Workgroup: Network Working Group
Internet-Draft:
draft-fu-bess-evpn-umr-application-01
Published: 17 November 2023
Intended Status: Standards Track
Expires: 20 May 2024
Authors: Z. Fu
Huawei Technologies
T. Zhu
Huawei Technologies
H. Wang
Huawei Technologies
J. Dai
Huawei Technologies
D. Wang
Huawei Technologies

UMR application in Ethernet VPN(EVPN)

Abstract

This document describes an application scenario that how unknown MAC-route(UMR) is used in the EVPN network. In particular, this document describes how MAC address route and UMR route are advertised on DC's GW or NVE. This document also describes the soloution that MAC mobility issue due to the lack of advertisement of specific MAC routes. However, some incremental work is required, which will be covered in a separate document.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

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 20 May 2024.

Table of Contents

1. Introduction

In DCI scenario, if multiple DCs are interconnected into a single EVI, each DC will have to import all of the MAC addresses from each of the other DCs. [RFC9014]. In addition, in user authentication scenario, a large number of users send authentication packets to the aggregation device through the access device, as a result, there are large scale of MAC addresses on RRs and aggregation devices. This document describes the use of the Unknown MAC-route(UMR). The solution advertises an unknown MAC-route (UMR) route[RFC9014] instead of advertising all specific MAC routes and reducing the MAC scale.

However, since the solution only sends UMR routes instead of advertising specific MAC routes, the MAC mobility function of EVPN cannot take effect normally. In particular, this document describes a MAC mobility procedure in UMR scenario.

Also, This document discusses how the functional requirements for E-Tree service[RFC8317] can be met with a solution based on UMR application in EVPN. The details of this function are described in Section 5.

2. Terminology

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 BCP14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

"GW": Gateway or Data Center Gateway

"DC": Data Center

"NVE": Network Virtualization Edge

"UMR": Unknown MAC Route

"E-Tree": Ethernet-Tree

"I-ES and I-ESI": Interconnect Ethernet Segment and Interconnect Ethernet Segment Identifier. An I-ES is defined on the GWs for multihoming to/from the WAN.

3. The procedure of UMR

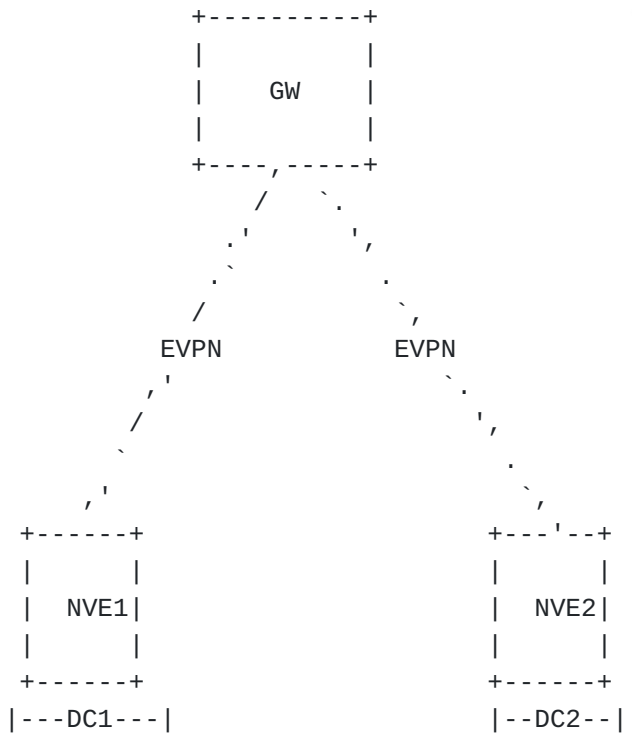

Figure 1

1. All the MAC addresses are learned on NVE1/NVE2 within DC should advertised to DC's GW device accrording EVPN MAC/IP routes in the control plane.

2. All the MAC addresses are learned on NVE within DC should advertised to the other NVE that in the same DC, so that the NVE to NVE that in the same DC communication is always direct and does not go through the GW[RFC7543].

3. The MAC addresses are learned on NVE should not advertised to the other NVE that in the different DC.

4. The DC's GW advertise UMR route to NVE1/NVE2 instead of advertising the specific MAC in order to reduce the device's routes pressure. The UMR route is defined in [RFC7543] [RFC9014]and is a regular EVPN MAC/IP advertisement route in which the MAC address length is set to 48, the MAC address is set to 0, and the ESI field is set to DC's GW I-ES.

5. NVE1/NVE2 need to understand and process the UMR route, send frame to GW. Then GW will forward the packet to correct NVE.

4. MAC Mobility for UMR

As shown above, since GW only sends UMR routes to NVE devices, NVE will not import the MAC addresses of NVEs in different DCs. When the MAC of DC1 migrates from NVE1 to DC2's NVE2, NVE1 will not perceive this migration and keep learning the MAC that has migrated to NVE2. As a result, the frame traffic to MAC from GW may go to wrong site.

4.1. MAC Mobility Issue

Step1: The user first goes online from NVE1, NVE1 learns the user's MAC1, and advertise EVPN MAC1 route to GW.

Step2: The GW receives the MAC1 route from NVE1, installs MAC1 to the local MAC-VRF table which the next hop of MAC1 is NVE1. Since it only sends UMR routes to NVE, it will not send EVPN MAC1 route to NVE2.

Step3: The user migrates to NVE2 and goes online. NVE2 learns the user's MAC1 and advertise EVPN MAC1 route to GW.

Step4: The GW receives the MAC1 route from NVE2, which has the same prefix as the MAC1 route from NVE1, as a result, the GW will form load balancing MAC-VRF table.

Step5: As a result, the frame traffic sent to MAC1 via the GW may be sent to NVE1 by mistake until MAC1 on NVE1 ages out.

4.2. MAC Mobility Solution

In order to solve this mac migration issue, the GW SHOULD advertise the MAC route to the NVE when the GW detect the MAC has been migrated. There are two scenarios as follows.

1. One of the scenario:

Step1: When the GW receives MAC routes that have the same prefix, rather than different next hop and different ESI, the following conclusion can be drawn, which the MAC has been migrated. At the same time, the GW only send UMR route.

Step2: If MAC route from NVE1 is selected as the best, the GW advertise MAC1 route to NVE2 with a MAC mobility extended community[RFC7432], that carrying the increased seq number.

Step3: The NVE2 receives the MAC1 route with MAC mobility extended community, and will select the MAC1 from the GW as the best, and withdraw the MAC1 originally sent to the GW.

Step4: The traffic from user will re-triggers NVE2 to learn the local MAC1, which resulting in migration, and the NVE2 will advertise MAC1 route with MAC mobility extended community that carrying the seq + 1.

Step5: When the GW receives the MAC1 route with MAC mobility extended community that carrying seq + 1, the GW will select the MAC1 from NVE2 as best, and send MAC1 route with seq + 1 to NVE1.

Step6: After receiving the MAC1 route with MAC mobility extended community that carrying seq + 1, the NVE1 will select the MAC1 from the GW as the best, and withdraw the MAC1 originally sent to the GW.

2. The other scenario:

Step1: When the GW receives MAC routes that have the same prefix, rather than different next hop and different ESI, the following conclusion can be drawn, which the MAC has been migrated. At the same time, the GW only send UMR route.

Step2: If MAC route from NVE2 is selected as the best, the GW advertise MAC1 route to NVE1 with a MAC mobility extended community, that carrying the increased seq number.

Step3: After receiving the MAC1 route with MAC mobility extended community that carrying seq + 1, the NVE1 will select the MAC1 from the GW as the best, and withdraw the MAC1 originally sent to the GW.

5. E-tree for UMR

In this scenario, since PE only sends UMR routes to remote PE devices instead of advertising the specific MAC, In this case, it is not possible to identify whether the UMR routes originates from Root ACs or Leaf ACs. In this way, unicast traffic isolation between Leafs cannot be achieved.

5.1. Scenario 1: Leaf or Root Site(s) per EVI

In this scenario, a given EVPN Instance (EVI) on PE device is either associated with Root(s) or Leaf(s), but not both. PE may receive traffic from either Root ACs or Leaf ACs for a given MAC-VRF/bridge table with UMR route. So the UMR route to Leaf ACs or Root ACs of a given EVPN Instance need to be colored with a Root or Leaf-Indication before advertising to remote PE. E-Tree Extended Community[RFC8317] can be used for such coloring. The leaf-indication indicates the UMR route has a leaf attribute. When the E-Tree extended community is advertised with the UMR advertisement route, the Leaf-Indication flag MUST be set to one and the Leaf label SHOULD be set to zero.

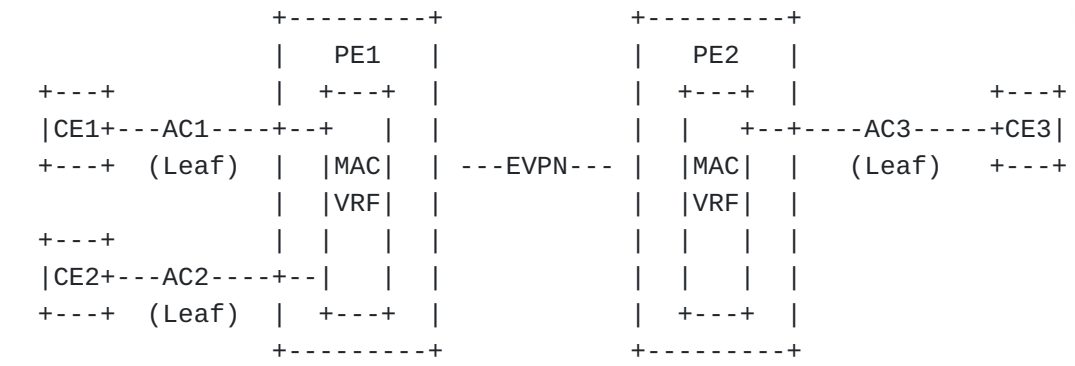

Figure 2

5.2. Scenario 2: Leaf or Root Site(s) per AC

In this scenario, a given EVPN Instance (EVI) on PE device can be associated with both Root(s) and Leaf(s). For example, in the figure below, PE for an EVPN Instance (EVI) has both Leaf and Root ACs. In this scenario, ingress filtering for known unicast traffic is not performed just like scenario-1 and thus need to receive the unicast traffic and perform egress filtering.

In order to perform egress filtering for unicast traffic received at the egress PE, the ingress PE need to color the unicast traffic in data-plane to indicate if the traffic is coming from a Root or Leaf AC. E-Tree Extended Community[RFC8317] also can be used for such scenario. When the E-Tree extended community is advertised with the UMR advertisement route for such scenario, the Leaf-Indication flag MUST be set to zero and the Leaf label MUST be valid.

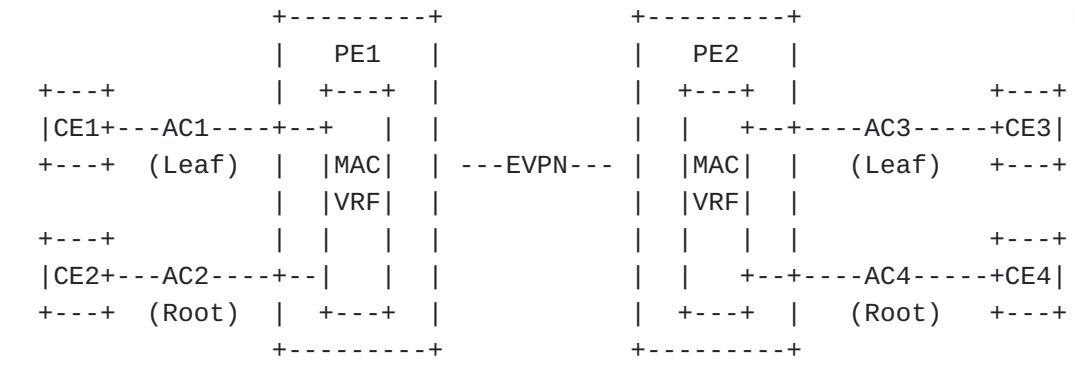

Figure 3

5.3. Known Unicast Traffic For Leaf or Root Site(s) per EVI

To provide the ingress filtering for known unicast traffic, a PE MUST indicate to other PEs what kind of sites (Root or Leaf) its UMR MAC address are associated with. This is done by advertising a Leaf-Indication flag via E-Tree extended community [RFC8317] along with its UMR MAC/IP Advertisement routes learned from a Leaf site. This E-Tree extended community MUST be advertised with UMR MAC/IP

Advertisement routes learned from a Leaf site. The lack of such a flag indicates that the UMR MAC address is associated with a Root site. This scheme applies to scenario-1 (Leaf or Root Site(s) per EVI) described in Section 5.

Tagging UMR MAC address with a Leaf-Indication enables remote PEs to perform ingress filtering for known unicast traffic. After receiving the UMR, PE2 generates a default MAC address entry comprising a full zero MAC address and the a leaf-indication. So, on the ingress PE, the MAC destination address lookup yields (in addition to the UMR forwarding adjacency) a flag than indicates whether or not the target UMR MAC is associated with a Leaf site. The ingress PE(e.g. PE2) cross-checks this flag with the status of the originating AC, and if both are Leafs, then the packet is dropped. Otherwise, if both are not leafs, then the packet is forwarded.

5.4. Known Unicast Traffic For Leaf or Root Site(s) per AC

This section specifies the procedure for egress filtering of known unicast traffic with MPLS encapsulation. To support scenario-2 efficiently, egress filtering of known unicast traffic is required as described below. In order to apply the proper egress filtering, which varies based on whether a packet is sent from a Leaf AC or a Root AC, the MPLS-encapsulated frames MUST be tagged with an indication of when they originated from a Leaf AC. This Leaf label allows for disposition PE (e.g., egress PE) to perform the necessary egress filtering function in a data plane similar to the MPLS label1 in [RFC7432].

If a PE receive UMR route with E-Tree extended community that has both Root-Indication and Leaf-Indication set along with a valid Leaf label, then the receiving PE (Root-only, Root-and-Leaf, or Leaf-only) add both MPLS label1 [RFC7432] and Leaf label to MAC-VRF/ bridge table.

The ingress PE cross-checks this flag with the status of the originating AC. If the originating AC is Leaf, then the packet is encapsulated in the EVPN Leaf label advertised by the remote PE, for that MAC address, and in the MPLS LSP label stack to reach the remote PE[RFC7432]. Accroding to receiving the packet with Leaf label, the egress PE checks the MAC-VRF/bridge table according to the destination MAC and filtering the Leaf AC before forwarding. If the originating AC is Root, then the packet is encapsulated in the EVPN MPLS label addvertised by the remote PE, for that MAC address, and in the MPLS LSP label stack to reach the remote PE. If the top MPLS label ends up being an EVPN label that was advertised in the unicast MAC advertisements, then the PE either forwards the packet based on CE next-hop forwarding information associated with the

label or does a destination MAC address lookup to forward the packet to a CE[RFC7432].

6. IANA considerations

TBD

7. Security Considerations

TBD

8. References

8.1. 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>.

[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>.

[RFC7543] Jeng, H., Jalil, L., Bonica, R., Patel, K., and L. Yong, "Covering Prefixes Outbound Route Filter for BGP-4", RFC 7543, DOI 10.17487/RFC7543, May 2015, <https://www.rfc-editor.org/info/rfc7543>.

[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>.

[RFC8317] Sajassi, A., Ed., Salam, S., Drake, J., Uttaro, J., Boutros, S., and J. Rabadan, "Ethernet-Tree (E-Tree) Support in Ethernet VPN (EVPN) and Provider Backbone Bridging EVPN (PBB-EVPN)", RFC 8317, DOI 10.17487/RFC8317, January 2018, <https://www.rfc-editor.org/info/rfc8317>.

[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>.

8.2. Informative References

Authors' Addresses

Zheng Fu
Huawei Technologies
No.101 Software Avenue, Yuhuatai District
Nanjing
210012
China

Email: fuzheng7@huawei.com

Tong Zhu
Huawei Technologies
No.101 Software Avenue, Yuhuatai District.
Nanjing
210012
China

Email: zhu.tong@huawei.com

Haibo Wang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China

Email: rainsword.wang@huawei.com

Jian Dai
Huawei Technologies
No.101 Software Avenue, Yuhuatai District
Nanjing
210012
China

Email: daijian2@huawei.com

Dawei Wang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China

Email: wang.dawei@huawei.com