BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2026

J. Rabadan, Ed.
K. Nagaraj
Nokia
A. Nichol
Arista
A. Sajassi
Cisco Systems
W. Lin
Juniper Networks
J. Tantsura
Nvidia
2 March 2026

EVPN Anycast Multi-Homing
draft-rabnag-bess-evpn-anycast-aliasing-05

Abstract

The current Ethernet Virtual Private Network (EVPN) all-active multi-homing procedures in Network Virtualization Over Layer-3 (NVO3) networks provide the required Split Horizon filtering, Designated Forwarder Election and Aliasing functions that the network needs in order to handle the traffic to and from the multi-homed CE in an efficient way. In particular, the Aliasing function supports load balancing of unicast traffic from remote Network Virtualization Edge (NVE) devices to NVEs that are multi-homed to the same CE, regardless of whether the CE's MAC/IP information has been learned on all those NVEs. This document introduces an optional enhancement to the EVPN multi-homing Aliasing function, referred to as EVPN Anycast Multi-homing. This optimization is specific to EVPN deployments over NVO3 tunnels (i.e., IP-based tunnels) and may offer benefits in typical data center designs, which are discussed herein.

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 3 September 2026.

Table of Contents

1. Introduction

Ethernet Virtual Private Network (EVPN) is the de-facto standard control plane in Network Virtualization Over Layer-3 (NVO3) networks deployed in multi-tenant Data Centers [RFC8365][RFC9469]. EVPN enables Network Virtualization Edge (NVE) auto-discovery, tenant MAC/IP dissemination, and advanced capabilities required by Network Virtualization over Layer 3 (NVO3) networks, such as all-active multi-homing. The current EVPN all-active multi-homing procedures in NVO3 networks provide the required Split Horizon filtering, Designated Forwarder Election and Aliasing functions that the network needs in order to handle the traffic to and from the multi-homed CE in an efficient way. In particular, the Aliasing function supports load balancing of unicast traffic from remote NVEs to NVEs that are multi-homed to the same CE, regardless of whether the CE's MAC/IP information has been learned on all those NVEs. This document introduces an optional enhancement to the EVPN multi-homing Aliasing function, referred to as EVPN Anycast Multi-homing. This optimization is specific to EVPN deployments over NVO3 tunnels (i.e., IP-based tunnels) and may offer benefits in typical data center designs, which are discussed herein.

1.1. 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.

* A-D per EVI route: EVPN route type 1, Auto-Discovery per EVPN Instance route. Route used for aliasing or backup signaling in EVPN multi-homing procedures [RFC7432].

* A-D per ES route: EVPN route type 1, Auto-Discovery per Ethernet Segment route. Route used for mass withdraw in EVPN multi-homing procedures [RFC7432].

* BUM traffic: Broadcast, Unknown unicast and Multicast traffic.

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

* Clos: a multistage network topology described in [CLOS1953], where all the edge nodes (or Leaf routers) are connected to all the core nodes (or Spines). Typically used in Data Centers.

* ECMP: Equal Cost Multi-Path.

* ES: Ethernet Segment. When a Tenant System (TS) is connected to one or more NVEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'. Each ES is represented by a unique Ethernet Segment Identifier (ESI) in the NVO3 network and the ESI is used in EVPN routes that are specific to that ES.

* EVI: or EVPN Instance. It is a Layer-2 Virtual Network that uses an EVPN control-plane to exchange reachability information among the member NVEs. It corresponds to a set of MAC-VRFs of the same tenant. See MAC-VRF in this section.

* GENEVE: Generic Network Virtualization Encapsulation, an NVO3 encapsulation defined in [RFC8926].

* IP-VRF: an IP Virtual Routing and Forwarding table, as defined in [RFC4364]. It stores IP Prefixes that are part of the tenant's IP space, and are distributed among NVEs of the same tenant by EVPN. Route Distinguisher (RD) and Route Target(s) (RTs) are required properties of an IP-VRF. An IP-VRF is instantiated in an NVE for a given tenant, if the NVE is attached to multiple subnets of the tenant and local inter-subnet-forwarding is required across those subnets.

* IRB: Integrated Routing and Bridging interface. It refers to the logical interface that connects a Broadcast Domain instance (or a BT) to an IP-VRF and allows to forward packets with destination in a different subnet.

* MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [RFC7432]. The instantiation of an EVI (EVPN Instance) in an NVE. Route Distinguisher (RD) and Route Target(s) (RTs) are required properties of a MAC-VRF and they are normally different from the ones defined in the associated IP-VRF (if the MAC-VRF has an IRB interface).

* MPLS and non-MPLS NVO3 tunnels: refer to Multi-Protocol Label Switching (or the absence of it) Network Virtualization Overlay tunnels. Network Virtualization Overlay tunnels use an IP encapsulation for overlay frames, where the source IP address identifies the ingress NVE and the destination IP address the egress NVE.

* NLRI: BGP Network Layer Reachability Information.

* NVE: Network Virtualization Edge device, a network entity that sits at the edge of an underlay network and implements Layer-2 and/or Layer-3 network virtualization functions. The network-

facing side of the NVE uses the underlying Layer-3 network to tunnel tenant frames to and from other NVEs. The tenant-facing side of the NVE sends and receives Ethernet frames to and from individual Tenant Systems. In this document, an NVE could be implemented as a virtual switch within a hypervisor, a switch or a router, and runs EVPN in the control-plane. This document uses the terms NVE and "Leaf router" interchangeably.

* NVO3 tunnels: Network Virtualization Over Layer-3 tunnels. In this document, NVO3 tunnels refer to a way to encapsulate tenant frames or packets into IP packets whose IP Source Addresses (SA) or Destination Addresses (DA) belong to the underlay IP address space, and identify NVEs connected to the same underlay network. Examples of NVO3 tunnel encapsulations are VXLAN [RFC7348], GENEVE [RFC8926] or MPLSoUDP [RFC7510].

* SBD: Supplementary Broadcast Domain [RFC9136].

* SRv6: Segment routing with an IPv6 data plane, [RFC8986].

* TS: Tenant System. A physical or virtual system that can play the role of a host or a forwarding element such as a router, switch, firewall, etc. It belongs to a single tenant and connects to one or more Broadcast Domains of that tenant.

* VNI: Virtual Network Identifier. Irrespective of the NVO3 encapsulation, the tunnel header always includes a VNI that is added at the ingress NVE (based on the mapping table lookup) and identifies the BT at the egress NVE. This VNI is called VNI in VXLAN or GENEVE, VSID in nvGRE or Label in MPLSoGRE or MPLSoUDP. This document will refer to VNI as a generic Virtual Network Identifier for any NVO3 encapsulation.

* VTEP: VXLAN Termination End Point. A loopback IP address of the destination NVE that is used in the outer destination IP address of VXLAN packets directed to that NVE.

* VXLAN: Virtual eXtensible Local Area Network, an NVO3 encapsulation defined in [RFC7348].

1.2. Problem Statement

Figure 1 depicts the typical Clos topology in multi-tenant Data Centers, only simplified to show three Leaf routers and two Spines, forming a 3-stage Clos topology. The NVEs or Leaf routers run EVPN for NVO3 tunnels, as in [RFC8365]. This document assumes VXLAN is used as the NVO3 tunnel encapsulation, given its widespread adoption in multi-tenant data center environments. The diagram below serves

as a reference throughout the document. It is important to note that in large-scale data center deployments, the number of Tenant Systems, leaf routers, and spine layers may be significantly higher than what is depicted in Figure 1.

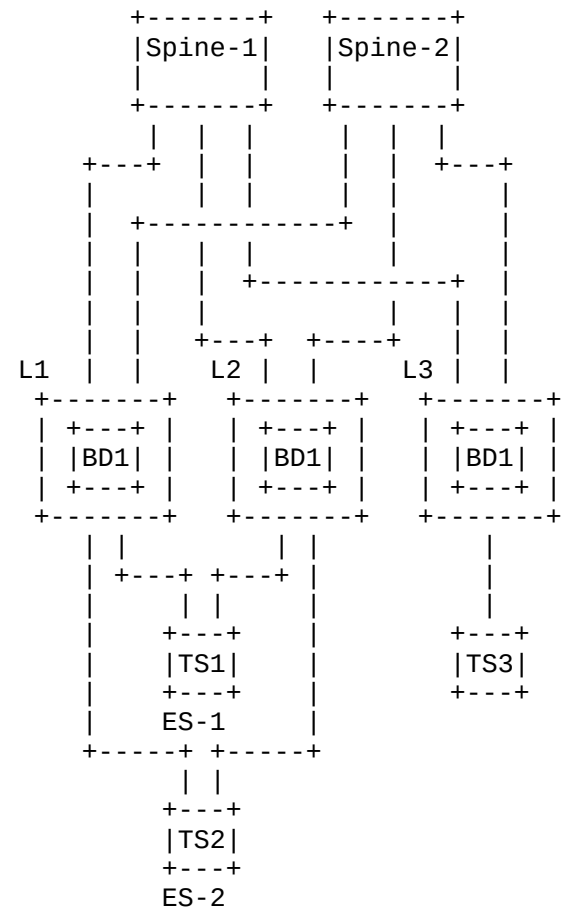


Figure 1: Simplified Clos topology in Data Centers

In the example of Figure 1 the Tenant Systems TS1 and TS2 are multi-homed to Leaf routers L1 and L2, and Ethernet Segments Identifiers ESI-1 and ESI-2 are the representation of TS1 and TS2 Ethernet Segments in the EVPN control plane for the Split Horizon filtering, Designated Forwarder and Aliasing functions [RFC8365].

Taking Tenant Systems TS1 and TS3 as an example, the EVPN all-active multi-homing procedures guarantee that, when TS3 sends unicast traffic to TS1, Leaf L3 does per-flow load balancing towards Leaf

routers L1 and L2. As explained in [RFC7432] and [RFC8365] this is possible due to L1 and/or L2 Leaf routers advertising TS1's MAC address in an EVPN MAC/IP Advertisement route that includes ESI-1 in the Ethernet Segment Identifier field. When the route is imported into Leaf L3, TS1's MAC address is programmed with a destination associated with the ESI-1 next-hop list. This ESI-1 next-hop list is created based on the reception of the EVPN A-D per ES and A-D per EVI routes for ESI-1 that are received from Leaf routers L1 and L2.

Assuming the Ethernet Segment ES-1 links are operationally active, Leaf routers L1 and L2 advertise the EVPN A-D per ES/EVI routes for ESI-1. Leaf L3 then adds L1 and L2 to its next-hop list for ESI-1. As a result, unicast flows from TS3 to TS1 are load-balanced across L1 and L2. This ESI-1 next-hop list in Leaf L3 is referred to as the "overlay ECMP set" for ESI-1. In addition, once Leaf L3 selects one of the next hops in the overlay ECMP set (e.g., L1), it performs a route lookup for L1's address in the base router route table. This lookup yields a list of two next hops — Spine-1 and Spine-2 — which is referred to as the "underlay ECMP set." Therefore, for any given unicast flow to TS1, Leaf L3 performs per-flow load balancing at two levels: it first selects a next hop from the overlay ECMP set (e.g., L1), and then selects a next hop from the underlay ECMP set (e.g., Spine-1).

While aliasing [RFC7432] offers an efficient way to load balance unicast traffic across Leaf routers attached to the same all-active Ethernet Segment, it introduces challenges in very large data centers where the number of Ethernet Segments and Leaf routers is substantial:

a. Control Plane Scale: In a large data center environment, the number of multi-homed compute nodes can grow substantially into the thousands. Each compute node requires a unique Ethernet Segment (ES) and hosts dozens of EVIs per ES. Under the aliasing model defined in [RFC7432], there is a requirement to advertise EVPN A-D per EVI routes for every active EVI on each Ethernet Segment. As a result, the volume of EVPN state that Route Reflectors, Data Center Gateways, and Leaf routers must process becomes significant, and it only increases as additional Ethernet Segments, Broadcast Domains, and Leaf routers are deployed. Eliminating the need to advertise these EVPN A-D per EVI routes would therefore provide a substantial benefit in reducing overall route scale and processing overhead.

b. Convergence and Processing overhead: In accordance with [RFC8365] each node in an Ethernet Segment operates as an independent VTEP and therefore acts as a separate EVPN next hop. In a typical data center leaf-spine topology, this results in ECMP being

applied both in the underlay ECMP set and in the overlay ECMP set. As a consequence, convergence at scale during a failure can be slow and CPU intensive. All leaf routers must process the overlay state changes triggered by the withdrawal of EVPN route(s) at the point of failure and update their overlay ECMP sets accordingly. By performing load balancing solely within the underlay ECMP set, it is possible to significantly reduce this network-wide state churn and processing overhead. This approach also enables faster convergence at scale by limiting re-convergence to only the intermediate spine nodes.

c. Inefficient underlay forwarding during a failure: Another consequence of using ECMP with the overlay ECMP set is the potential for in-flight packets sent by remote leaf routers to be rerouted inefficiently. For example, suppose the link between L1 and Spine-1 (shown in Figure 1) fails. In-flight VXLAN packets already sent from L3 with the destination VTEP set to L1 arrive at Spine-1 and are rerouted along a suboptimal path — for instance, through L2 -> Spine-2 -> L1 -> TS1 — even though they could have been forwarded directly via L2 -> TS1, since TS1 is also connected to Leaf L2. Once the underlay routing protocol converges, all VXLAN packets destined for VTEP L1 are correctly forwarded to Spine-2, and Leaf L3 removes Spine-1 from the underlay ECMP set for Leaf L1.

There are existing proprietary multi-chassis Link Aggregation Group implementations, collectively and commonly known as MC-LAG, that attempt to address the challenges described above by using the concept of "Anycast VTEPs". This involves assigning a shared loopback IP address that the leaf routers connected to the same multi-homed tenant system use to terminate VXLAN packets. For example, in Figure 1, if Leaf routers L1 and L2 were to use an Anycast VTEP address (e.g., anycast-IP1), they could identify VXLAN packets destined for multi-homed tenant systems using that shared address:

* Leaf L3 would not need to create an overlay ECMP set for packets destined to TS1 or TS2, since the use of anycast-IP1 in the underlay ECMP set guarantees per-flow load balancing across the two leaf routers.

* In the same failure scenario described earlier — where the link between L1 and Spine-1 fails — Spine-1 would reroute VXLAN packets directly to Leaf L2. This is possible because L2 also advertises the anycast-IP1 address that Leaf L3 uses to forward packets to TS1 or TS2.

* Additionally, if Leaf routers L1 and L2 used proprietary MC-LAG techniques, no EVPN A-D per EVI routes would be required. As a result, the number of EVPN routes would be significantly reduced in a large-scale data center.

However, the use of proprietary MC-LAG technologies in EVPN NVO3 networks is being abandoned due to the superior capabilities offered by EVPN Multi-Homing. These include features such as mass withdraw [RFC7432], advanced Designated Forwarding election [RFC8584] or weighted load balancing [I-D.ietf-bess-evpn-unequal-lb], among others.

1.3. Solution Overview

This document specifies an EVPN Anycast Multi-Homing extension that can be used as an alternative to EVPN aliasing ([RFC7432]). The EVPN Anycast Multi-Homing procedures described here may optionally replace per-flow overlay ECMP load balancing with simplified per-flow underlay ECMP load balancing. This approach works similarly to proprietary MC-LAG solutions but provides a standardized method that retains the superior advantages of EVPN Multi-Homing — such as Designated Forwarder Election, Split Horizon filtering, and the mass withdraw function (all described in [RFC8365] and [RFC7432]).

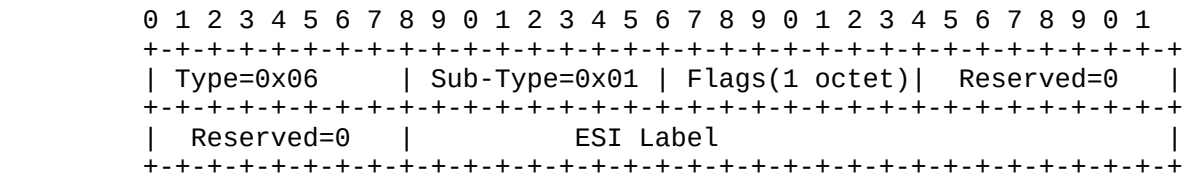
The solution uses A-D per ES routes to advertise the Anycast VTEP address to be used when sending traffic to the Ethernet Segment, and it suppresses the use of A-D per EVI routes for Ethernet Segments configured in this mode. This design addresses the challenges outlined in Section 1.2.

The solution is applicable to all NVO3 tunnels and even to IP tunnels in general. While VXLAN is often used as an example in this document due to its widespread adoption in multi-tenant data centers, the examples and procedures are equally valid for any NVO3 or IP tunnel type.

2. BGP EVPN Extensions

This specification makes use of two BGP extensions that are used along with some EVPN routes.

The first extension is the flag "A" or "Anycast Multi-homing mode" and it is requested to IANA to be allocated in bit 2 of the EVPN ESI Multihoming Attributes registry for the 1-octect Flags field in the ESI Label Extended Community, as follows:



Flags field:

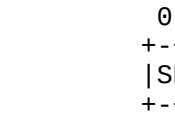


Figure 2: ESI Label Extended Community and Flags

Where the following Flags are defined:

Name	Meaning	Reference
RED	Multihomed redundancy mode	[RFC9746]
SHT	Split Horizon type	[RFC9746]
A	Anycast Multi-homing mode	This document

Table 1: Flags Field

When the NVE advertises an A-D per ES route with the A flag set, it indicates the Ethernet Segment is working in Anycast Multi-homing mode. The A flag is set only if the RED = 00 (All-Active redundancy mode), and MUST NOT be set if RED is different from 00.

The second extension that this document introduces is the encoding of the "Anycast VTEP" address in the BGP Tunnel Encapsulation Attribute, Tunnel Egress Endpoint Sub-TLV (code point 6) [RFC9012], that is advertised along with:

* The EVPN A-D per ES routes for an Ethernet Segment working in Anycast Multi-homing mode or

* The EVPN IP Prefix routes for multi-homed IP prefixes.

Refer to [RFC9012] for the error handling procedures related to the BGP Tunnel Encapsulation Attribute.

When the BGP Tunnel Encapsulation Attribute and the Tunnel Egress Endpoint Sub-TLV are advertised along with EVPN routes for the purpose of this specification, the BGP Encapsulation extended community [RFC9012] is also advertised with the EVPN routes.

For NVO3 tunnel types (e.g., VXLAN, GENEVE), the 'Anycast VTEP' MUST be encoded in the BGP Tunnel Encapsulation Attribute and advertised with the A-D per ES or IP Prefix routes. However, when using SRv6 encapsulation, the BGP Tunnel Encapsulation Attribute is not applicable. Refer to Section 6 for details about SRv6.

3. Anycast Multi-Homing Solution

This document proposes an optional "EVPN Anycast Multi-homing" procedure that provides a solution to optimize network behavior if the challenges described in Section 1.2 become significant. The description uses the terms "Ingress NVE" and "Egress NVE". In this document, Egress NVE refers to an NVE that is attached to a group of Ethernet Segments operating in Anycast Multi-homing mode. Ingress NVE refers to the NVE transmitting unicast traffic to a MAC address associated with a remote Ethernet Segment that also operates in Anycast Multi-Homing mode. In addition, the concepts of Unicast VTEP and Anycast VTEP are introduced:

* A Unicast VTEP is a loopback IP address unique within the data center fabric and owned by a single NVE that terminates VXLAN (or other NVO3 or IP tunnel) traffic.

* An Anycast VTEP is a loopback IP address shared among NVEs attached to the same group of Ethernet Segments in Anycast Multi-Homing mode and is used to terminate VXLAN (or other NVO3 or IP tunnel) traffic on those NVEs.

An Anycast VTEP in this document MUST NOT be used as the BGP next hop of any EVPN route NLRI. This restriction is necessary because the Multi-Homing procedures require the originator of EVPN routes to be uniquely identified by their NLRI next hops

The solution consists of the following modifications of the [RFC7432] EVPN Aliasing function:

1. The [RFC8365] Designated Forwarder and Split Horizon filtering procedures remain unmodified. However, the Aliasing procedure is modified in this Anycast Multi-homing mode.

2. The forwarding of BUM traffic and related procedures are not modified by this document. Only the procedures related to the forwarding of unicast traffic to a remote Ethernet Segment are updated.

3. Any two or more Egress NVEs attached to the same Ethernet Segment working in Anycast Multi-homing mode MUST use the same VNI or label to identify the Broadcast Domain associated with that Ethernet Segment. For non-MPLS NVO3 tunnels, using the same VNI is implicit if global VNIs are used ([RFC8365] section 5.1.1). If locally significant values are used for the VNIs, at least all the Egress NVEs sharing Ethernet Segments MUST use the same VNI for the Broadcast Domain. For MPLS NVO3 tunnels, the Egress NVEs sharing Anycast Multi-homing Ethernet Segments MUST use Domain-wide Common Block labels [RFC9573] so that all can be configured with the same unicast label for the same Broadcast Domain. Note that this requirement only affects unicast labels (i.e., the labels advertised in EVPN MAC/IP Advertisement routes) and does not affect the Ingress Replication labels for BUM traffic, which are advertised via EVPN Inclusive Multicast Ethernet Tag routes.

4. The default behavior for an Egress NVE attached to an Ethernet Segment follows [RFC8365]. The Anycast Multi-homing mode MUST be explicitly configured for a given all-active Ethernet Segment. When the Egress NVE Ethernet Segment is configured to follow the Anycast Multi-homing behavior for at least one Ethernet Segment, the egress NVE:

    a. Is configured with an Anycast VTEP address. A single Anycast VTEP address is allocated for all the Anycast Aliasing Ethernet Segments shared among the same group of Egress NVEs. This is the only additional address whose reachability needs to be advertised in the underlay routing protocol. If "m" Egress NVEs are attached to the same "n" Ethernet Segments, all the "m" Egress NVEs advertise the same Anycast VTEP address in the A-D per ES routes for those "n" Ethernet Segments.

    b. Is assumed to advertise reachability for the Anycast VTEP in the underlay routing protocol, either by announcing an exact match route for the Anycast VTEP address (using a /32 mask for IPv4 or a /128 mask for IPv6) or by advertising a shorter prefix that includes the Anycast VTEP IP address.

    c. Advertises EVPN A-D per ES routes for each Ethernet Segment with:

* an "Anycast Multi-homing mode" flag that indicates to the remote NVEs that the EVPN MAC/IP Advertisement routes with matching Ethernet Segment Identifier are resolved by only A-D per ES routes for the Ethernet Segment. In other words, this flag signals to the ingress NVE that no A-D per EVI routes are advertised for the Ethernet Segment.

* an Anycast VTEP that identifies the Ethernet Segment and is encoded in a BGP tunnel encapsulation attribute [RFC9012] attached to the route.

d. Does not modify the procedures for the EVPN MAC/IP Advertisement routes.

e. Suppresses the advertisement of the A-D per EVI routes for the Ethernet Segment configured in Anycast Multi-homing mode.

f. In case of a failure on the Ethernet Segment link, the Egress NVE withdraws the A-D per ES route(s), as well as the ES route for the Ethernet Segment. The Egress NVE cannot withdraw the Anycast VTEP address from the underlay routing protocol as long as there is at least one Ethernet Segment left that makes use of the Anycast VTEP. The Anycast VTEP address is withdrawn from the Egress NVE only if the entire Egress NVE fails or all Ethernet Segments associated with the Anycast VTEP go down.

g. Unicast traffic for a failed local Ethernet Segment may still be attracted by the Egress NVE, given that the Anycast VTEP address is still advertised in the underlay routing protocol. In this case, the Egress NVE SHOULD support the procedures in Section 4 so that unicast traffic can be rerouted to another Egress NVE attached to the Ethernet Segment.

5. The Ingress NVE that supports this document:

a. Follows the regular [RFC8365] Aliasing procedures for the Ethernet Segments of the received in A-D per ES routes without the Anycast Multi-homing mode Flag set.

b. Identifies the imported EVPN A-D per ES routes with the Anycast Multi-homing flag set and process them for Anycast Multi-homing.

c. Upon receiving and importing (on a Broadcast Domain) an EVPN MAC/IP Advertisement route for MAC-1 with a non-zero Ethernet Segment Identifier ESI-1, the NVE searches for an A-D per ES route with the same ESI-1 imported into the same Broadcast

Domain. If at least one A-D per ES route for ESI-1 is present, the NVE checks whether the Anycast Multi-Homing flag is set.

* If the flag is not set, the ingress NVE follows the procedures defined in [RFC8365].

* If the Anycast Multi-Homing flag is set, the ingress NVE programs MAC-1 to be associated with destination ESI-1.

The ESI-1 destination is resolved to the Ethernet Segment Anycast VTEP, which is derived from the A-D per ES routes, along with the VNI (e.g., VNI-1) received in the MAC/IP Advertisement route.

d. When the ingress NVE receives a frame with destination MAC address MAC-1 on any of the Attachment Circuits of the Broadcast Domain, the MAC lookup resolves to ESI-1 as the destination. The frame is then encapsulated into a VXLAN (or other IP tunnel) packet with the destination VTEP set to the Anycast VTEP and the VNI set to VNI-1. Because all Egress NVEs attached to the Ethernet Segment have previously advertised reachability to the Anycast VTEP, the ingress NVE creates an underlay ECMP set for the Anycast VTEP (assuming multiple equal-cost underlay paths). As a result, per-flow load balancing is achieved.

e. The Ingress NVE MUST NOT use an Anycast VTEP as the outer source IP address of the VXLAN (or NVO3) tunnel, unless the ingress NVE also functions as an egress NVE that re-encapsulates the traffic into a tunnel for the purpose of Fast Reroute (see Section 4).

f. The reception of one or more MP_UNREACH_NLRI messages for the A-D per ES routes for Ethernet Segment Identifier ESI-1 does not change the programming of the MAC addresses associated to ESI-1 as long as there is at least one valid A-D per ES route for ESI-1 in the Bridge Domain. The reception of the MP_UNREACH_NLRI message for the last A-D per ES route for ESI-1 triggers the mass withdraw procedures for all MAC entries pointing at ESI-1.

g. As an OPTIONAL optimization, if an ingress node receives an MP_UNREACH_NLRI message for the A-D per ES route from one of the NVEs on the Ethernet Segment - and only one NVE remains active on that Ethernet Segment - the ingress node may update the Ethernet Segment destination resolution from the Anycast VTEP to the Unicast VTEP, derived from the next hop of the A-D per ES routes of the Ethernet Segment remaining NVE.

6. The procedures on the Ingress NVE for Anycast Multi-homing assume that all the Egress NVEs attached to the same Ethernet Segment advertise the Anycast Multi-homing flag value set to 1 and the same Anycast VTEP address in their A-D per ES routes for the Ethernet Segment. The following error-handling procedures are applied on the ingress NVE when the Ethernet Segment egress NVEs do not consistently advertise the same Anycast Multi-Homing flag and Anycast VTEP address values:

   * An A-D per ES route received with the Anycast Multi-Homing flag set to 1, but without an Anycast VTEP address or with an Anycast VTEP address that cannot be resolved, MUST NOT be considered for the Anycast procedures described in this section. If the ingress NVE receives other A-D per ES routes with the Anycast Multi-Homing flag set and advertising the same (resolvable) Anycast VTEP, it applies the procedures described in point 5 above, excluding the invalid route.

   * An A-D per ES route received with the Anycast Multi-Homing flag set to 0 MUST NOT be treated as belonging to an Anycast Multi-Homing Ethernet Segment, regardless of whether the route includes an Anycast VTEP value. In this case, the ingress NVEs SHOULD program the Ethernet Segment destination using the next hops derived from the A-D per ES routes, i.e., the corresponding unicast VTEP addresses.

   * If all A-D per ES routes received for a given Ethernet Segment have the Anycast Multi-Homing flag set and include an Anycast VTEP value, but the advertised Anycast VTEP values differ, the ingress NVEs SHOULD program the Ethernet Segment destination using the next hops derived from the A-D per ES routes, i.e., the corresponding unicast VTEP addresses.

Non-upgraded NVEs ignore the Anycast Multi-homing flag value and the BGP tunnel encapsulation attribute.

3.1. Anycast Multi-Homing Example

Consider the example in Figure 1 where three Leaf routers run EVPN over VXLAN tunnels. Suppose Leaf routers L1, L2 and L3 support Anycast Multi-homing as described in Section 3, and Ethernet Segments ES-1 and ES-2 are configured as Anycast Ethernet Segments in all-active mode, using the Anycast VTEP IP12.

Leaf routers L1 and L2 each advertise an A-D per ES route for ESI-1 and an A-D per ES route for ESI-2 (in addition to the ES routes). Both routes include the Anycast Aliasing flag set and the same Anycast VTEP IP12. Upon receiving MAC/IP Advertisement routes for the two Ethernet Segments, Leaf L3 programs the MAC addresses associated with their respective destination Ethernet Segment. Therefore, when sending unicast packets to Tenant Systems TS1 or TS2, L3 uses the Anycast VTEP address as the outer IP destination. All A-D per EVI routes for ES-1 and ES-2 are suppressed.

Suppose only Leaf L1 learns TS1 MAC address, hence only L1 advertises a MAC/IP Advertisement route for TS1 MAC with ESI-1. In that case:

* Leaf L3 has the Anycast VTEP IP12 programmed in its routing table, associated with an underlay ECMP set composed of Spine-1 and Spine-2. The TS1 MAC address is programmed with the destination ESI-1, resolved to Anycast VTEP IP12.

* When Tenant System TS3 sends unicast traffic to TS1, Leaf L3 encapsulates the frames into VXLAN packets with the destination VTEP set to Anycast VTEP IP12. Leaf L3 can perform per-flow load balancing using only the underlay ECMP set, without needing to create an overlay ECMP set.

* Spine-1 and Spine-2 also create underlay ECMP-sets for Anycast VTEP IP12 with next hops L1 and L2. Therefore, in case of:

  - A failure of the link between L1 and Spine-1, Spine-1 immediately removes L1 from the ECMP set for IP12, and packets are rerouted faster than when regular aliasing is used.

  - A failure of the link between TS1 and L1, Leaf L1 sends an MP_UNREACH_NLRI for the A-D per ES route for ESI-1. Upon receiving this message, Leaf L3 does not change the resolution of the ESI-1 destination, because the A-D per ES route for ESI-1 from L2 remains active. Packets sent to TS1 that arrive at Leaf L1 are “fast-rerouted” to Leaf L2 as described in Section 4.

* As per Section 3, point 5g, Leaf L3 can optionally be configured to change the resolution of the ESI-1 destination to the unicast VTEP (derived from the A-D per ES routes) upon receiving an MP_UNREACH_NLRI for the A-D per ES route from L1. Even so, in-flight packets destined for TS1 and arriving at Leaf L1 are still “fast-rerouted” to Leaf L2.

4. EVPN Fast Reroute Extensions For Anycast Multi-Homing

The procedures in Section 3 may result in situations where known unicast traffic destined for an Anycast VTEP of an Ethernet Segment arrives at an Egress NVE whose Ethernet Segment link is in a failed state. In that case, the Egress NVE SHOULD re-encapsulate the traffic into a NVO3 tunnel following the procedures described in [I-D.ietf-bess-evpn-fast-reroute], section 7.1, with the following modifications:

1. The Egress NVEs in this document do not advertise A-D per EVI routes, therefore there is no signaling of specific redirect labels or VNIs. The Egress NVE uses the VNI assigned to the Egress NVEs attached to the Ethernet Segment in Anycast mode, or Domain-wide Common Block label of the Ethernet Segment NVEs when re-encapsulating the traffic into an NVO3 tunnel (Section 3, point 3).

2. Additionally, when rerouting traffic, the Egress NVE uses the Anycast VTEP of the Ethernet Segment as the outer source IP address of the NVO3 tunnel. Note this is the only scenario in this document where using the Anycast VTEP as the source IP address is permitted. Receiving NVO3-encapsulated packets with a local Anycast VTEP indicates that those packets have been "fast-rerouted". Therefore, they MUST NOT be forwarded into another tunnel.

5. Applicability of Anycast Multi-Homing to Inter-Subnet Forwarding

Anycast Multi-Homing can also be applied to inter-subnet forwarding scenarios. The diagram in Figure 3 illustrates such a scenario where Anycast Multi-Homing is used. This diagram serves as a reference throughout this section.

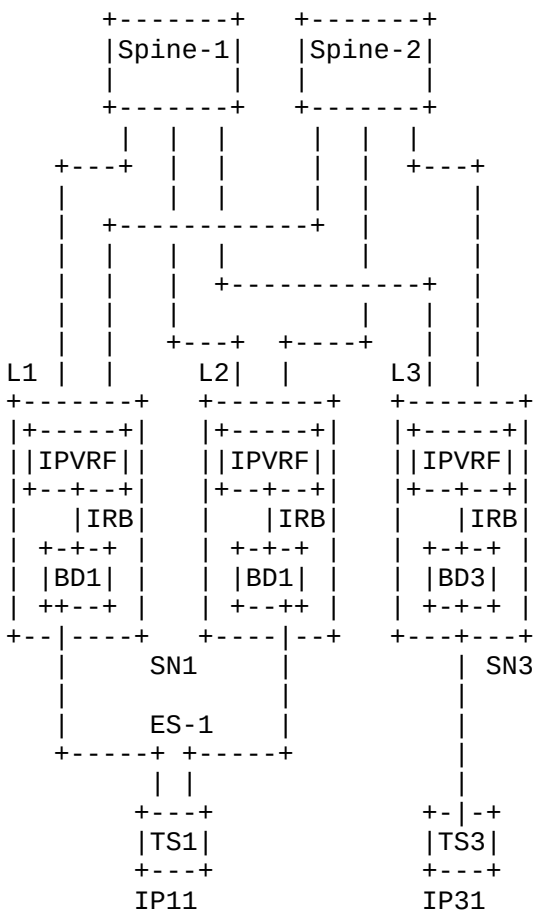


Figure 3: Anycast Multi-Homing for Inter-Subnet Forwarding

5.1. Anycast Multi-Homing and Multi-Homed IP Prefixes

Multi-homed IP Prefixes (subnets or hosts attached to two or more Leaf routers) can also benefit from Anycast Multi-Homing. These multi-homed IP prefixes are advertised via EVPN IP Prefix routes, as described in [RFC9136], section 4.4.

Not all the challenges described in Section 1.2 apply to the Anycast Multi-homing scenario for IP prefixes. For example, multi-homed IP prefixes advertised via EVPN IP Prefix routes do not require the use of Ethernet Segments, so the challenge described in Section 1.2 (a) does not arise here. However, using Anycast VTEPs for EVPN IP Prefix routes advertised from the Leaf routers attached to the same IP prefix can still improve the convergence, reduce processing overhead, and address inefficient underlay forwarding, as explained in Section 1.2 (b) and (c).

The solution consists of the following modifications of the IP-VRF-to-IP-VRF model ([RFC9136] section 4.4):

1. Similar to Section 3 bullet 3, any two or more Egress NVEs attached to the same IP prefix working in Anycast Multi-homing mode MUST use the same VNI or label to identify the IP-VRF (or SBD) associated with that IP prefix. The same considerations described in that bullet also apply to EVPN IP Prefix routes for multi-homed IP prefixes.

2. The use of Anycast VTEPs by Egress NVEs multi-homed to the same IP prefix is as follows:

    a. A single Anycast VTEP is shared by the group of NVEs associated with the same IP prefix(es) and these NVEs advertise reachability for the Anycast VTEP in the underlay routing protocol, as explained in Section 3, bullet 4b.

    b. In an Interface-less IP-VRF-to-IP-VRF model ([RFC9136] section 4.4.1), the Anycast VTEP is encoded in a BGP tunnel encapsulation attribute [RFC9012] and advertised together with the EVPN IP Prefix route for the IP prefix. The presence of the Anycast VTEP in the IP Prefix route indicates to the ingress NVE that the prefix is using Anycast Multi-Homing. Additionally, the same MAC address SHOULD be encoded in the EVPN Router' s MAC extended community of the IP Prefix routes advertised for the same multi-homed IP prefix by all the Egress NVEs using Anycast Multi-Homing.

    c. In an Interface-ful IP-VRF-to-IP-VRF with SBD IRB model ([RFC9136] section 4.4.2), the Anycast VTEP is advertised along with the IP Prefix route and also with the EVPN MAC/IP Advertisement route for the SBD IRB interface. In this model, when Anycast Multi-Homing is used, all Egress NVEs attached to the same IP prefix MUST use the same Anycast SBD IRB IP and MAC addresses. Specifically, the SBD IRB is configured with an Anycast MAC and IP address that are shared by all Egress NVEs operating in Anycast Multi-Homing mode.

The IP Prefix routes for multi-homed IP prefixes are advertised using these Anycast Gateway IP addresses in the Gateway IP field.

d. In an Interface-ful IP-VRF-to-IP-VRF with Unnumbered SBD IRB model ([RFC9136] section 4.4.3), Anycast Multi-Homing operates similarly to the interface-ful SBD IRB model (bullet c). The difference is that the SBD IRBs do not have an IP address, and the Anycast SBD MAC address is used as the overlay index for IP Prefix resolution.

3. The Ingress NVEs supporting this document operate as follows:

a. Upon receiving and importing an IP Prefix route with an Anycast VTEP, the Ingress NVE checks whether the same prefix has been received from another egress NVE and whether that IP Prefix route contains a matching Anycast VTEP and a matching Router’s MAC. If both conditions are met, the prefix is programmed to use Anycast forwarding - by using the Anycast VTEP as the outer destination IP address and the shared Router’s MAC as inner destination MAC address. If multiple IP Prefix routes for the same prefix exist but their Anycast VTEPs or Router’s MAC do not match, the IP Prefix routes are processed as described in [RFC9136] and MUST NOT be programmed to use Anycast forwarding. In the interface-less model, this means the prefix is programmed using the next hop from the IP Prefix route. In the interface-ful models, the prefix is resolved to the EVPN MAC/IP Advertisement routes associated with the non-Anycast SBD IRB.

b. When using Anycast forwarding, regardless of the IP-VRF-to-IP-VRF implemented model, the Ingress NVE encapsulates the packets destined for a multi-homed IP prefix into a VXLAN (or other NVO3) packet with the destination VTEP set to the Anycast VTEP and the VNI set to the VNI of the IP-VRF (Interface-less model) or the SBD (Interface-ful models). Because all Egress NVEs attached to the multi-homed IP prefix have previously advertised reachability to the Anycast VTEP, the ingress NVE creates an underlay ECMP set for the Anycast VTEP (assuming multiple equal-cost underlay paths).

c. The Ingress NVE MUST NOT use an Anycast VTEP as the outer source IP address of the tunnel, unless the ingress NVE also functions as an egress NVE that re-encapsulates the traffic into a tunnel for the purpose of Fast Reroute (Section 4).

In the example shown in Figure 3, IP prefix SN1 is multi-homed to Leaf routers L1 and L2. Assuming the interface-less model ([RFC9136], Section 4.4.1) and following the procedure described above, L1 and L2 advertise SN1 in IP Prefix routes that use the same Anycast VTEP IP12 and the same Router's MAC M12. The ingress NVE, L3, programs SN1 with VTEP IP12 as the outer destination IP and M12 as the inner destination MAC. Packets destined for SN1 are then load-balanced based on the underlay ECMP set associated with Anycast VTEP IP12.

Although Figure 3 assumes that SN1 is associated with IRB interfaces, Anycast Multi-Homing can also be used when SN1 is associated with non-IRB Layer 3 interfaces. It is important to note that if SN1 is associated with IRB interfaces, a link failure between TS1 and L1 does not trigger the advertisement of an MP_UNREACH_NLRI message for SN1. As a result, L1 continues to attract traffic destined for SN1, which must then be fast-rerouted to L2.

5.2. Anycast Multi-Homing and EVPN IP Aliasing

IP Aliasing is described in [I-D.ietf-bess-evpn-ip-aliasing] and leverages Ethernet Segments to provide fast convergence multi-homing for host routes ([I-D.ietf-bess-evpn-ip-aliasing] sections 1.1 and 1.2) or IP Prefix routes ([I-D.ietf-bess-evpn-ip-aliasing] section 1.3) programmed in an IP-VRF. Anycast Multi-homing can also be applied to these Ethernet Segments to address all the challenges described in Section 1.2, but specifically in the context of IP-VRFs rather than MAC-VRFs. The procedures described in Section 3 and Section 4 of this document apply to IP Aliasing, with the following considerations:

1. The Egress NVEs attached to the Anycast Multi-homing Ethernet Segments:

    a. Advertise both sets of Ethernet A-D per ES and IP A-D per ES routes with the Anycast Multi-homing mode flag and the Anycast VTEP.

    b. Suppress the advertisement of both Ethernet A-D per EVI and IP A-D per EVI routes for Ethernet Segment configured in Anycast Multi-homing mode.

c. Include the EVPN Router's MAC Extended Community along with the IP A-D per ES routes if the encapsulation used between the PEs for inter-subnet forwarding is an Ethernet NVO tunnel [RFC9136]. When advertised with the IP A-D per ES routes, the EVPN Router's MAC extended community SHOULD contain the same MAC address value in all the IP A-D per ES routes advertised by all the Egress NVEs attached to the same Anycast Multi-homing Ethernet Segment.

d. Apply the same procedures to IP A-D per ES routes as those described for Ethernet A-D per ES routes in Section 3 for Egress NVEs.

2. The Ingress NVEs:

a. Upon receiving and importing an EVPN MAC/IP Advertisement route ([RFC9135]) or an IP Prefix route ([RFC9136]) with a non-zero Ethernet Segment Identifier (ESI), the NVE searches for an IP A-D per ES route with the same ESI imported into the same IP-VRF. If at least one IP A-D per ES route for the ESI is present, the NVE checks whether the Anycast Multi-Homing flag is set.

* If the flag is not set, the ingress NVE follows the procedures described in [I-D.ietf-bess-evpn-ip-aliasing].

* If the flag is set, the ingress NVE programs the host route entry (from the MAC/IP Advertisement route with two VNIs or from the EVPN IP Prefix route) or the IP Prefix entry (from the EVPN IP Prefix route) to be associated with the ES destination that uses an Anycast VTEP

b. Other than the above, apply the same procedures to IP A-D per ES routes as those described for Ethernet A-D per ES routes in Section 3 for Ingress NVEs.

6. Applicability of Anycast Multi-Homing to SRv6 tunnels

To be added.

7. Applicability of Anycast Multi-Homing to Inter-Domain Service Gateways

The procedures described in this document apply not only to egress NVEs attached to the same CE Ethernet Segment or multi-homed prefix, but also to Service Gateways that interconnect domains, as per [RFC9014] or [I-D.ietf-bess-evpn-ipvpn-interworking].

7.1. Anycast Multi-Homing Inter-Domain Service Gateways for Broadcast Domains

This document extends the Anycast Multi-homing Ethernet Segment concept to also the I-ES (Interconnect Ethernet Segments) defined in [RFC9014], when used to interconnect EVPN domains. When Anycast Multi-homing is applied to I-ESes, such I-ESes are referred to as Anycast Multi-Homing I-ESes.

Section 4.4 in [RFC9014] applies to Anycast Multi-homing with the following considerations:

1. As specified in [RFC9014], an I-ES is configured in all the Service Gateways of the redundancy group. The I-ES - in this case an Anycast Multi-homing I-ES - represents the WAN PEs to the DC but also the DC EVPN NVEs to the WAN.

2. Anycast Multi-Homing Service Gateways follow the procedures specified in Section 3. They behave as ingress NVEs with respect to Anycast Multi-Homing Ethernet Segments located on NVEs in either domain, and as egress NVEs for the Anycast Multi-Homing I-ES to which they are attached.

3. The EVPN ES routes, Inclusive Multicast Ethernet Tag routes and MAC/IP Advertisement routes are advertised and processed as specified in [RFC9014]. Ethernet A-D per ES routes for the Anycast Multi-homing I-ES are advertised with the Anycast Multi-homing flag set and the Anycast VTEP, as described in Section 3. Ethernet A-D per EVI routes are suppressed when working on this mode.

4. Anycast Multi-Homing Service Gateways may provide interconnection between two domains, where only one of the domains supports Anycast Multi-homing. This situation may arise, for example, when one domain uses an encapsulation supported in Section 3 (e.g., VXLAN), while the other domain uses MPLS tunnels. Therefore, there are two supported scenarios:

    * When both domains support the Anycast Multi-homing procedures, the Service Gateways suppress the advertisement of A-D per EVI routes toward either domain. In addition, they advertise A-D per ES routes to both domains with the Anycast Multi-Homing flag set to 1 and including the Anycast VTEP value.

    * When only one of the two domains supports Anycast Multi-homing, the Service Gateways suppress the advertisement of A-D per EVI routes toward the domain that supports Anycast Multi-homing, while continuing to advertise them to the other

domain. Similarly, the Anycast Multi-Homing flag set to 1 and the Anycast VTEP value are included only in the A-D per ES routes advertised to the domain that supports the Anycast Multi-Homing procedures.

5. As in the case of the egress NVEs described in Section 3, Anycast Multi-homing Service Gateways do not modify their DF Election, Split Horizon, or BUM forwarding procedures. Service Gateways may also have local attachment circuits. The procedures for DF Election, Split Horizon, BUM forwarding, and handling of local nonI-ES attachment circuits are specified in [RFC9014].

6. Extensions for [RFC9014] Service Gateways MAY also be enabled if Anycast Multi-homing I-ESes are used. Examples include the use of the Unknown MAC Route, as specified in [I-D.ietf-bess-evpn-umr-mobility], and the use of D-PATH, as specified in [I-D.ietf-bess-evpn-dpath].

7.2. Anycast Multi-Homing Inter-Domain Service Gateways for Inter-Subnet-Forwarding

This document also introduces Anycast Multi-homing for Service Gateways that interconnect Layer-3 EVPN domains, as specified in [I-D.ietf-bess-evpn-ipvpn-interworking], section 8. When redundant Gateways are deployed between two EVPN domains, Anycast Multi-Homing for IP Prefixes, as described in Section 5 MAY be used. The procedures specified in section 8 of [I-D.ietf-bess-evpn-ipvpn-interworking] are applied, subject to the following considerations:

1. Similarly to the case described in Section 7.1, Service Gateways may interconnect EVPN domains that both support the Anycast Multi-Homing procedures, or EVPN domains where only one domain supports these procedures. When the export conditions of [I-D.ietf-bess-evpn-ipvpn-interworking] section 8.1 are satisfied, the Gateway PE advertises a given prefix P in an IP Prefix route that follows [I-D.ietf-bess-evpn-ipvpn-interworking] section 8.2, but in addition, if the destination domain supports Anycast Multi-homing, the Gateway PE includes the Anycast VTEP in the IP Prefix route, as defined in Section 5.

2. The Gateway PE MUST NOT propagate the Anycast VTEP encoded in the Tunnel Egress Endpoint Sub-TLV of the BGP Encapsulation Attribute, irrespective of whether Uniform Propagation Mode ([I-D.ietf-bess-evpn-ipvpn-interworking], section 5.2) is enabled. The Anycast VTEP is always originated by the Gateway PE, and matches a local Gateway IP addresses, as described in this document.

3. This document specifies the use of Anycast Multi-Homing VTEPs only for the EVPN address family. Therefore, a Gateway PE MUST NOT advertise an Anycast VTEP to a BGP peer using any AFI/SAFI other than EVPN.

8. Operational Considerations

"Underlay convergence" — that is, convergence handled by the underlay routing protocol in the event of a failure — is generally considered faster than "overlay convergence," where EVPN processes network convergence when failures occur.

The use of Anycast Multi-Homing is especially valuable in scenarios where the operator aims to optimize convergence. In this model, a node failure affecting an Ethernet Segment Egress NVE simply results in the underlay routing protocol rerouting traffic to another Egress NVE that advertises the same Anycast VTEP. This underlay rerouting to a different owner of the Anycast VTEP is extremely fast and efficient, particularly in data center designs that use BGP in the underlay and follow the Autonomous System allocation recommended in [RFC7938] for loop protection.

To illustrate this, consider a link failure between L1 and Spine-1, as shown in Figure 1. If Spine-1 and Spine-2 are assigned the same Autonomous System Number for their underlay BGP peering sessions and no "Allowas-in" is configured (per [RFC7938]), packets destined for the Anycast VTEP IP12 received by Spine-1 are immediately rerouted to L2 when the L1-Spine-1 link fails. In contrast, if unicast VTEPs are used (as in regular all-active Ethernet Segments), in-flight packets destined for the unicast VTEP on L1 that arrive at Spine-1 would be dropped if the L1-Spine-1 link is unavailable. This example highlights how Anycast Multi-Homing achieves significantly faster convergence.

Another benefit of Anycast Multi-homing is the reduction of EVPN control plane pressure (due to the suppression of the A-D per EVI routes).

However, operators should consider the following operational factors before deploying this solution:

1. Troubleshooting Anycast Multi-Homing Ethernet Segments differs from troubleshooting regular all-active Ethernet Segments. In traditional setups, operators rely on the withdrawal of an A-D per EVI route as an indication that the Ethernet Segment has failed in the specific Broadcast Domain associated with that route. With Anycast Multi-Homing, however, the suppression of A-D per EVI routes means that logical failures affecting only a subset of Broadcast Domains on the Ethernet Segment — while others remain operational — are more difficult to detect.

2. Anycast Multi-homing Ethernet Segments MUST NOT be used in in the following cases:

    a. If the Ethernet Segment multi-homing redundancy mode is not All-Active mode.

    b. If the Ethernet Segment is used on EVPN VPWS Attachment Circuits [RFC8214].

    c. If the Attachment Circuit Influenced Designated Forwarded capability is needed in the Ethernet Segment [RFC8584].

    d. If advanced multi-homing features that make use of the signaling in EVPN A-D per EVI routes are needed. An example would be per EVI mass withdraw [RFC8365].

    e. If unequal load balancing is needed [I-D.ietf-bess-evpn-unequal-lb].

    f. If the tunnels used by EVPN in the Broadcast Domains associated with the Ethernet Segment are not IP tunnels (i.e., they are not NVO3 tunnels).

    g. If the NVEs attached to the Ethernet Segment do not use the same VNI or label to identify the same Broadcast Domain.

3. Using the procedure in Section 3 may result in packets being permanently fast-rerouted in the event of a link failure. To illustrate this, consider three Egress NVEs — L1, L2, and L3 — attached to ES-1. In this scenario, a failure of ES-1 on L1 does not prevent the network from continuing to send packets to L1 with the Anycast VTEP as the destination. When L1 receives these packets, it re-encapsulates them and forwards them to, for instance, L2. This rerouting persists for as long as ES-1 remains in a failed state on L1. In such cases, operators may consider deploying direct inter-node links between the Egress NVEs to optimize fast reroute forwarding. In the example above, rerouted packets are handled more efficiently if L1, L2, and L3 are directly connected.

9. Security Considerations

To be added.

10. IANA Considerations

IANA is requested to allocate the flag "A" or "Anycast Multi-homing mode" in bit 2 of the EVPN ESI Multihoming Attributes registry for the 1-octect Flags field in the ESI Label Extended Community.

11. Contributors

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

Nick Morris, Verizon

nicklous.morris@verizonwireless.com

12. Acknowledgments

13. Annex - Potential Multi Ethernet Segment Anycast Multi-Homing optimizations

This section is here for documentation purposes only, and it will be removed from the document before publication. While these procedures were initially included in the document, they introduce additional complexity and are therefore excluded, as they undermine the primary goal of using anycast VTEPs, which is to simplify EVPN operations. However, the section is included as an annex for completeness.

As described in Section 7, the use of Anycast Multi-Homing may mean that packets are permanently fast rerouted in case of a link failure. Some potential additional extensions on the Ingress NVE may mitigate the permanent "fast rerouting", as follows:

1. On the Ingress NVEs, an "anycast-aliasing-threshold" and a "collect-timer" can be configured. The "anycast-aliasing-threshold" represents the number of active Egress NVEs per Ethernet Segment under which the ingress PE no longer uses the Anycast VTEP address to resolve the Ethernet Segment destination (and uses the Unicast VTEP instead, derived from the MAC/IP Advertisement route next hop). The "collect-timer" is triggered upon the creation of the Ethernet Segment destination, and it is needed to settle on the number of Egress NVEs for the Ethernet Segment against which the "anycast-aliasing-threshold" is compared.

2. Upon expiration of the "collect-timer", the Ingress NVE computes the number of Egress NVEs for the Ethernet Segment based on the next hop count of the received A-D per ES routes. If the number of Egress NVEs for the Ethernet Segment is greater than or equal to the "anycast-aliasing-threshold" integer, the Ethernet Destination is resolved to the Anycast VTEP address. If lower than the threshold, the Ethernet Destination is resolved to the unicast VTEP address.

In most of the use cases in multi-tenant Data Centers, there are two Leaf routers per rack that share all the Ethernet Segments of Tenant Systems in the rack. In this case, the "anycast-aliasing-threshold" is set to 2 and in case of link failure on the Ethernet Segment, this limits the amount of "fast-rerouted" traffic to only the in-flight packets.

As an example, consider Figure 1. Suppose Leaf router L3 supports these additional extensions. Leaf routers L1 and L2 both advertise an A-D per ES route for ESI-1, and an A-D per ES route for ESI-2. Both routes will carry the Anycast Multi-homing flag set and the same Anycast VTEP IP12. Following the described procedure, Leaf L3 is configured with anycast-aliasing-threshold = 2 and collect-timer = t. Upon receiving MAC/IP Advertisement routes for the two Ethernet Segments and the expiration of "t" seconds, Leaf L3 determines that the number of NVEs for ESI-1 and ESI-2 is equal to the threshold. Therefore, when sending unicast packets to Tenant Systems TS1 or TS2, L3 uses the Anycast VTEP address as outer IP address. Suppose now that the link TS1-L1 fails. Leaf L1 then sends an MP_UNREACH_NLRI for the A-D per ES route for ESI-1. Upon reception of the message, Leaf L3 changes the resolution of the ESI-1 destination from the Anycast VTEP to the Unicast VTEP derived from the MAC/IP

Advertisement route next hop. Packets sent to Tenant System TS2 (on ES-2) still use the Anycast VTEP. In-flight packets sent to TS1 but still arriving at Leaf L1 are "fast-rerouted" to Leaf L2 as per Section 4.

Another potential optimization is to use different Anycast VTEPs per ES. The proposal in Section 3 uses a shared VTEP for all the Ethernet Segments in a common Egress NVE group. In case the number of Egress NVEs sharing the group of Ethernet Segments is limited to two, an alternative proposal is to use a different Anycast VTEP per Ethernet Segment, however allocate all those Anycast VTEP addresses from the same subnet. A single IP Prefix for such subnet is announced in the underlay routing protocol by the Egress NVEs. The benefit of this proposal is that, in case of link failure in one individual Ethernet Segment, e.g., link TS1-L1 in Figure 1, Leaf L2 detects the failure (based on the withdraw of the A-D per ES and ES routes) and can immediately announce the specific Anycast VTEP address (/32 or /128) into the underlay. Based on a Longest Prefix Match when routing NVO3 packets, Spines can immediately reroute packets (with destination the Anycast VTEP for ESI-1) to Leaf L2. This may reduce the amount of fast-rerouted VXLAN packets and spares the Ingress NVE from having to change the resolution of the Ethernet Segment destination from the Anycast VTEP to the Unicast VTEP.

14. References

14.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[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>.

[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>.

[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>.

[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-13, 24 June 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-rfc7432bis-13>.

[RFC9573] Zhang, Z., Rosen, E., Lin, W., Li, Z., and IJ. Wijnands, "MVPN/EVPN Tunnel Aggregation with Common Labels", RFC 9573, DOI 10.17487/RFC9573, May 2024, <https://www.rfc-editor.org/info/rfc9573>.

[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>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

[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>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[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>.

[I-D.ietf-bess-evpn-ipvpn-interworking]
Rabadan, J., Sajassi, A., Rosen, E. C., Drake, J., Lin, W., Uttaro, J., and A. Simpson, "Interconnecting EVPN and IPVPN Domains", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-ipvpn-interworking-16, 28 February 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-ipvpn-interworking-16>.

14.2. Informative References

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[RFC8926] Gross, J., Ed., Ganga, I., Ed., and T. Sridhar, Ed., "Geneve: Generic Network Virtualization Encapsulation", RFC 8926, DOI 10.17487/RFC8926, November 2020, <https://www.rfc-editor.org/info/rfc8926>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", RFC 7510, DOI 10.17487/RFC7510, April 2015, <https://www.rfc-editor.org/info/rfc7510>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[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>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC9469] Rabadan, J., Ed., Bocci, M., Boutros, S., and A. Sajassi, "Applicability of Ethernet Virtual Private Network (EVPN) to Network Virtualization over Layer 3 (NVO3) Networks", RFC 9469, DOI 10.17487/RFC9469, September 2023, <https://www.rfc-editor.org/info/rfc9469>.

[I-D.ietf-bess-evpn-ip-aliasing] Sajassi, A., Rabadan, J., Pasupula, S., Krattiger, L., and J. Drake, "EVPN Support for L3 Fast Convergence and Aliasing/Backup Path", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-ip-aliasing-03, 7 May 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-ip-aliasing-03>.

[I-D.ietf-bess-evpn-unequal-lb] Malhotra, N., Sajassi, A., Rabadan, J., Drake, J., Lingala, A. R., and S. Thoria, "Weighted Multi-Path Procedures for EVPN Multi-Homing", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-unequal-lb-32, 27 February 2026, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-unequal-lb-32>.

[I-D.ietf-bess-evpn-fast-reroute] Burdet, L. A., Brissette, P., Miyasaka, T., Rabadan, J., Liu, Y., and C. Lin, "EVPN Fast Reroute", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-fast-reroute-00, 9 June 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-fast-reroute-00>.

[RFC9746] Rabadan, J., Nagaraj, K., Lin, W., and A. Sajassi, "BGP EVPN Multihoming Extensions for Split-Horizon Filtering", RFC 9746, DOI 10.17487/RFC9746, March 2025, <https://www.rfc-editor.org/info/rfc9746>.

[I-D.ietf-bess-evpn-dpath] Rabadan, J., Sathappan, S., Gautam, M., Brissette, P., and W. Lin, "Domain Path (D-PATH) for Ethernet VPN (EVPN) Interconnect Networks", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-dpath-03, 16 October 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-dpath-03>.

[I-D.ietf-bess-evpn-umr-mobility]
Sajassi, A., Rabadan, J., Nichol, A., Krattiger, L., and K. Ananthamurthy, "Applications and Procedures for Unknown MAC Route in EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-umr-mobility-00, 25 November 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-umr-mobility-00>.

[CLOS1953] Clos, C., "A Study of Non-Blocking Switching Networks", March 1953.

Authors' Addresses

Jorge Rabadan (editor)
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: jorge.rabadan@nokia.com

Kiran Nagaraj
Nokia
520 Almanor Avenue
Sunnyvale, CA 94085
United States of America
Email: kiran.nagaraj@nokia.com

Alex Nichol
Arista
Email: anichol@arista.com

Ali Sajassi
Cisco Systems
Email: sajassi@cisco.com

Wen Lin
Juniper Networks
Email: wlin@juniper.net

Jeff Tantsura
Nvidia
Email: jefftant.ietf@gmail.com