]) or Segment Routing IPv6 SID (Segment Identifier [RFC9252]) that identifies the corresponding IP-VRF. In case of Segment Routing IPv6 (SRv6), the Service SID (Segment Identifier) is enclosed in an SRv6 Service TLV of type L3 within the BGP Prefix-SID attribute, where the SRv6 Endpoint

Behavior SHOULD be one of these: End.DT46, End.DT4, End.DT6, End.DX4, or End.DX6 [RFC9252].

* The route MUST carry the PE's MAC Extended Community if the encapsulation used between the PEs for inter-subnet forwarding is an Ethernet NVO tunnel [RFC9136].

* The route SHOULD carry the EVPN Layer 2 Extended Community [I-D.ietf-bess-rfc7432bis]. For all-active multihoming, all PEs attached to the specified ES will advertise P=1. For backup path, the Primary PE will advertise P=1 and the Backup PE will advertise P=0, B=1.

  - The Primary PE SHOULD be a PE with a routing adjacency to the attached CE.

  - The Primary PE MAY be determined by policy or MAY be elected by a DF Election as in [RFC8584] as described in Section 2.

4. Fast Convergence for Routed Traffic

Host or Prefix reachability is learned via the BGP-EVPN control plane over the MPLS/NVO network. EVPN IP routes for a given ES are advertised by one or more of the PEs attached to that ES. When one of these PEs fails, a remote PE needs to quickly invalidate the EVPN IP routes received from it.

To accomplish this, EVPN defined the fast convergence function specified in [RFC7432]. This document extends fast convergence to inter-subnet forwarding by having each PE advertise a set of one or more IP A-D per ES routes for each locally attached Ethernet segment (refer to Section 4.1 below for details on how these routes are constructed). A PE may need to advertise more than one IP A-D per ES route for a given ES because the ES may be in a multiplicity of IP-VRFs and the Route Targets for all of these IP-VRFs may not fit into a single route. Advertising a set of IP A-D per ES routes for the ES allows each route to contain a subset of the complete set of Route Targets. Each IP A-D per ES route is differentiated from the other routes in the set by a different Route Distinguisher (RD).

Upon failure in connectivity to the attached ES, the PE withdraws the corresponding set of IP A-D per ES routes. This triggers all PEs that receive the withdrawal to update their next-hop adjacencies for all IP addresses associated with the Ethernet Segment in question, across IP-VRFs. If no other PE has advertised an IP A-D per ES route for the same Ethernet Segment, then the PE that received the withdrawal simply invalidates the IP entries for that segment. Otherwise, the PE updates its next-hop adjacencies accordingly.

These routes should be processed with higher priority than EVPN IP route withdrawals upon failure. Similar priority processing is needed even on the intermediate Route Reflectors.

4.1. Constructing IP A-D per Ethernet Segment Route

This section describes the procedures used to construct the IP A-D per ES route, which is used for fast convergence (as discussed in Section 4). The usage/construction of this route remains similar to that described in section 8.2.1. of [RFC7432] with a few notable exceptions as explained in following sections.

4.1.1. IP A-D per ES Route Targets

Each IP A-D per ES route MUST carry one or more Route Targets. The set of IP A-D per ES routes MUST carry the entire set of IP-VRF Route Targets for all the IP-VRFs defined on that ES.

4.1.2. IP A-D per ES SRv6 Transport

When an SRv6 transport is used, each IP A-D per ES route MUST carry an SRv6 L3 Service TLV within the BGP Prefix-SID attribute [RFC9252]. The Service SID MUST be of value 0. The SRv6 Endpoint Behavior SHOULD be one of these End.DT46, End.DT4, End.DT6, End.DX4, or End.DX6.

4.2. Avoiding convergence issues by synchronizing IP prefixes

Consider a pair of multi-homing PEs, PE1 and PE2. Let there be a host H1 attached to them. Consider PE3 and a host H3 attached to it.

If the host H1 is learned on both the PEs, the ECMP path list is formed on PE3 pointing to (PE1/PE2). Traffic from H3 to H1 is not impacted even if one of the PEs fails as the path list gets corrected upon receiving the withdrawal of the fast convergence route(s) (IP A-D per ES routes).

In a case where H1 is locally learned only on PE1 due to LAG hashing or a single routing protocol adjacency to PE1, at PE3, H1 has ECMP path list (PE1/PE2) using Aliasing as described in this document. Traffic from H3 can reach H1 via either PE1 or PE2.

PE2 should install local forwarding state for EVPN IP routes advertised by other PEs attached to the same ES (i.e., PE1) but not advertise them as local routes. When the traffic from H3 reaches PE2, PE2 will be able forward the traffic to H1 without any convergence delay (caused by triggering ARP/ND to H1 or to the next-hop to reach H1). The synchronization of the EVPN IP routes across all PEs of the same Ethernet Segment is important to solve convergence issues.

4.3. Handling Silent Host MAC/IP route for IP Aliasing

Consider the example of Figure 1 for IP aliasing. If PE1 fails, PE3 will receive the withdrawal of the fast convergence route(s) and update the ECMP list for H1 to be just PE2. When the EVPN IP route for H1 is also withdrawn, neither PE2 nor PE3 will have a route to H1, and traffic from H3 to H1 is blackholed until PE2 learns H1 and advertises an EVPN IP route for it.

This blackholing can be much worse if the H1 behaves like a silent host. IP address of H1 will not be re-learned on PE2 till H1 ARP/ND messages or some traffic triggers ARP/ND for H1.

PE2 can detect the failure of PE1's reachability in different ways:

a. When PE1 fails, the next hop tracking to PE1 in the underlay routing protocols can help detect the failure.

b. Upon the failure of its link to CE1, PE1 will withdraw its IP A-D route(s) and PE2 can use this as a trigger to detect failure.

Thus to avoid blackholing, when PE2 detects loss of reachability to PE1, it should trigger ARP/ND requests for all remote IP prefixes received from PE1 across all affected IP-VRFs. This will force host H1 to reply to the solicited ARP/ND messages from PE2 and refresh both MAC and IP for the corresponding host in its tables.

Even in core failure scenario on PE1, PE1 must withdraw all its local layer-2 connectivity, as Layer-2 traffic should not be received by PE1. So when ARP/ND is triggered from PE2 the replies from host H1 can only be received by PE2. Thus H1 will be learned as local route and also advertised from PE2.

It is recommended to have a staggered or delayed deletion of the EVPN IP routes from PE1, so that ARP/ND refresh can happen on PE2 before the deletion.

4.4. MAC Aging

In the same example as in Section 4.3, PE1 would do ARP/ND refresh for H1 before it ages out. During this process, H1 can age out genuinely or due to the ARP/ND reply landing on PE2. PE1 must withdraw the local entry from BGP when H1 entry ages out. PE1 deletes the entry from the local forwarding only when there are no remote synced entries.

5. Determining Reachability to Unicast IP Addresses

5.1. Local Learning

The procedures for local learning do not change from [RFC7432] or [RFC9136].

5.2. Remote Learning

The procedures for remote learning do not change from [RFC7432] or [RFC9136].

5.3. Constructing the EVPN IP Routes

The procedures for constructing MAC/IP Address or IP Prefix Advertisements do not change from [RFC7432] or [RFC9136].

5.3.1. Route Resolution

If the ESI field is set to reserved values of 0 or MAX-ESI, the EVPN IP route resolution MUST be based on the EVPN IP route alone.

If the ESI field is set to a non-reserved ESI, the EVPN IP route resolution MUST happen only when both the EVPN IP route and the associated set of IP A-D per ES routes have been received. To illustrate this with an example, consider a pair of multi-homed PEs, PE1 and PE2, connected to an all-active Ethernet Segment. A given host with IP address H1 is learned by PE1 but not by PE2. When the EVPN IP route from PE1 and a set of IP A-D per ES and IP A-D per EVI routes from PE1 and PE2 are received, then (1) PE3 can forward traffic destined to H1 to both PE1 and PE2.

If after (1) PE1 withdraws the IP A-D per ES route, then PE3 will forward the traffic to PE2 only.

If after (1) PE2 withdraws the IP A-D per ES route, then PE3 will forward the traffic to PE1 only.

If after (1) PE1 withdraws the EVPN IP route, then PE3 will do delayed deletion of H1, as described in Section 4.3.

If after (1) PE2 advertised the EVPN IP route, but PE1 withdraws it, PE3 will continue forwarding to both PE1 and PE2 as long as it has the IP A-D per ES and the IP A-D per EVI route from both.

6. Forwarding Unicast Packets

Refer to Section 5 in [RFC9135] and [RFC9136].

7. Load Balancing of Unicast Packets

The procedures for load balancing of Unicast Packets do not change from [RFC7432]

8. IP Aliasing and Unequal ECMP for IP Prefix Routes

[I-D.ietf-bess-evpn-unequal-lb] specifies the use of the EVPN Link bandwidth extended community to achieve weighted load balancing to an ES or Virtual ES for unicast traffic. The procedures in [I-D.ietf-bess-evpn-unequal-lb] MAY be used along with the procedures described in this document for any of the three cases described in Section 1, with the following considerations:

* The ES weight is signaled by the multi-homed PEs in the IP A-D per ES routes.

* The remote ingress PE learning an EVPN IP Route to prefix/host P that is associated to a weighted load balancing ES, will follow the procedures in [I-D.ietf-bess-evpn-unequal-lb] to influence the load balancing for traffic to P.

* [I-D.ietf-bess-evpn-unequal-lb] also allows the use of the EVPN Link Bandwidth Extended Community along with RT-5s. If the ingress PE learns a prefix P via a non-reserved ESI RT-5 route with a weight (for which IP A-D per ES routes also signal a weight) and a zero ESI RT-5 that includes a weight, the ingress PE will consider all the PEs attached to the ES as a single PE when normalizing weights.

  As an example, consider PE1 and PE2 are attached to ES-1 and PE1 advertises an RT-5 for prefix P with ESI-1 (and EVPN Link Bandwidth of 1). Consider PE3 advertises an RT-5 for P with ESI=0 and EVPN Link Bandwidth of 2. If PE1 and PE2 advertise an EVPN Link Bandwidth of 1 and 2, respectively, in the IP A-D per ES routes for ES-1, an ingress PE4 SHOULD assign a normalized weight of 1 to ES-1 and a normalized weight of 2 to PE3. When PE4 sprays the flows to P, it will send twice as many flows to PE3. For the flows sent to ES-1, the individual PE EVPN Link Bandwidths advertised in the IP A-D per ES routes will be considered.

9. Security Considerations

The mechanisms in this document use EVPN control plane as defined in [RFC7432]. Security considerations described in [RFC7432] are equally applicable. 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.

10. IANA Considerations

This document does not require any actions from IANA.

11. Contributors

12. Acknowledgments

The authors would like to thank Kiran Nagaraj, Mallika Gautam, Senthil Sathappan and Sasha Vainshtein for their comments and feedback.

13. References

13.1. Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., Henderickx, W., and RFC Publisher, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[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>.

[RFC2119] Bradner, S. and RFC Publisher, "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. and RFC Publisher, "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>.

[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>.

[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>.

[I-D.ietf-bess-rfc7432bis] Sajassi, A., Burdet, L. A., Drake, J., and J. Rabadan, "BGP MPLS-Based Ethernet VPN", Work in Progress, Internet-Draft, draft-ietf-bess-rfc7432bis-06, 5 January 2023, <https://www.ietf.org/archive/id/draft-ietf-bess-rfc7432bis-06.txt>.

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI 10.17487/RFC9252, July 2022, <https://www.rfc-editor.org/info/rfc9252>.

13.2. Informative References

[I-D.ietf-bess-evpn-virtual-eth-segment]
Sajassi, A., Brissette, P., Schell, R., Drake, J., and J. Rabadan, "EVPN Virtual Ethernet Segment", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-virtual-eth-segment-07, 6 July 2021, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-virtual-eth-segment-07.txt>.

[I-D.ietf-bess-evpn-unequal-lb]
Malhotra, N., Sajassi, A., Rabadan, J., Drake, J., Lingala, A., and S. Thoria, "Weighted Multi-Path Procedures for EVPN Multi-Homing", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-unequal-lb-17, 28 November 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-unequal-lb-17.txt>.

Authors' Addresses

A. Sajassi (editor)
Cisco Systems

Email: sajassi@cisco.com

G. Badoni
Cisco Systems

Email: gbadoni@cisco.com

P. Warade
Cisco Systems

Email: pwarade@cisco.com

S. Pasupula
Cisco Systems

Email: surpasup@cisco.com

L. Krattiger
Cisco Systems

Email: lkrattig@cisco.com

J. Drake (editor)
Juniper

Email: jdrake@juniper.net

J. Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America

Email: jorge.rabadan@nokia.com