BESS Working Group
Internet-Draft
Updates: 9014 (if approved)
Intended status: Standards Track
Expires: 26 December 2025

A. Sajassi
Cisco
J. Rabadan
Nokia
A. Nichol
Arista
L. Krattiger
K. Ananthamurthy
Cisco

24 June 2025

Applications and Procedures for Unknown MAC Route in EVPN
draft-sajassi-bess-evpn-umr-mobility-03

Abstract

The Interconnect Solution for Ethernet VPN defines Unknown MAC (Media Access Control) Route (UMR) utilization for Data Center Interconnect (DCI) when EVPN MPLS or EVPN VXLAN is an overlay network for such interconnects. This scenario impacts MAC mobility procedures and needs to be addressed. This document describes additional changes and enhancements required for MAC mobility procedures when using UMR..

Requirements Language

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.

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 26 December 2025.

Table of Contents

1. Terminology

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN. An EVI may be comprised of one BD (VLAN-based, VLAN Bundle, or Port-based services) or multiple BDs (VLAN-aware Bundle or Port-based VLAN-Aware services).

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'.

VID: VLAN Identifier.

Ethernet Tag: Used to represent a BD that is configured on a given ES for the purposes of DF election and <EVI, BD> identification for frames received from the CE. Note that any of the following may be used to represent a BD: VIDs (including Q-in-Q tags), configured IDs, VNIs (Virtual Extensible Local Area Network (VXLAN) Network Identifiers), normalized VIDs, I-SIDs (Service Instance Identifiers), etc., as long as the representation of the BDs is configured consistently across the multihomed PEs attached to that ES.

Ethernet Tag ID: Normalized network wide ID that is used to identify a BD within an EVI and carried in EVPN routes.

MP2MP: Multipoint to Multipoint.

MP2P: Multipoint to Point.

P2MP: Point to Multipoint.

P2P: Point to Point.

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.

BUM: Broadcast, unknown unicast, and multicast.

DF: Designated Forwarder.

Backup-DF (BDF): Backup-Designated Forwarder.

Non-DF (NDF): Non-Designated Forwarder.

AC: Attachment Circuit.

NVO: Network Virtualization Overlay as described in [RFC8365]

IRB: Integrated Routing and Bridging interface, with EVPN procedures described in [RFC9135]

2. Introduction

The Unknown MAC Route (UMR) is specified in RFC9014 and it is defined as 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 the Data Center Gateway’s Interconnect Ethernet Segment Identifier (I-ESI).

The use of a UMR for Layer 2 traffic is analogous to using a default IP route (e.g., 0.0.0.0/0 for IPv4 or ::/0 for IPv6) in Layer 3 forwarding. Just as a default IP route in the Forwarding Information Base (FIB) ensures that packets are forwarded to a remote PE when no more specific route exists, a UMR ensures that unknown Layer 2 destinations are forwarded to a remote PE. To conserve FIB resources on NVEs, Data Center Gateways advertise UMRs instead of numerous individual EVPN MAC/IP Advertisement routes - just as they advertise default IP routes instead of all specific IP prefixes. In addition, the UMR helps suppress unknown unicast flooding within EVPN Broadcast Domains. When combined with Proxy ARP/ND as defined in [RFC 9161], it significantly reduces overall flooding in the data center.

Although the basic procedures related to the UMR are described in [RFC 9014], several important applications of UMR are not covered which can lead to interoperability issues. These applications and use cases include inter-domain and its related mobility, Proxy ARP/ND [RFC 9161], hybrid UMR/non-UMR deployments, or symmetric and asymmetric IRB scenarios. Accordingly, this document updates [RFC 9014] to clarify and extend the proceduresfor UMR usage.

3. Baseline UMR Mechanisms for EVPN Broadcast Domains

The baseline UMR procedures can be categorized into:

1. Advertisement Procedures

2. Processing Procedures

3.1. UMR Advertisement Procedures

The UMR is typically advertised by Data Center Gateways or Service Gateways (generically GWs, hereafter) that extend Broadcast Domains to PEs in remote data centers. As specified in [RFC9014], GWs rely on local configuration to trigger the advertisement of a locally generated UMR. This configuration determines whether the gateway advertises only the UMR within its local domain, or the UMR along with EVPN MAC/IP Advertisement routes learned from remote domains. [RFC9014] recommends advertising only the UMR when MAC addresses in the local domain are learned through the control plane or management plane.

UMR routes are not subject to EVPN mobility procedures, and therefore SHOULD NOT be advertised along with the MAC Mobility Extended Community. Only non-UMR EVPN MAC/IP Advertisement routes follow mobility procedures, as specified later in this document. Additionally, UMRs are always originated by the gateways and MUST NOT be redistributed across domains.

UMR routes support the same EVPN multi-homing mechanisms as any other EVPN MAC/IP Advertisement routes. In particular, UMRs are advertised with a non-reserved ESI if the GWs are attached to an Interconnect Ethernet Segment (I-ES), as specified in [RFC9014]. The ESI associated with a UMR MUST always correspond to an Interconnect Ethernet Segment (I-ES) and SHOULD NOT represent the ESI of an Ethernet Segment directly connected to a CE.

3.2. UMR Processing Procedures

A PE within the data center that supports and processes the UMR forwards unknown unicast frames to the GW that advertised the UMR. If there are multiple valid UMR routes received with a zero ESI, or multiple UMRs received with different non-zero ESI, the PE selects the UMR based on best path selection, as in [I-D.ietf-bess-rfc7432bis]. PEs that do not support UMR will fallback to handling unknown unicast traffic as specified in [I-D.ietf-bess-rfc7432bis].

When multiple UMRs are received with the same non-reserved ESI, the standard multi-homing procedures defined in [I-D.ietf-bess-rfc7432bis] for all-active and single-active multi-homing apply, just as they do for any other EVPN MAC/IP Advertisement routes. As noted in Section 3.5.1 of [RFC 9014], the use of UMR helps resolve certain transient packet duplication issues that can occur in all-active multi-homing scenarios.

4. UMR Mix-Mode Operation

The procedures described previously illustrate the operation of the UMR route to provide improved MAC scaling when layer 2 only connectivity is required between domains. The deployment of a GW is not, however, limited to a layer 2 only topology, it can provide a mix of layer 2 and 3 connectivity or layer 3 only connectivity between domains.

This section of the document defines the GW and PE procedures for these different layer 2 and 3 deployment scenarios, which can be summarised into three categories.

* Layer 3 only: where a subnet(s) only exists within a single domain, but it’s associated IP-VRF is present across multiple domains

* Layer 2 only: where a MAC-VRF is present across multiple domains, but no IRBs are enabled for any of the BDs of the associated with the MAC-VRF

* Layer 2 and 3: where a MAC-VRF is present across multiple domains, with an IRB(s) enabled for BDs of the associated with the MAC-VRF

4.1. Layer 3 Only

In the Layer 3 only scenario, a subnet(s) will only be present within a single domain, with the associated IP-VRF for the subnet stretched across domains. In this scenario, type-5 (IP-Prefix) routes will be sufficient to provide layer 3 connectivity between the domains and externally. There is no requirement for layer 2 between domains or ARP suppression of non-local hosts in this scenario. The GW node that is providing the L3 connectivity between domains is therefore not required to advertise a UMR route or any type-2 (MAC-only, MAC-IP) route for hosts learnt in the remote domain.

4.2. Layer 2 Only

In the Layer 2 only scenario, where a MAC-VRF is present across domains but with no IRB enabled, the GW can advertise the UMR route into the local domain to represent any type-2 (MAC) route learnt in the remote domain. If ARP suppression for non-local hosts is required on the PE nodes themselves, even though they are operating in a L2 only mode (no IRB), the GW can follow the procedures outlined in section 5.1 of the document.

4.3. Layer 2 and 3

In a mixed Layer 2 and 3 topology, where MAC-VRFs are present across domains, with IRBs enabled on the associated BDs, for UMR operation there are additional procedures required on the GW and PE nodes to ensure correct IRB forwarding.

In a Symmetric IRB mode of operation, for any BD that is stretched across domains the EVPN GW will be required to advertise both a UMR route and a type-2 (MAC-IP) for any non-local host in the BD. Where the type-2 (MAC-IP) route is advertised with dual labels and route-targets which would be associated with the host’s BD and IP-VRF. For a UMR aware PE node, the UMR route will be installed in the L2 FIB, any subsequent type-2 (MAC-IP) route advertised by the GW for the BD will be installed in the RIB for ARP suppression purposes (see section 5.1) and in the L3 FIB for routing, the MAC address of the type-2 route should not be installed in the L2 FIB. For a UMR unaware PE node, the UMR route would not be installed in the L2 FIB, any type-2 (MAC-IP) route advertised by the GW for the BD would follow normal procedures, where the MAC would be installed in the L2 FIB , the RIB for ARP suppression purposes (see section 5.1) and in the L3 FIB for routing.

5. UMR MAC Mobility procedures

As discussed in the introduction section, the host IP default route and host unknown MAC route within a DC can be used to ensure that leaf nodes within a DC only learn and store host MAC and IP addresses for that DC. All other hosts MAC and IP addresses from remote DCs are learned and stored in DC Gateway (GW) nodes thus alleviating leaf nodes from learning host MAC and IP addresses from the remote DCs and potentially improving the scale of MAC and IP addresses on leaf nodes by one to two orders of magnitude.

Interconnect Solution for Ethernet VPN [RFC9014] defines Unknown MAC Route (UMR) utilization for Data Center Interconnect (DCI) when EVPN MPLS or EVPN VXLAN is used as an overlay network for such interconnects. The introduction of UMR for such scenarios impacts

the MAC mobility procedures that are not discussed in [RFC9014]. This document describes additional changes and enhancements needed for MAC mobility procedures when UMR is used.

The following paragraphs lists the requirements for MAC mobility solution when UMR is used.

* When an EVPN overlay network (i.e., DC, Enterprise, or SP network) is enabled for UMR operation, then all the Provider Edges (PEs) or leaf nodes and all the Gateways (border leaf nodes) in that network MUST support UMR operation - e.g., a DC network cannot have some leaf nodes supporting UMR operation and some other leaf nodes incapable of UMR operation. This means when upgrading a DC network for UMR capability, all leaf nodes (PEs), all border leaf nodes (Gateways), and all Route Reflectors (RRs) in that network needs to be upgraded before turning on the UMR capability. If desired, a physical DC network can be partitioned into two logical networks with one supporting UMR operation and the other not supporting it.

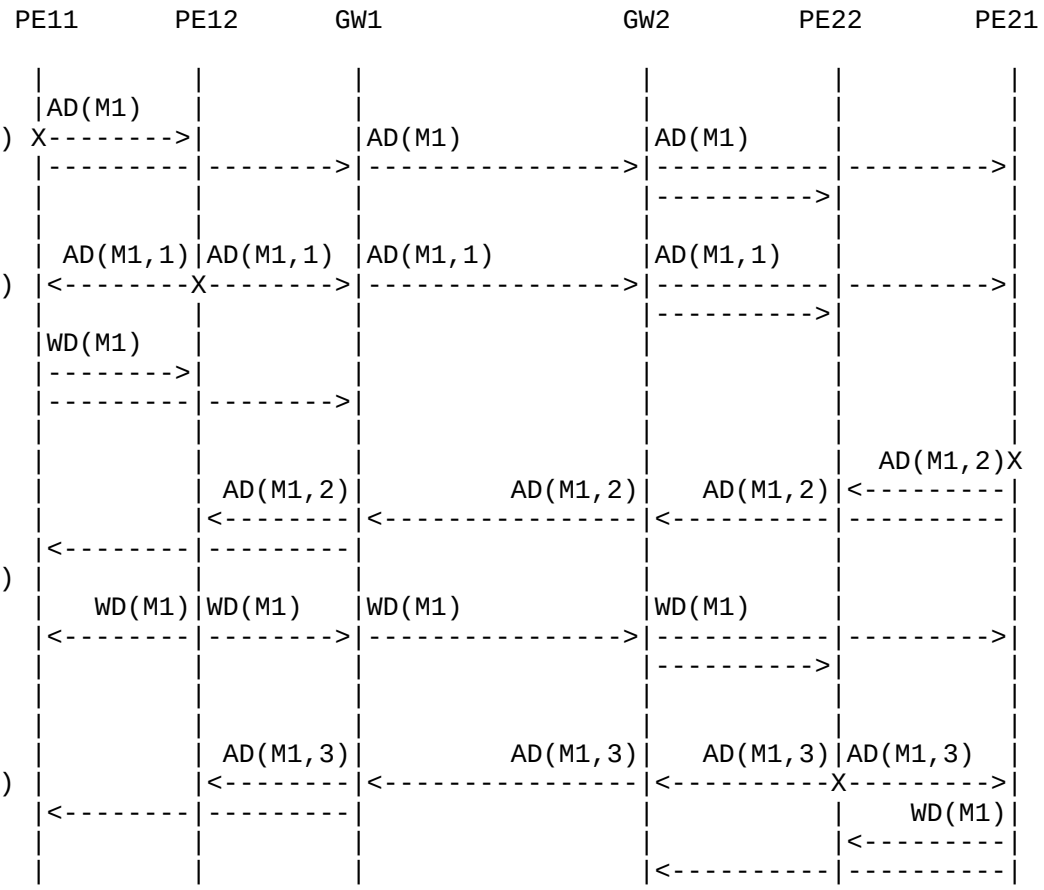
* UMR MAC mobility procedures for DC networks that are UMR capable MUST operate seamlessly with DC networks that are UMR incapable.

* UMR operation is optional and a PE device (a leaf node) that supports UMR procedure but doesn’t receive the UMR route from its Gateway (its border leaf), SHALL operate per baseline [RFC7432]. This does not mean that a DC network can partially operate in UMR mode but rather it means that a DC network can gradually be upgraded for UMR capability and once the entire network is upgraded, then it can operate in UMR mode.

Section 5.1 ("Inter-network MAC Mobility procedures without UMR") discusses MAC Mobility for DCI operation without UMR utilization. Section 5 discusses MAC Mobility for DCI operation with UMR and the modifications and enhancements needed on top of the baseline operation discussed in section 4.

5.1. Inter-network MAC Mobility procedures without UMR

In order to better differentiate the enhancements needed to MAC Mobility procedures for networks interconnect scenarios with utilization of UMR which is missing in [RFC9014], we first start with the description of the baseline MAC Mobility procedures (without utilization of UMR) in this section and then we describe the changes needed on top of this baseline scenario in the next section. The following ladder diagram is used to help in describing baseline MAC Mobility operation.



AD(M1,x) = MAC/IP Route Advertisement for MAC M1 with seq# x
WD(M1) = MAC/IP Route Withdrawl for MAC M1
X = Host M1 being local to that PE

This diagram depicts MAC Mobility procedures between two overlay networks which in turn are connected via a WAN network; where, GW1 and GW2 sit at the edge of the WAN. The first network consists of PE11, PE12, and GW1. The second network consists of PE21, PE22, and GW2. EVPN control plane is used both within each network as well as between them.

1. PE11 learns host M1 for the first time and it advertises it via MAC/IP Route Advertisement to other nodes in its DC network participating in that EVI. Since this is the first time that PE11 learns of M1, it sends the advertisement without MAC Mobility extended community attribute per section 15 of

[RFC7432]. When the local Gateway (GW1 of DC1) receives this advertisement, it readvertises this MAC address per procedures of [RFC9014] without MAC Mobility extended community attribute. The remote Gateway (GW2 of DC2) receives this advertisement and it in turn readvertises it to its PEs participating in that EVI. At this point, remote PE21 and PE22 know that the next hop for M1 is GW2, and GW2 knows that the next hop for M1 is GW1.

2. In this step, host M1 makes an intra-DC move within DC1 network and moves from PE11 to PE12, the PE12 follows the MAC mobility procedures in [RFC7432] and advertises the MAC/IP Advertisement route for M1 with a sequence number which is incremented by one (in this case seq = 1). PE11 and GW1 receive this advertisement and update their next hop for M1 to point to PE12. GW1 follows the procedures in [RFC9014] and it readvertises this route with this new sequence number received from PE12. Upon receiving this route, GW2 updates its sequence number for M1 and in turn it readvertises this route to its DC. PE21 and PE22 receive this advertisement and update their sequence numbers for M1 (seq = 1), but there is no change to the next hop for M1 and they keep it as GW2.

   Furthermore, after verifying that M1 is no longer present locally, PE11 sends a withdrawal message for M1 to all local PEs and GWs that are participating in that EVI per MAC Mobility procedures of [RFC7432]. When PE12 and GW1 receive this withdrawal message, they clean up their BGP tables and remove BGP EVPN route for M1 received from PE11. Since BGP table in GW1 has at least one BGP EVPN route learned from its DC1 (and host M1 has as its next hop one of the PEs in DC1), GW1 does not readvertise the withdrawal message for M1 received from PE11.

3. In this step, host M1 moves from PE12 in DC1 to PE21 in DC2. PE21 upon learning M1 locally, it advertises the MAC/IP Advertisement route for M1 with a sequence number which is incremented by one (in this case seq = 2). PE22 and GW2 receive this advertisement, recognize the move, and update their next hop for M1 to point to PE21. GW2 follows the procedures in [RFC9014] and it readvertises this route with this new sequence number received from PE21. Upon receiving this route, GW1 updates its sequence number for M1 and in turn it readvertises this route to its DC1 network. PE11 and PE12 receive this advertisement, recognize the move, and update their next hop to point to GW1, and their sequence numbers for M1.

   Furthermore, after verifying that M1 is no longer present locally, PE12 sends a withdrawal message for M1 to all local PEs and GWs that are participating in that EVI. When PE11 and GW1

receive this withdrawal message, they clean up their BGP tables and remove BGP EVPN route for M1 received from PE12. Since there is no more local BGP EVPN routes for M1 in BGP table of GW1 (i.e., no more routes from its local PEs), it readvertises this withdrawal message to other GWs over WAN. When other GWs over WAN (including GW2) receive this withdrawal message, they remove the BGP EVPN route for M1 received from GW1. At this point the only BGP EVPN route entry in GW1 is the one received from GW2, and for GW2 is the one received from its local PE21.

4. This step is similar to that of step 2 and demonstrates what it takes place when an intra-DC move happens but this time within DC2 where the host M1 moves from PE21 to PE22. The PE22 follows the MAC mobility procedures in [RFC7432] and advertises the MAC/IP Advertisement route for M1 with a sequence number which is incremented by one (in this case seq = 3). PE21 and GW2 receive this advertisement and update their next hop for M1 to point to PE22. GW2 follows the procedures in [RFC9014] and it readvertises this route with this new sequence number received from PE22. Upon receiving this route, GW1 updates its sequence number for M1 and in turn it readvertises this route to its DC. PE11 and PE12 receive this advertisement and update their sequence numbers for M1 (seq = 3), but there is no change to the next hop for M1 and they keep it as GW1.

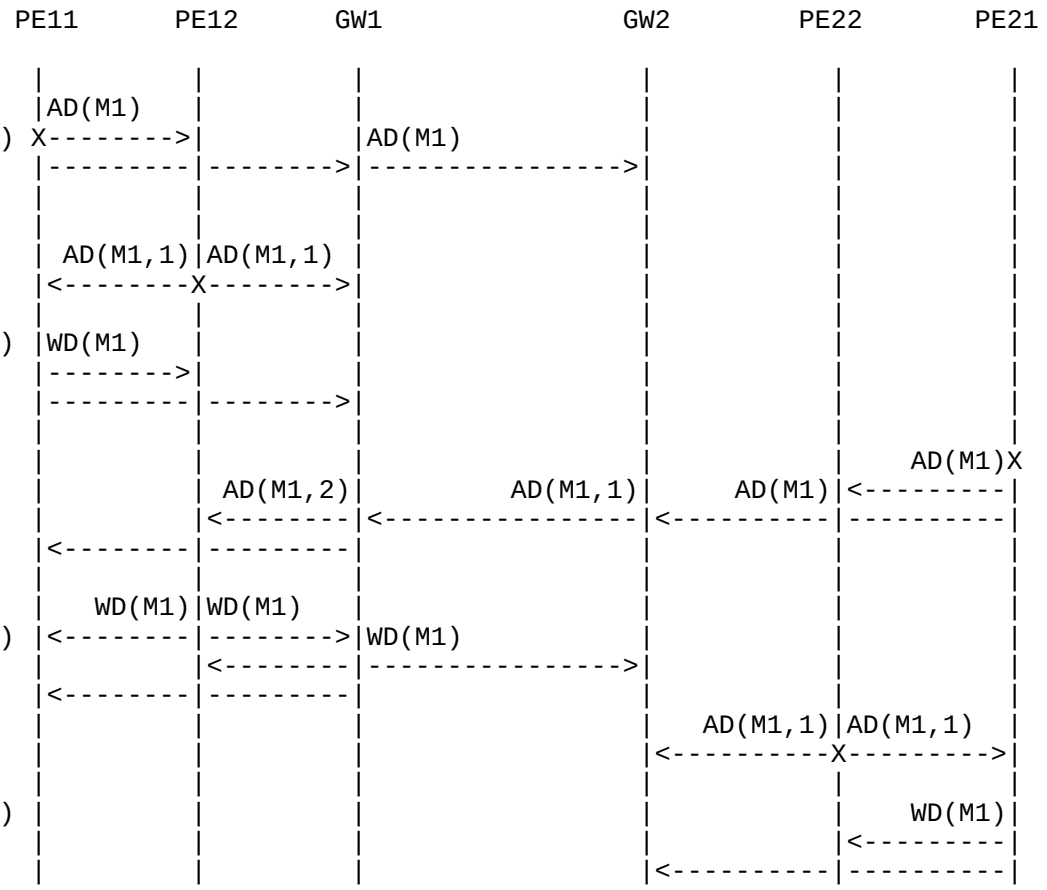
   Furthermore, after verifying that M1 is no longer present locally, PE21 sends a withdrawal message for M1 to all local PEs and GWs that are participating in that EVI per MAC Mobility procedures of [RFC7432]. When PE22 and GW2 receive this withdrawal message, they clean up their BGP tables and remove BGP EVPN route for M1 received from PE21. Since BGP table in GW2 has at least one BGP EVPN route learned from its DC2 (and host M1 has as its next hop one of the PEs in DC2), GW2 does not readvertise the withdrawal message for M1 received from PE21.

5.2. Inter-network MAC Mobility Procedures for UMR

This section describes the changes needed to MAC Mobility procedures when UMR is utilized in EVPN overlay networks and hosts are allowed to move between these networks. Since advertisement of MAC/IP addresses from the local overlay network are not propagated all the way to the PEs of the remote overlay network, the baseline MAC mobility procedures described in the previous section, cannot be used as is and it needs to be modified. When a host M1 moves from one overlay network (e.g., DC1) to anoher one (e.g., DC2), the PE in DC2 (PE21) that learns the host locally, it learns it for the very first time because of UMR operation - i.e., M1 never previously got advertised by DC2 GW (GW2). Therefore, the PE21 advertises EVPN MAC/

IP route for M1 without any sequence number which breaks the baseline MAC mobility procedures described in the previous section. In order to accommodate MAC mobility in the presence of UMR, each needs to maintain two sequence numbers per host MAC address - one for its local overlay network (e.g., its DC network) and another one for the interconnect network (e.g., WAN network). Only Gateways need to maintain both MAC Mobility sequence numbers. The PEs that are enabled for UMR operation, only need to maintain a single MAC Mobility sequence number per MAC address. These two sequence numbers operate independently (i.e., they get incremented independently) so that a local MAC move within an overlay network (e.g., DC1) does not impact other overlay networks (e.g., other DCs) and the interconnect network (e.g., WAN network) - i.e., when the host mobility is confined to a DC (aka intra-DC host mobility), then only the intra-DC MAC Mobility counter for that DC is incremented upon host move without any changes to inter-DC MAC Mobility counter or any other intra-DC MAC Mobility counters in other DCs. However, when the host moves from one DC to another, then the inter-DC MAC Mobility counter is impacted.

The solution described in this section optimizes based on convergence time and number of BGP EVPN route advertisements - i.e., it tries to minimize the convergence time upon a host move and to minimize the number of EVPN route advertisements. Whenever these two factors are in conflict, the preference is given to minimizing the convergence time. The following ladder diagram is used to help in describing MAC Mobility procedures for UMR operation.



AD(M1,x) = MAC/IP Route Advertisement for MAC M1 with seq# x
WD(M1) = MAC/IP Route Withdrawl for MAC M1
X = Host M1 being local to that PE

This diagram depicts MAC Mobility procedures between two overlay networks which in turn are connected via a WAN network where UMR operation is utilized. To avoid repeating the text verbatim from previous section and put the emphasis on the new procedures, we mainly elaborate on the changes relative to the baseline MAC mobility procedures described in the previous section.

When UMR operation is enabled for a given EVI, all Gateways participating in that EVI for that overlay network, advertise the UMR route to their local overlay network. A PE that is capable of UMR processing, upon receiving the UMR route, activates its UMR procedure as described below. When a Gateway receives the UMR route from

another Gateway for one of its EVI for which UMR operation is enabled, it should simply discard it (i.e., not to add it to its BGP table and MAC-VRF).

1. When the host M1 is learned in DC1 for the first time, the baseline MAC Mobility procedure described in step (1) of Section 5.1 is executed in DC1 among PE11, PE12, and GW1. When the remote Gateway (GW2 of DC2) receives this advertisement from GW1, it processes it just as step (1) of Section 5.1 and adds it to its BGP table and MAC-VRF for that EVI. However, it does not readvertise it into its own DC network because it is configured for UMR operation and no remote MAC/IP Advertisement routes (routes received from remote GWs) are ever readvertised locally.

2. When the host M1 makes an intra-DC move within DC1 network, the baseline MAC Mobility procedure described in step (2) of Section 5.1 is executed in DC1 among PE11, PE12, and GW1. GW1 realizes that this is an intra-overlay network MAC move (intra-DC MAC move) and thus it does not readvertises this route to other GWs in the WAN network. It should be noted that GW1 maintains two sequence numbers for M1 and it increments its intra-DC sequence number by one (seq = 1); however, it leaves its inter-DC sequence number unchanged (seq = 0).

   Generation of the withdrawal message by PE11 and processing of this message by other EVPN devices in DC1 (i.e., PE12 and GW1) are the same as the ones described in step (2) of Section 5.1. Since after receiving the withdrawal message and cleaning up its BGP table, GW1 still has at least one BGP EVPN route from its local DC1 (and host M1 has as its next hop one of the PEs in DC1), GW1 does not readvertise the withdrawal message for M1 received from PE11 to remote Gateways (e.g., GW2 of DC2).

3. When the host M1 moves from DC1 to DC2 and its presence is detected locally by the PE21, the PE21 learns M1 for the very first time since its Gateway (GW2) never advertised MAC/IP route for M1 to its local PEs (because of UMR operation). The PE21 advertises MAC/IP route for M1 without any sequence number. All PEs and Gateways in DC2 upon receiving this advertisement update their BGP and MAC-VRF tables. In addition to this update, the GW2 recognizes that there has been a MAC move, increments its inter-DC MAC Mobility counter for M1, and it readvertises this MAC/IP route along with the updated MAC Mobility extended community.

   GW1 receives this MAC/IP advertisement for M1 and it also recognizes that M1 has moved to DC2. GW1 increments its intra-DC MAC Mobility sequence number and it readvertises this route along

with the updated MAC Mobility extended community (seq = 2) to its local DC for that EVI. As the result, all the PEs participating for that EVI in DC1, receive this MAC/IP advertisement and update their BGP and MAC-VRF tables. They also update the BGP next hop to point to GW1 for MAC address M1. Besides this update, PE12 recognizes the MAC move and advertises a withdrawal message for M1. Furthermore, it verifies that M1 has actually moved and is no longer present locally.

When the Gateways and other PEs in DC1 receive this withdrawal message from PE12, they cleanup their BGP tables and remove the corresponding M1 entry from their tables. After this cleanup, GW1 realizes that there is no more entry for M1 in its BGP table from its local PEs and thus it sends a withdrawal message for M1 to all its local PEs and remote Gateways (e.g., GW2). Furthermore, GW1 must reset its intra-DC MAC mobility counter for M1 to zero because M1 no longer exist among its local PEs. When the local PEs (PE11 and PE12) receive this withdrawal message, they clean up their BGP and MAC-VRF tables for M1. After the cleanup, there should be no entry in BGP and MAC-VRF tables for M1 and thus the forwarding for M1 must follow UMR operation - i.e., the packet with the destination MAC address of M1 must be load balanced to one of the GWs that has advertised UMR route. When the remote Gateways receive this withdrawal message, they clean up their BGP tables for M1 and the only entry in BGP table for M1 should be that of the one received from GW2.

4. This step demonstrates an intra-DC MAC move for DC2. The procedure for the PEs (PE21 and PE22) and the corresponding Gateway (GW2) is the same as the one described in step 2 and thus no further explanation is needed.

Redundant Gateways are supported by the described procedure. All the redundant Gateways attached to a given BD advertise the EVPN MAC/IP Advertisement routes with the same Interconnect ESI [RFC9014], and all the redundant Gateways MUST use the same sequence numbers when advertising MAC addresses to either their local overlay network or their interconnect network.

5.3. Duplicate MAC Address Detection

Duplicate MAC addresses can occur as described in section 15.1 of [RFC7432]. MAC address duplication can happen within the same DC network (e.g., DC1) or across different DC networks (e.g., DC1 and DC2) where UMR is utilized. In either case, the procedure is the same as the one described in section 15.1 of [RFC7432]. More specifically, the timer and the move counter for a given MAC are kept only at the PEs - i.e., there is no need to maintain such timer and

move counter for a given MAC unless that MAC is learned locally on that GW.

5.4. MAC Mobility for Gateway Local Attachment Circuits

This section describes MAC Mobility procedures for hosts sitting behind local Attachment Circuts (ACs) of a Gateway and moving to/from a local PE, or another Gateway in a remote DC, or a remote PE in a remote DC. TBD.

6. Impact of UMR on EVPN Use Cases

This section examines and specifies how the UMR interacts with other EVPN procedures and specifications.

6.1. UMR and EVPN Proxy ARP/ND

The Proxy ARP/ND functionality, as defined in [RFC9161], is widely deployed in EVPN Broadcast Domains. When enabled, it optimizes IP address resolution, enhances security by mitigating ARP/ND spoofing attacks, and significantly reduces - or even eliminates - ARP/ND flooding within the Broadcast Domain.

Proxy ARP/ND is also supported in Broadcast Domains that span interconnected data centers. In interconnect solutions as described in [RFC9014], Proxy ARP/ND relies on the advertisement and end-to-end propagation of EVPN MAC/IP Advertisement routes for all hosts. This allows a PE to populate its Proxy ARP/ND table with IP-MAC bindings for both local domain and remote domain hosts. As a result, when a local ARP Request or Neighbor Solicitation is received for a remote domain host IP, the PE can respond locally, effectively preventing ARP/ND flooding across data center domains. However, if the GWs are configured to advertise the UMR while suppressing the redistribution of EVPN MAC/IP Advertisement routes from remote domains into the local domain, the Proxy ARP/ND function will not be able to prevent ARP/ND flooding toward those remote domains. [RFC9014] does not define the interaction between Proxy ARP/ND and the UMR, which may lead readers to assume that Proxy ARP/ND cannot be used in conjunction with the UMR when the goal is to suppress ARP/ND flooding to PEs in remote domains.

This document proposes two solutions that enable the use of the UMR while still suppressing ARP/ND flooding to remote domains:

1. Advertise the UMR and redistribute EVPN MAC/IP Advertisement routes, while suppressing the programming of MAC addresses into the Bridge Table of leaf routers in the local domain. This approach is referred to as the "distributed Proxy ARP/ND solution".

2. Advertise only the UMR, and enable Proxy ARP/ND on the gateway to handle remote IP resolution. This approach is referred to as the "GW-based Proxy ARP/ND solution"

The two methods are described in the sections 6.1.1 and 6.1.2. By default, the GW operates in either mode based on local configuration. However, a dynamic mode - described in Section 6.1.3 - allows the GW to select its behavior based on the UMR capability advertised by other routers within the local domain.

6.1.1. Distributed Proxy ARP/ND Solution

In this approach, the GW is configured to advertise both the UMR and the EVPN MAC/IP Advertisement routes for hosts in remote domains to nodes within the local domain. When operating in this mode, the GW MUST redistribute only those EVPN MAC/IP Advertisement routes from remote domains that include a valid IP address. Routes containing only a MAC address with a zero IP address SHOULD NOT be redistributed, as they are not usable by routers in the local domain for performing Proxy ARP/ND functions.

When a domain router receives and programs a UMR from a GW, it continues to import EVPN MAC/IP Advertisement routes as usual. However, for routes received from the same GW that advertised the UMR, the router does not program the MAC addresses into the Bridge Table. Instead, it uses the information solely to populate the Proxy ARP/ND table with IP-MAC bindings. This approach preserves Bridge Table FIB resources - a key objective of the UMR mechanism - while still enabling Proxy ARP/ND functionality, thereby preventing ARP/ND flooding within the local domain.

6.1.2. GW-based Proxy ARP/ND Solution

In this approach, the GW is configured to advertise only the UMR while suppressing the redistribution of all EVPN MAC/IP Advertisement routes received from remote domains. As a result, domain routers do not receive the IP-MAC bindings for hosts in remote domains. Consequently, when these routers receive an ARP Request or Neighbor Solicitation for a remote IP, they cannot respond via their local proxy ARP/ND function. Instead, the request is flooded within the EVPN Broadcast Domain. When the request reaches the GW, it performs a lookup in its local proxy ARP/ND table. If a matching entry is

found, the GW responds with an ARP Reply or Neighbor Advertisement for the target IP. While this method does not suppress ARP/ND flooding entirely within the local domain, it does prevent flooding across the EVPN Broadcast Domain to remote PEs, thus achieving partial flooding suppression.

6.1.3. Dynamic Redistribution of EVPN MAC/IP Advertisement Routes based on UMR capability

The methods described in Sections 6.1.1 and 6.1.2 rely on static configuration of the GWs and domain routers to operate in a specific mode. As an alternative, dynamic signaling can be used to allow the GW to adjust its mode of operation based on received UMR capability information. In this dynamic approach, domain routers signal their support for UMR processing using the UMR Capability flag in the Inclusive Multicast Ethernet Tag route for the Broadcast Domain. The GW inspects this flag across all participating routers in the local domain (for the Broadcast Domain).

* If all local domain routers in the Broadcast Domain have the UMR Capability flag set, the GW MAY operate in either mode (Distributed or GW-based proxy ARP/ND).

* If any router in the local domain does not set the UMR Capability flag, the GW MUST operate in Distributed Proxy ARP/ND mode. This requirement exists because routers that do not support UMR processing depend on receiving EVPN MAC/IP Advertisement routes from remote domains to prevent constant ARP/ND flooding for traffic destined to remote hosts.

6.2. UMR and Silent Hosts

This section analyzes the behavior and potential implications of using UMR procedures when silent hosts exist in the local domain.

Consider the case where a host (H1), with IP address IP1 and MAC address M1, initiates communication with another host (H2), which has been silent until that moment and is identified by IP address IP2 and MAC address M2. When H1 attempts to reach H2, it sends an ARP Request for IP2. This request is flooded across all routers in the domain and eventually reaches H2, which responds with an ARP Reply containing its MAC address (M2). At the same time the ARP Reply is forwarded to the EVPN router connected to H1 (PE1), the EVPN router attached to H2 (PE2) triggers the advertisement of an EVPN MAC/IP Advertisement route for M2 and IP2. Because the ARP Reply typically reaches PE1 before the corresponding MAC/IP Advertisement route is received and programmed, PE1 may receive a unicast packet from H1 to H2 (i.e., M2) while M2 is still unknown in the Bridge Table. During this brief window, PE1 may treat the packet as unknown unicast.

If PE1 supports the UMR procedure, it will forward the unknown unicast packet to the GW that advertised the UMR. In this scenario, the GW may drop a few packets destined to H2 until the MAC address (M2) is learned and programmed at PE1. Once M2 is present in PE1’s Bridge Table, the traffic is no longer treated as unknown, and subsequent packets are forwarded directly to PE2 as expected. While the example is described for IPv4 and ARP resolution, the same happens for IPv6 and Neighbor Discovery.

This example highlights how the use of UMR can introduce race conditions in environments with silent hosts, potentially resulting in temporary packet loss for initial traffic flows.

If silent hosts are located in remote domains, unknown unicast traffic will be forwarded to the GW by design, so no unexpected packet drops occur in this case.

7. Security Considerations

Since this document describes how to address MAC mobility issue as the result of using UMR for interconnection solutions of [RFC9014], and since no new requiremnts with respect to mobility procedures are introduced at the edge devices (e.g., PEs or leafs), there is no additional security risks beyond the ones described in [RFC7432] and [RFC8365].

8. IANA Considerations

This document does not introduce any IANA requirements.

9. References

9.1. 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>.

[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>.

[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>.

9.2. Informative References

[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>.

[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>.

Authors' Addresses

Ali Sajassi
Cisco
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: sajassi@cisco.com

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

Alex Nichol
Arista
5453 Great America Parkway
Santa Clara, CA 95054
United States of America
Email: anichol@arista.com

Lukas Krattiger
Cisco
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: lkrattig@cisco.com

Krishnaswamy Ananthamurthy
Cisco
170 W. Tasman Drive
San Jose, CA 95134
United States of America
Email: kriswamy@cisco.com