Internet Engineering Task Force (IETF)
Request for Comments: 9574
Category: Standards Track
ISSN: 2070-1721

J. Rabadan, Ed.
S. Sathappan
Nokia
W. Lin
Juniper Networks
M. Katiyar
Versa Networks
A. Sajassi
Cisco Systems

May 2024

Optimized Ingress Replication Solution for Ethernet VPNs (EVPNs)

Abstract

Network Virtualization Overlay (NVO) networks using Ethernet VPNs (EVPNs) as their control plane may use trees based on ingress replication or Protocol Independent Multicast (PIM) to convey the overlay Broadcast, Unknown Unicast, or Multicast (BUM) traffic. PIM provides an efficient solution that prevents sending multiple copies of the same packet over the same physical link; however, it may not always be deployed in the NVO network core. Ingress replication avoids the dependency on PIM in the NVO network core. While ingress replication provides a simple multicast transport, some NVO networks with demanding multicast applications require a more efficient solution without PIM in the core. This document describes a solution to optimize the efficiency of ingress replication trees.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9574.

Copyright Notice

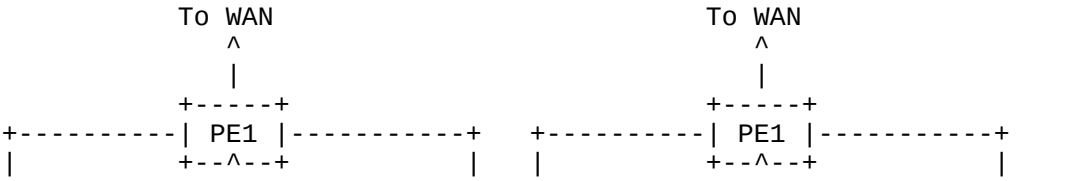
Table of Contents

1. Introduction

Ethernet Virtual Private Networks (EVPNs) may be used as the control plane for a Network Virtualization Overlay (NVO) network [RFC8365]. Network Virtualization Edge (NVE) and Provider Edge (PE) devices that are part of the same EVPN Broadcast Domain (BD) use Ingress Replication (IR) or PIM-based trees to transport the tenant's Broadcast, Unknown Unicast, or Multicast (BUM) traffic.

In the ingress replication approach, the ingress NVE receiving a BUM frame from the Tenant System (TS) will create as many copies of the frame as the number of remote NVEs/PEs that are attached to the BD. Each of those copies will be encapsulated into an IP packet where the outer IP Destination Address (IP DA) identifies the loopback of the egress NVE/PE. The IP fabric core nodes (also known as spines) will simply route the IP-encapsulated BUM frames based on the outer IP DA. If PIM-based trees are used instead of ingress replication, the NVEs/PEs attached to the same BD will join a PIM-based tree. The ingress NVE receiving a BUM frame will send a single copy of the frame, encapsulated into an IP packet where the outer IP DA is the multicast address that represents the PIM-based tree. The IP fabric core nodes are part of the PIM tree and keep multicast state for the multicast group, so that IP-encapsulated BUM frames can be routed to all the NVEs/PEs that joined the tree.

The two approaches are illustrated in Figure 1. On the left-hand side of the diagram, NVE1 uses ingress replication to send a BUM frame (originated from Tenant System TS1) to the remote nodes attached to the BD, i.e., NVE2, NVE3, and PE1. On the right-hand side, the same example is depicted but using a PIM-based tree, i.e., (S1,G1), instead of ingress replication. While a single copy of the tunneled BUM frame is generated in the latter approach, all the routers in the fabric need to keep multicast state, e.g., the spine keeps a PIM routing entry for (S1,G1) with an Incoming Interface (IIF) and three Outgoing Interfaces (OIFs).



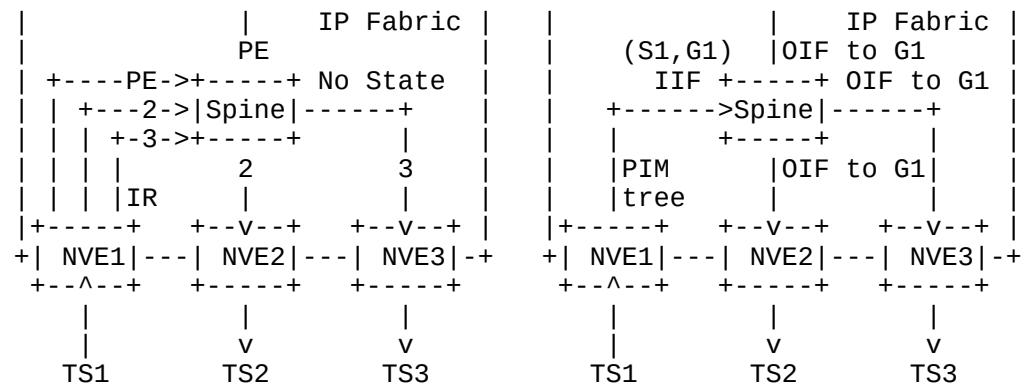


Figure 1: Ingress Replication vs. PIM-Based Trees in NVO Networks

In NVO networks where PIM-based trees cannot be used, ingress replication is the only option. Examples of these situations are NVO networks where the core nodes do not support PIM or the network operator does not want to run PIM in the core.

In some use cases, the amount of replication for BUM traffic is kept under control on the NVEs due to the following fairly common assumptions:

a. Broadcast traffic is greatly reduced due to the proxy Address Resolution Protocol (ARP) and proxy Neighbor Discovery (ND) capabilities supported by EVPNs [RFC9161] on the NVEs. Some NVEs can even provide Dynamic Host Configuration Protocol (DHCP) server functions for the attached TSs, reducing the broadcast traffic even further.

b. Unknown unicast traffic is greatly reduced in NVO networks where all the Media Access Control (MAC) and IP addresses from the TSs are learned in the control plane.

c. Multicast applications are not used.

If the above assumptions are true for a given NVO network, then ingress replication provides a simple solution for multi-destination traffic. However, statement c. above is not always true, and multicast applications are required in many use cases.

When the multicast sources are attached to NVEs residing in hypervisors or low-performance-replication Top-of-Rack (ToR) switches, the ingress replication of a large amount of multicast traffic to a significant number of remote NVEs/PEs can seriously degrade the performance of the NVE and impact the application.

This document describes a solution that makes use of two ingress replication optimizations:

1. Assisted Replication (AR)

2. Pruned Flooding Lists (PFLs)

Assisted Replication consists of a set of procedures that allows the ingress NVE/PE to send a single copy of a broadcast or multicast frame received from a TS to the BD without the need for PIM in the underlay. Assisted Replication defines the roles of AR-REPLICATOR and AR-LEAF routers. The AR-LEAF is the ingress NVE/PE attached to the TS. The AR-LEAF sends a single copy of a broadcast or multicast packet to a selected AR-REPLICATOR that replicates the packet multiple times to remote AR-LEAF or AR-REPLICATOR routers and is therefore "assisting" the ingress AR-LEAF in delivering the broadcast or multicast traffic to the remote NVEs/PEs attached to the same BD.

Assisted Replication can use a single AR-REPLICATOR or two AR-REPLICATOR routers in the path between the ingress AR-LEAF and the remote destination NVEs/PEs. The procedures that use a single AR-REPLICATOR (the non-selective Assisted Replication solution) are specified in Section 5, whereas Section 6 describes how multi-stage replication, i.e., two AR-REPLICATOR routers in the path between the ingress AR-LEAF and destination NVEs/PEs, is accomplished (the selective Assisted Replication solution). The procedures for Assisted Replication do not impact unknown unicast traffic, which follows the same forwarding procedures as known unicast traffic so that packet reordering does not occur.

PFLs provide a method for the ingress NVE/PE to prune or remove certain destination NVEs/PEs from a flooding list, depending on the interest of those NVEs/PEs in receiving BUM traffic. As specified in [RFC8365], an NVE/PE builds a flooding list for BUM traffic based on the next hops of the received EVPN Inclusive Multicast Ethernet Tag routes for the BD. While [RFC8365] states that the flooding list is used for all BUM traffic, this document allows pruning certain next hops from the list. As an example, suppose an ingress NVE creates a flooding list with next hops PE1, PE2, and PE3. If PE2 and PE3 did not signal any interest in receiving unknown unicast traffic in their Inclusive Multicast Ethernet Tag routes, when the ingress NVE receives an unknown unicast frame from a TS, it will replicate it only to PE1. That is, PE2 and PE3 are "pruned" from the NVE's flooding list for unknown unicast traffic. PFLs can be used with ingress replication or Assisted Replication and are described in Section 7.

Both optimizations -- Assisted Replication and PFLs -- may be used together or independently so that the performance and efficiency of the network to transport multicast can be improved. Both solutions require some extensions to the BGP attributes used in [RFC7432]; see Section 4 for details.

The Assisted Replication solution described in this document is focused on NVO networks (hence its use of IP tunnels). MPLS transport networks are out of scope for this document. The PFLs solution MAY be used in NVO and MPLS transport networks.

Section 3 lists the requirements of the combined optimized ingress replication solution, whereas Sections 5 and 6 describe the Assisted Replication solution for non-selective and selective procedures, respectively. Section 7 provides the PFLs solution.

2. 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.

The following terminology is used throughout this document:

AR-IP: Assisted Replication - IP. Refers to an IP address owned by the AR-REPLICATOR and used to differentiate the incoming traffic that must follow the AR procedures. The AR-IP is also used in the Tunnel Identifier and Next Hop fields of the Replicator-AR route.

AR-LEAF: Assisted Replication - LEAF. Refers to an NVE/PE that sends all the BM traffic to an AR-REPLICATOR that can replicate the traffic further on its behalf. An AR-LEAF is typically an NVE/PE with poor replication performance capabilities.

AR-REPLICATOR: Assisted Replication - REPLICATOR. Refers to an NVE/

PE that can replicate broadcast or multicast traffic received on overlay tunnels to other overlay tunnels and local Attachment Circuits (ACs). This document defines the control and data plane procedures that an AR-REPLICATOR needs to follow.

AR-VNI: Assisted Replication - VNI. Refers to a Virtual eXtensible Local Area Network (VXLAN) Network Identifier (VNI) advertised by the AR-REPLICATOR along with the Replicator-AR route. It is used to identify the incoming packets that must follow the AR procedures ONLY in the single-IP AR-REPLICATOR case (see Section 8).

Assisted Replication forwarding mode: In the case of an AR-LEAF, sending an AC Broadcast and Multicast (BM) packet to a single AR-REPLICATOR with a tunnel destination address AR-IP. In the case of an AR-REPLICATOR, this means sending a BM packet to a selected number of, or all of, the overlay tunnels when the packet was previously received from an overlay tunnel.

BD: Broadcast Domain, as defined in [RFC7432].

BD label: Defined as the MPLS label that identifies the BD and is advertised in Regular-IR or Replicator-AR routes, when the encapsulation is MPLS over GRE (MPLSoGRE) or MPLS over UDP (MPLSoUDP).

BM traffic: Refers to broadcast and multicast frames (excluding unknown unicast frames).

DF and NDF: Designated Forwarder and Non-Designated Forwarder. These are roles defined in NVEs/PEs attached to multihomed TSs, as per [RFC7432] and [RFC8365].

ES and ESI: Ethernet Segment and Ethernet Segment Identifier. EVPN multihoming concepts as specified in [RFC7432].

EVI: EVPN Instance. A group of Provider Edge (PE) devices participating in the same EVPN service, as specified in [RFC7432].

GRE: Generic Routing Encapsulation [RFC4023].

Ingress Replication forwarding mode: Refers to the ingress replication behavior explained in [RFC7432]. In this mode, an AC BM packet copy is sent to each remote PE/NVE in the BD, and an overlay BM packet is sent only to the ACs and not to other overlay tunnels.

IR-IP: Ingress Replication - IP. Refers to the local IP address of an NVE/PE that is used for the ingress replication signaling and procedures provided in [RFC7432]. Encapsulated incoming traffic with an outer destination IP address matching the IR-IP will follow the procedures for ingress replication and not the procedures for Assisted Replication. The IR-IP is also used in the Tunnel Identifier and Next Hop fields of the Regular-IR route.

IR-VNI: Ingress Replication - VNI. Refers to a VNI advertised along with the Inclusive Multicast Ethernet Tag route for the ingress replication tunnel type.

MPLS: Multi-Protocol Label Switching.

NVE: Network Virtualization Edge [RFC8365].

NVGRE: Network virtualization using Generic Routing Encapsulation [RFC7637].

PE: Provider Edge.

PMSI: P-Multicast Service Interface. A conceptual interface for a PE to send customer multicast traffic to all or some PEs in the same VPN [RFC6513].

RD: Route Distinguisher.

Regular-IR route: An EVPN Inclusive Multicast Ethernet Tag route [RFC7432] that uses the ingress replication tunnel type.

Replicator-AR route: An EVPN Inclusive Multicast Ethernet Tag route that is advertised by an AR-REPLICATOR to signal its capabilities, as described in Section 4.

RNVE: Regular NVE. Refers to an NVE that supports the procedures provided in [RFC8365] and does not support the procedures provided in this document. However, this document defines procedures to interoperate with RNVEs.

ToR switch: Top-of-Rack switch.

TS and VM: Tenant System and Virtual Machine. In this document, TSs and VMs are the devices connected to the ACs of the PEs and NVEs.

VNI: VXLAN Network Identifier. Used in VXLAN tunnels.

VSID: Virtual Segment Identifier. Used in NVGRE tunnels.

VXLAN: Virtual eXtensible Local Area Network [RFC7348].

3. Solution Requirements

The ingress replication optimization solution specified in this document meets the following requirements:

a. The solution provides an ingress replication optimization for BM traffic without the need for PIM while preserving the packet order for unicast applications, i.e., unknown unicast traffic should follow the same path as known unicast traffic. This optimization is required in low-performance NVEs.

b. The solution reduces the flooded traffic in NVO networks where some NVEs do not need broadcast/multicast and/or unknown unicast traffic.

c. The solution is compatible with [RFC7432] and [RFC8365] and has no impact on the Customer Edge (CE) procedures for BM traffic. In particular, the solution supports the following EVPN functions:

   * All-active multihoming, including the split-horizon and DF functions.

   * Single-active multihoming, including the DF function.

   * Handling of multi-destination traffic and processing of BM traffic as per [RFC7432].

d. The solution is backward compatible with existing NVEs using a non-optimized version of ingress replication. A given BD can have NVEs/PEs supporting regular ingress replication and optimized ingress replication.

e. The solution is independent of the NVO-specific data plane encapsulation and the virtual identifiers being used, e.g., VXLAN

VNIs, NVGRE VSIDs, or MPLS labels, as long as the tunnel is IP based.

4. EVPN BGP Attributes for Optimized Ingress Replication

The ingress replication optimization solution specified in this document extends the Inclusive Multicast Ethernet Tag routes and attributes described in [RFC7432] so that an NVE/PE can signal its optimized ingress replication capabilities.

The Network Layer Reachability Information (NLRI) of the Inclusive Multicast Ethernet Tag route [RFC7432] is shown in Figure 2 and is used in this document without any modifications to its format. The PMSI Tunnel Attribute's general format as provided in [RFC7432] (which takes it from [RFC6514]) is used in this document; only a new tunnel type and new flags are specified, as shown in Figure 3.

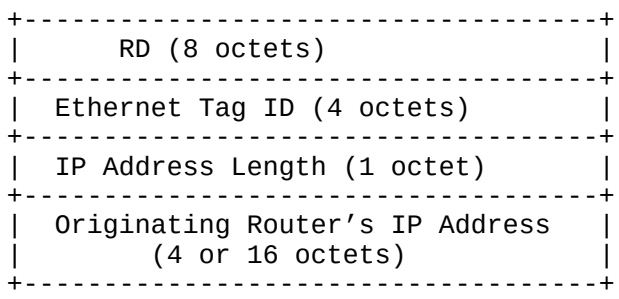


Figure 2: EVPN Inclusive Multicast Ethernet Tag Route's NLRI

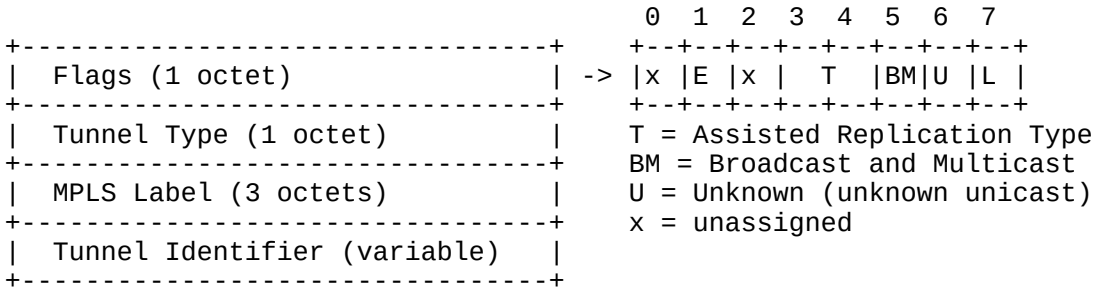


Figure 3: PMSI Tunnel Attribute

The Flags field in Figure 3 is 8 bits long as per [RFC7902]. The Extension (E) flag was allocated by [RFC7902], and the Leaf Information Required (L) flag was allocated by [RFC6514]. This document defines the use of 4 bits of this Flags field:

* Bits 3 and 4, which together form the Assisted Replication Type (T) field

* Bit 5, called the Broadcast and Multicast (BM) flag

* Bit 6, called the Unknown (U) flag

Bits 5 and 6 are collectively referred to as the Pruned Flooding Lists (PFLs) flags.

The T field and PFLs flags are defined as follows:

* T is the Assisted Replication Type field (2 bits), which defines the AR role of the advertising router:

  - 00 (decimal 0) = RNVE (non-AR support)

  - 01 (decimal 1) = AR-REPLICATOR

- 10 (decimal 2) = AR-LEAF

- 11 (decimal 3) = RESERVED

* The PFLs flags define the desired behavior of the advertising router for the different types of traffic:

  - Broadcast and Multicast (BM) flag. BM = 1 means "prune me from the BM flooding list". BM = 0 indicates regular behavior.

  - Unknown (U) flag. U = 1 means "prune me from the Unknown flooding list". U = 0 indicates regular behavior.

* The L flag (bit 7) is defined in [RFC6514] and will be used only in the selective AR solution.

Please refer to Section 11 for the IANA considerations related to the PMSI Tunnel Attribute flags.

In this document, the above Inclusive Multicast Ethernet Tag route (Figure 2) and PMSI Tunnel Attribute (Figure 3) can be used in two different modes for the same BD:

Regular-IR route: In this route, Originating Router's IP Address, Tunnel Type (0x06), MPLS Label, and Tunnel Identifier MUST be used as described in [RFC7432] when ingress replication is in use. The NVE/PE that advertises the route will set the Next Hop to an IP address that we denominate IR-IP in this document. When advertised by an AR-LEAF node, the Regular-IR route MUST be advertised with the T field set to 10 (AR-LEAF).

Replicator-AR route: This route is used by the AR-REPLICATOR to advertise its AR capabilities, with the fields set as follows:

  * Originating Router's IP Address MUST be set to an IP address of the advertising router that is common to all the EVIs on the PE (usually this is a loopback address of the PE).

    - The Tunnel Identifier and Next Hop fields SHOULD be set to the same IP address as the Originating Router's IP Address field when the NVE/PE originates the route -- that is, when the NVE/PE is not an ASBR; see Section 10.2 of [RFC8365]. Irrespective of the values in the Tunnel Identifier and Originating Router's IP Address fields, the ingress NVE/PE will process the received Replicator-AR route and will use the IP address setting in the Next Hop field to create IP tunnels to the AR-REPLICATOR.

    - The Next Hop address is referred to as the AR-IP and MUST be different from the IR-IP for a given PE/NVE, unless the procedures provided in Section 8 are followed.

  * Tunnel Type MUST be set to Assisted Replication Tunnel. Section 11 provides the allocated type value.

  * T (Assisted Replication type) MUST be set to 01 (AR-REPLICATOR).

  * L (Leaf Information Required) MUST be set to 0 for non-selective AR and MUST be set to 1 for selective AR.

An NVE/PE configured as an AR-REPLICATOR for a BD MUST advertise a Replicator-AR route for the BD and MAY advertise a Regular-IR route. The advertisement of the Replicator-AR route will indicate to the AR-LEAFs which outer IP DA, i.e., which AR-IP, they need to use for IP-

encapsulated BM frames that use Assisted Replication forwarding mode. The AR-REPLICATOR will forward an IP-encapsulated BM frame in Assisted Replication forwarding mode if the outer IP DA matches its AR-IP but will forward in Ingress Replication forwarding mode if the outer IP DA matches its IR-IP.

In addition, this document also uses the Leaf Auto-Discovery (Leaf A-D) route defined in [RFC9572] in cases where the selective AR mode is used. An AR-LEAF MAY send a Leaf A-D route in response to reception of a Replicator-AR route whose L flag is set. The Leaf A-D route is only used for selective AR, and the fields of such a route are set as follows:

* Originating Router's IP Address is set to the advertising router's IP address (the same IP address used by the AR-LEAF in Regular-IR routes). The Next Hop address is set to the IR-IP, which SHOULD be the same IP address as the advertising router's IP address, when the NVE/PE originates the route, i.e., when the NVE/PE is not an ASBR; see Section 10.2 of [RFC8365].

* Route Key [RFC9572] is the "Route Type Specific" NLRI of the Replicator-AR route for which this Leaf A-D route is generated.

* The AR-LEAF constructs an IP-address-specific Route Target, analogously to [RFC9572], by placing the IP address carried in the Next Hop field of the received Replicator-AR route in the Global Administrator field of the extended community, with the Local Administrator field of this extended community set to 0, and setting the Extended Communities attribute of the Leaf A-D route to that extended community. The same IP-address-specific import Route Target is auto-configured by the AR-REPLICATOR that sent the Replicator-AR route, in order to control the acceptance of the Leaf A-D routes.

* The Leaf A-D route MUST include the PMSI Tunnel Attribute with Tunnel Type set to Assisted Replication Tunnel (Section 11), T (Assisted Replication type) set to AR-LEAF, and Tunnel Identifier set to the IP address of the advertising AR-LEAF. The PMSI Tunnel Attribute MUST carry a downstream-assigned MPLS label or VNI that is used by the AR-REPLICATOR to send traffic to the AR-LEAF.

Each AR-enabled node understands and processes the T (Assisted Replication type) field in the PMSI Tunnel Attribute (Flags field) of the routes and MUST signal the corresponding type (AR-REPLICATOR or AR-LEAF type) according to its administrative choice. An NVE/PE following this specification is not expected to set the Assisted Replication Type field to decimal 3 (which is a RESERVED value). If a route with the Assisted Replication Type field set to decimal 3 is received by an AR-REPLICATOR or AR-LEAF, the router will process the route as a Regular-IR route advertised by an RNVE.

Each node attached to the BD may understand and process the BM/U flags (PFLs flags). Note that these BM/U flags may be used to optimize the delivery of multi-destination traffic; their use SHOULD be an administrative choice and independent of the AR role. When the PFL capability is enabled, the BM/U flags can be used with the Regular-IR, Replicator-AR, and Leaf A-D routes.

Non-optimized ingress replication NVEs/PEs will be unaware of the new PMSI Tunnel Attribute flag definition as well as the new tunnel type (AR), i.e., non-upgraded NVEs/PEs will ignore the information contained in the Flags field or an unknown tunnel type (type AR in this case) for any Inclusive Multicast Ethernet Tag route.

5. Non-selective Assisted Replication (AR) Solution Description

Figure 4 illustrates an example NVO network where the non-selective AR function is enabled. Three different roles are defined for a given BD: AR-REPLICATOR, AR-LEAF, and RNVE. The solution is called "non-selective" because the chosen AR-REPLICATOR for a given flow MUST replicate the BM traffic to all the NVEs/PEs in the BD except for the source NVE/PE. NVO tunnels, i.e., IP tunnels, exist among all the PEs and NVEs in the diagram. The PEs and NVEs in the diagram have TSs or VMs connected to their ACs.

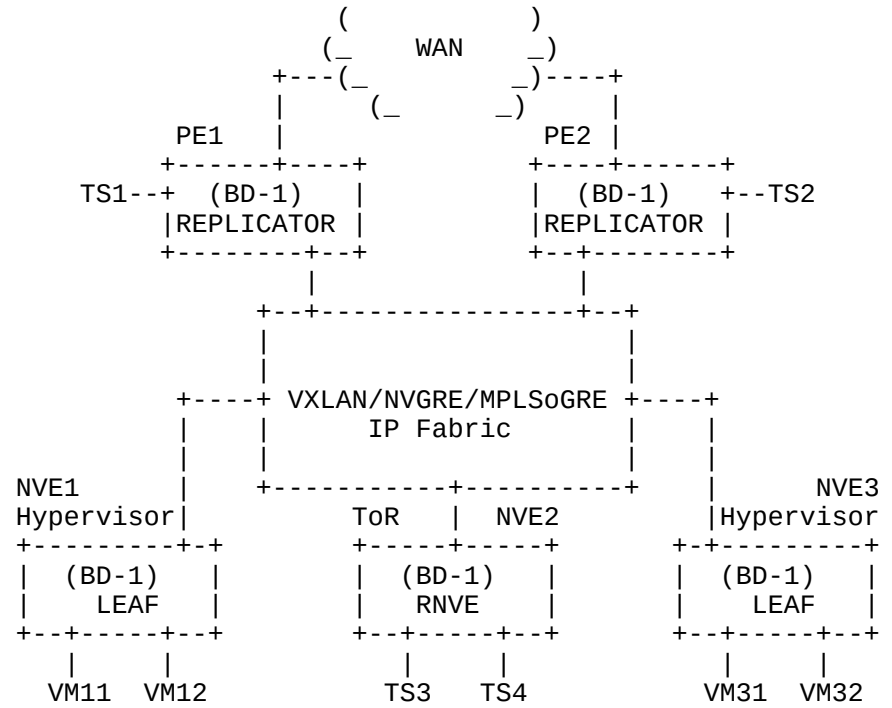


Figure 4: Non-selective AR Scenario

In AR BDs, such as BD-1 in Figure 4, BM traffic between two NVEs may follow a different path than unicast traffic. This solution recommends the replication of BM traffic through the AR-REPLICATOR node, whereas unknown/known unicast traffic will be delivered directly from the source node to the destination node without being replicated by any intermediate node.

Note that known unicast forwarding is not impacted by this solution, i.e., unknown unicast traffic SHALL follow the same path as known unicast traffic.

5.1. Non-selective AR-REPLICATOR Procedures

An AR-REPLICATOR is defined as an NVE/PE capable of replicating incoming BM traffic received on an overlay tunnel to other overlay tunnels and local ACs. The AR-REPLICATOR signals its role in the control plane and understands where the other roles (AR-LEAF nodes, RNVEs, and other AR-REPLICATORs) are located. A given AR-enabled BD service may have zero, one, or more AR-REPLICATORs. In our example in Figure 4, PE1 and PE2 are defined as AR-REPLICATORs. The following considerations apply to the AR-REPLICATOR role:

a. The AR-REPLICATOR role SHOULD be an administrative choice in any NVE/PE that is part of an AR-enabled BD. This administrative option to enable AR-REPLICATOR capabilities MAY be implemented as a system-level option as opposed to a per-BD option.

b. An AR-REPLICATOR MUST advertise a Replicator-AR route and MAY advertise a Regular-IR route. The AR-REPLICATOR MUST NOT generate a Regular-IR route if it does not have local ACs. If

the Regular-IR route is advertised, the Assisted Replication Type field of the Regular-IR route MUST be set to 0.

c. The Replicator-AR and Regular-IR routes are generated according to Section 4. The AR-IP and IR-IP are different IP addresses owned by the AR-REPLICATOR.

d. When a node defined as an AR-REPLICATOR receives a BM packet on an overlay tunnel, it will do a tunnel destination IP address lookup and apply the following procedures:

   * If the destination IP address is the AR-REPLICATOR IR-IP address, the node will process the packet normally as discussed in [RFC7432].

   * If the destination IP address is the AR-REPLICATOR AR-IP address, the node MUST replicate the packet to local ACs and overlay tunnels (excluding the overlay tunnel to the source of the packet). When replicating to remote AR-REPLICATORs, the tunnel destination IP address will be an IR-IP. This will indicate to the remote AR-REPLICATOR that it MUST NOT replicate to overlay tunnels. The tunnel source IP address used by the AR-REPLICATOR MUST be its IR-IP when replicating to AR-REPLICATOR or AR-LEAF nodes.

An AR-REPLICATOR MUST follow a data path implementation compatible with the following rules:

* The AR-REPLICATORs will build a flooding list composed of ACs and overlay tunnels to remote nodes in the BD. Some of those overlay tunnels MAY be flagged as non-BM receivers based on the BM flag received from the remote nodes in the BD.

* When an AR-REPLICATOR receives a BM packet on an AC, it will forward the BM packet to its flooding list (including local ACs and remote NVEs/PEs), skipping the non-BM overlay tunnels.

* When an AR-REPLICATOR receives a BM packet on an overlay tunnel, it will check the destination IP address of the underlay IP header and:

  - If the destination IP address matches its IR-IP, the AR-REPLICATOR will skip all the overlay tunnels from the flooding list, i.e., it will only replicate to local ACs. This is the regular ingress replication behavior described in [RFC7432].

  - If the destination IP address matches its AR-IP, the AR-REPLICATOR MUST forward the BM packet to its flooding list (ACs and overlay tunnels), excluding the non-BM overlay tunnels. The AR-REPLICATOR will ensure that the traffic is not sent back to the originating AR-LEAF.

  - If the encapsulation is MPLSoGRE or MPLSoUDP and the received BD label that the AR-REPLICATOR advertised in the Replicator-AR route is not at the bottom of the stack, the AR-REPLICATOR MUST copy all the labels below the BD label and propagate them when forwarding the packet to the egress overlay tunnels.

* The AR-REPLICATOR/LEAF nodes will build an unknown unicast flooding list composed of ACs and overlay tunnels to the IR-IP addresses of the remote nodes in the BD. Some of those overlay tunnels MAY be flagged as non-U (unknown unicast) receivers based on the U flag received from the remote nodes in the BD.

  - When an AR-REPLICATOR/LEAF receives an unknown unicast packet on an AC, it will forward the unknown unicast packet to its

flooding list, skipping the non-U overlay tunnels.

- When an AR-REPLICATOR/LEAF receives an unknown unicast packet on an overlay tunnel, it will forward the unknown unicast packet to its local ACs and never to an overlay tunnel. This is the regular ingress replication behavior described in [RFC7432].

5.2. Non-selective AR-LEAF Procedures

An AR-LEAF is defined as an NVE/PE that, given its poor replication performance, sends all the BM traffic to an AR-REPLICATOR that can replicate the traffic further on its behalf. It MAY signal its AR-LEAF capability in the control plane and understands where the other roles are located (AR-REPLICATORs and RNVEs). A given service can have zero, one, or more AR-LEAF nodes. In Figure 4, NVE1 and NVE3 (both residing in hypervisors) act as AR-LEAF nodes. The following considerations apply to the AR-LEAF role:

a. The AR-LEAF role SHOULD be an administrative choice in any NVE/PE that is part of an AR-enabled BD. This administrative option to enable AR-LEAF capabilities MAY be implemented as a system-level option as opposed to a per-BD option.

b. In this non-selective AR solution, the AR-LEAF MUST advertise a single Regular-IR Inclusive Multicast Ethernet Tag route as described in [RFC7432]. The AR-LEAF SHOULD set the Assisted Replication Type field to AR-LEAF. Note that although this field does not affect the remote nodes when creating an EVPN destination to the AR-LEAF, this field is useful from the standpoint of ease of operation and troubleshooting of the BD.

c. In a BD where there are no AR-REPLICATORs due to the AR-REPLICATORs being down or reconfigured, the AR-LEAF MUST use regular ingress replication based on the remote Regular-IR Inclusive Multicast Ethernet Tag routes as described in [RFC7432]. This may happen in the following cases:

   * The AR-LEAF has a list of AR-REPLICATORs for the BD, but it detects that all the AR-REPLICATORs for the BD are down (via next-hop tracking in the IGP or some other detection mechanism).

   * The AR-LEAF receives updates from all the former AR-REPLICATORs containing a non-REPLICATOR AR type in the Inclusive Multicast Ethernet Tag routes.

   * The AR-LEAF never discovered an AR-REPLICATOR for the BD.

d. In a service where there are one or more AR-REPLICATORs (based on the received Replicator-AR routes for the BD), the AR-LEAF can locally select which AR-REPLICATOR it sends the BM traffic to:

   * A single AR-REPLICATOR MAY be selected for all the BM packets received on the AR-LEAF ACs for a given BD. This selection is a local decision and does not have to match other AR-LEAFs' selections within the same BD.

   * An AR-LEAF MAY select more than one AR-REPLICATOR and do either per-flow or per-BD load balancing.

   * In the case of failure of the selected AR-REPLICATOR, another AR-REPLICATOR SHOULD be selected by the AR-LEAF.

   * When an AR-REPLICATOR is selected for a given flow or BD, the AR-LEAF MUST send all the BM packets targeted to that AR-

REPLICATOR using the forwarding information given by the Replicator-AR route for the chosen AR-REPLICATOR, with Tunnel Type = 0x0A (AR tunnel). The underlay destination IP address MUST be the AR-IP advertised by the AR-REPLICATOR in the Replicator-AR route.

* An AR-LEAF MAY change the selection of AR-REPLICATOR(s) dynamically due to an administrative or policy configuration change.

* AR-LEAF nodes SHALL send service-level BM control plane packets, following the procedures for regular ingress replication. An example would be IGMP, Multicast Listener Discovery (MLD), or PIM packets, and, in general, any packets using link-local scope multicast IPv4 or IPv6 packets. The AR-REPLICATORs MUST NOT replicate these control plane packets to other overlay tunnels, since they will use the IR-IP address.

e. The use of an AR-REPLICATOR-activation-timer (in seconds, with a default value of 3) on the AR-LEAF nodes is RECOMMENDED. Upon receiving a new Replicator-AR route where the AR-REPLICATOR is selected, the AR-LEAF will run a timer before programming the new AR-REPLICATOR. In the case of a newly added AR-REPLICATOR or if an AR-REPLICATOR reboots, this timer will give the AR-REPLICATOR some time to program the AR-LEAF nodes before the AR-LEAF sends BM traffic. The AR-REPLICATOR-activation-timer SHOULD be configurable in seconds, and its value needs to account for the time it takes for the AR-LEAF Regular-IR Inclusive Multicast Ethernet Tag route to get to the AR-REPLICATOR and be programmed. While the AR-REPLICATOR-activation-timer is running, the AR-LEAF node will use regular ingress replication.

f. If the AR-LEAF has selected an AR-REPLICATOR, whether or not to change to a new preferred AR-REPLICATOR for the existing BM traffic flows is a matter of local policy.

An AR-LEAF MUST follow a data path implementation compatible with the following rules:

* The AR-LEAF nodes will build two flooding lists:

  Flooding list #1: Composed of ACs and an AR-REPLICATOR-set of overlay tunnels. The AR-REPLICATOR-set is defined as one or more overlay tunnels to the AR-IP addresses of the remote AR-REPLICATOR(s) in the BD. The selection of more than one AR-REPLICATOR is described in item d. above and is a local AR-LEAF decision.

  Flooding list #2: Composed of ACs and overlay tunnels to the remote IR-IP addresses.

* When an AR-LEAF receives a BM packet on an AC, it will check the AR-REPLICATOR-set:

  - If the AR-REPLICATOR-set is empty, the AR-LEAF MUST send the packet to flooding list #2.

  - If the AR-REPLICATOR-set is NOT empty, the AR-LEAF MUST send the packet to flooding list #1, where only one of the overlay tunnels of the AR-REPLICATOR-set is used.

* When an AR-LEAF receives a BM packet on an overlay tunnel, it will forward the BM packet to its local ACs and never to an overlay tunnel. This is the regular ingress replication behavior described in [RFC7432].

* AR-LEAF nodes process unknown unicast traffic in the same way AR-REPLICATORS do, as described in Section 5.1.

5.3. RNVE Procedures

An RNVE is defined as an NVE/PE without AR-REPLICATOR or AR-LEAF capabilities that does ingress replication as described in [RFC7432]. The RNVE does not signal any AR role and is unaware of the AR-REPLICATOR/LEAF roles in the BD. The RNVE will ignore the flags in the Regular-IR routes and will ignore the Replicator-AR routes (due to an unknown tunnel type in the PMSI Tunnel Attribute) and the Leaf A-D routes (due to the IP-address-specific Route Target).

This role provides EVPNs with the backward compatibility required in optimized ingress replication BDs. In Figure 4, NVE2 acts as an RNVE.

6. Selective Assisted Replication (AR) Solution Description

Figure 5 is used to describe the selective AR solution.

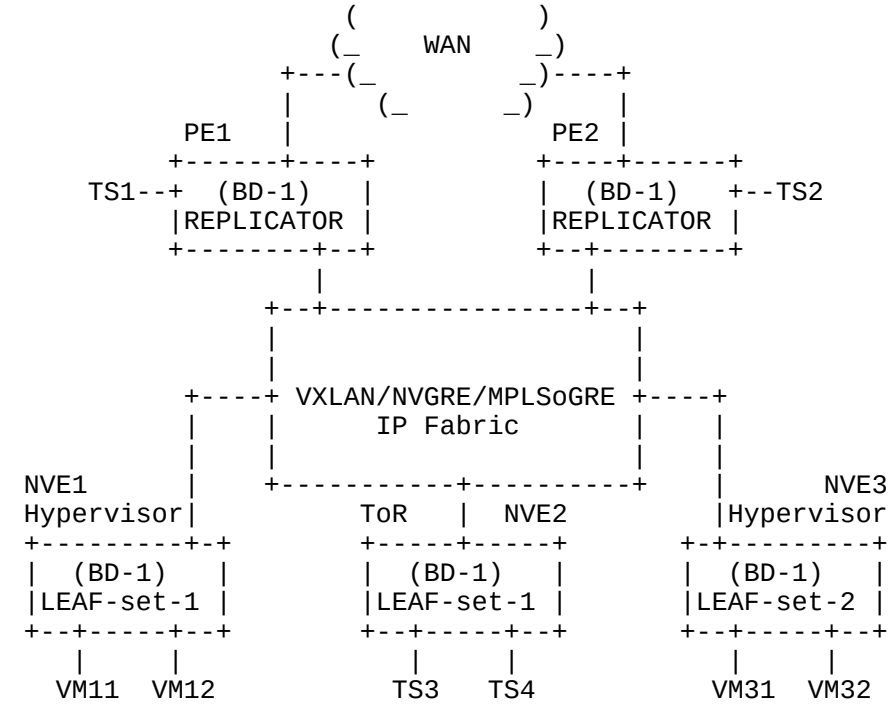


Figure 5: Selective AR Scenario

The solution is called "selective" because a given AR-REPLICATOR MUST replicate the BM traffic to only the AR-LEAFs that requested the replication (as opposed to all the AR-LEAF nodes) and MUST replicate the BM traffic to the RNVEs (if there are any). The same AR roles as those defined in Sections 4 and 5 are used here; however, the procedures are different.

The selective AR procedures create multiple AR-LEAF-sets in the EVPN BD and build single-hop trees among AR-LEAFs of the same set (AR-LEAF->AR-REPLICATOR->AR-LEAF) and two-hop trees among AR-LEAFs of different sets (AR-LEAF->AR-REPLICATOR->AR-REPLICATOR->AR-LEAF). Compared to the selective solution, the non-selective AR method assumes that all the AR-LEAFs of the BD are in the same set and always creates single-hop trees among AR-LEAFs. While the selective solution is more efficient than the non-selective solution in multi-stage IP fabrics, the trade-off is additional signaling and an additional outer source IP address lookup.

The following subsections describe the differences in the procedures for AR-REPLICATORs/LEAFs compared to the non-selective AR solution. There are no changes applicable to RNVEs.

6.1. Selective AR-REPLICATOR Procedures

In our example in Figure 5, PE1 and PE2 are defined as selective AR-REPLICATORs. The following considerations apply to the selective AR-REPLICATOR role:

a. The selective AR-REPLICATOR role SHOULD be an administrative choice in any NVE/PE that is part of an AR-enabled BD. This administrative option MAY be implemented as a system-level option as opposed to a per-BD option.

b. Each AR-REPLICATOR will build a list of AR-REPLICATOR, AR-LEAF, and RNVE nodes. In spite of the "selective" administrative option, an AR-REPLICATOR MUST NOT behave as a selective AR-REPLICATOR if at least one of the AR-REPLICATORs has the L flag NOT set. If at least one AR-REPLICATOR sends a Replicator-AR route with L = 0 (in the BD context), the rest of the AR-REPLICATORs will fall back to non-selective AR mode.

c. The selective AR-REPLICATOR MUST follow the procedures described in Section 5.1, except for the following differences:

   * The AR-REPLICATOR MUST have the L flag set to 1 when advertising the Replicator-AR route. This flag is used by the AR-REPLICATORs to advertise their "selective" AR-REPLICATOR capabilities. In addition, the AR-REPLICATOR auto-configures its IP-address-specific import Route Target as described in the third bullet of the procedures for Leaf A-D routes in Section 4.

   * The AR-REPLICATOR will build a "selective" AR-LEAF-set with the list of nodes that requested replication to its own AR-IP. For instance, assuming that NVE1 and NVE2 advertise a Leaf A-D route with PE1’s IP-address-specific Route Target and NVE3 advertises a Leaf A-D route with PE2’s IP-address-specific Route Target, PE1 will only add NVE1/NVE2 to its selective AR-LEAF-set for BD-1 and exclude NVE3. Likewise, PE2 will only add NVE3 to its selective AR-LEAF-set for BD-1 and exclude NVE1/NVE2.

   * When a node defined and operating as a selective AR-REPLICATOR receives a packet on an overlay tunnel, it will do a tunnel destination IP lookup, and if the destination IP address is the AR-REPLICATOR AR-IP address, the node MUST replicate the packet to:

     - Local ACs.

     - Overlay tunnels in the selective AR-LEAF-set, excluding the overlay tunnel to the source AR-LEAF.

     - Overlay tunnels to the RNVEs if the tunnel source IP address is the IR-IP of an AR-LEAF. In any other case, the AR-REPLICATOR MUST NOT replicate the BM traffic to remote RNVEs. In other words, only the first-hop selective AR-REPLICATOR will replicate to all the RNVEs.

     - Overlay tunnels to the remote selective AR-REPLICATORs if the tunnel source IP address (of the encapsulated packet that arrived on the overlay tunnel) is an IR-IP of its own AR-LEAF-set. In any other case, the AR-REPLICATOR MUST NOT

replicate the BM traffic to remote AR-REPLICATORs. When doing this replication, the tunnel destination IP address is the AR-IP of the remote selective AR-REPLICATOR. The tunnel destination address AR-IP will indicate to the remote selective AR-REPLICATOR that the packet needs further replication to its AR-LEAFs.

A selective AR-REPLICATOR data path implementation MUST be compatible with the following rules:

* The selective AR-REPLICATORs will build two flooding lists:

  Flooding list #1: Composed of ACs and overlay tunnels to the remote nodes in the BD, always using the IR-IPs in the tunnel destination IP addresses.

  Flooding list #2: Composed of ACs, a selective AR-LEAF-set, and a selective AR-REPLICATOR-set, where:

  - The selective AR-LEAF-set is composed of the overlay tunnels to the AR-LEAFs that advertise a Leaf A-D route for the local AR-REPLICATOR. This set is updated with every Leaf A-D route received/withdrawn from a new AR-LEAF.

  - The selective AR-REPLICATOR-set is composed of the overlay tunnels to all the AR-REPLICATORs that send a Replicator-AR route with L = 1. The AR-IP addresses are used as tunnel destination IP addresses.

* Some of the overlay tunnels in the flooding lists MAY be flagged as non-BM receivers based on the BM flag received from the remote nodes in the routes.

* When a selective AR-REPLICATOR receives a BM packet on an AC, it MUST forward the BM packet to its flooding list #1, skipping the non-BM overlay tunnels.

* When a selective AR-REPLICATOR receives a BM packet on an overlay tunnel, it will check the destination and source IPs of the underlay IP header and:

  - If the destination IP address matches its AR-IP and the source IP address matches an IP of its own selective AR-LEAF-set, the AR-REPLICATOR MUST forward the BM packet to its flooding list #2, unless some AR-REPLICATOR within the BD has advertised L = 0. In the latter case, the node reverts to Non-selective mode, and flooding list #1 MUST be used. Non-BM overlay tunnels are skipped when sending BM packets.

  - If the destination IP address matches its AR-IP and the source IP address does not match any IP address of its selective AR-LEAF-set, the AR-REPLICATOR MUST forward the BM packet to flooding list #2, skipping the AR-REPLICATOR-set. Non-BM overlay tunnels are skipped when sending BM packets.

  - If the destination IP address matches its IR-IP, the AR-REPLICATOR MUST use flooding list #1 but MUST skip all the overlay tunnels from the flooding list, i.e., it will only replicate to local ACs. This is the regular ingress replication behavior described in [RFC7432]. Non-BM overlay tunnels are skipped when sending BM packets.

* In any case, the AR-REPLICATOR ensures that the traffic is not sent back to the originating source. If the encapsulation is MPLSoGRE or MPLSoUDP and the received BD label (the label that the AR-REPLICATOR advertised in the Replicator-AR route) is not at the

bottom of the stack, the AR-REPLICATOR MUST copy the rest of the labels when forwarding them to the egress overlay tunnels.

6.2. Selective AR-LEAF Procedures

A selective AR-LEAF chooses a single selective AR-REPLICATOR per BD and:

* Sends all the BD's BM traffic to that AR-REPLICATOR and

* Expects to receive all the BM traffic for a given BD from the same AR-REPLICATOR (except for the BM traffic from the RNVEs, which comes directly from the RNVEs)

In the example in Figure 5, we consider NVE1/NVE2/NVE3 as selective AR-LEAFs. NVE1 selects PE1 as its selective AR-REPLICATOR. If that is so, NVE1 will send all its BM traffic for BD-1 to PE1. If other AR-LEAFs/REPLICATORs send BM traffic, NVE1 will receive that traffic from PE1. A selective AR-LEAF and a non-selective AR-LEAF behave differently, as follows:

a. The selective AR-LEAF role SHOULD be an administrative choice in any NVE/PE that is part of an AR-enabled BD. This administrative option to enable AR-LEAF capabilities MAY be implemented as a system-level option as opposed to a per-BD option.

b. The AR-LEAF MAY advertise a Regular-IR route if there are RNVEs in the BD. The selective AR-LEAF MUST advertise a Leaf A-D route after receiving a Replicator-AR route with L = 1. It is RECOMMENDED that the selective AR-LEAF wait for a period specified by an AR-LEAF-join-wait-timer (in seconds, with a default value of 3) before sending the Leaf A-D route, so that the AR-LEAF can collect all the Replicator-AR routes for the BD before advertising the Leaf A-D route. If the Replicator-AR route with L = 1 is withdrawn, the corresponding Leaf A-D route is withdrawn too.

c. In a service where there is more than one selective AR-REPLICATOR, the selective AR-LEAF MUST locally select a single selective AR-REPLICATOR for the BD. Once selected:

    * The selective AR-LEAF MUST send a Leaf A-D route, including the route key and IP-address-specific Route Target of the selected AR-REPLICATOR.

    * The selective AR-LEAF MUST send all the BM packets received on the ACs for a given BD to that AR-REPLICATOR.

    * In the case of failure of the selected AR-REPLICATOR (detected when the Replicator-AR route becomes infeasible as a result of any of the underlying BGP mechanisms), another AR-REPLICATOR will be selected and a new Leaf A-D update will be issued for the new AR-REPLICATOR. This new route will update the selective list in the new selective AR-REPLICATOR. In the case of failure of the active selective AR-REPLICATOR, it is RECOMMENDED that the selective AR-LEAF revert to ingress replication behavior for an AR-REPLICATOR-activation-timer (in seconds, with a default value of 3) to mitigate the traffic impact. When the timer expires, the selective AR-LEAF will resume its AR mode with the new selective AR-REPLICATOR. The AR-REPLICATOR-activation-timer MAY be the same configurable parameter as the parameter discussed in Section 5.2.

    * A selective AR-LEAF MAY change the selection of AR-REPLICATOR(s) dynamically due to an administrative or policy configuration change.

All the AR-LEAFs in a BD are expected to be configured as either selective or non-selective. A mix of selective and non-selective AR-LEAFs SHOULD NOT coexist in the same BD. If a non-selective AR-LEAF is present, its BM traffic sent to a selective AR-REPLICATOR will not be replicated to other AR-LEAFs that are not in its selective AR-LEAF-set.

A selective AR-LEAF MUST follow a data path implementation compatible with the following rules:

* The selective AR-LEAF nodes will build two flooding lists:

  Flooding list #1: Composed of ACs and the overlay tunnel to the selected AR-REPLICATOR (using the AR-IP as the tunnel destination IP address).

  Flooding list #2: Composed of ACs and overlay tunnels to the remote IR-IP addresses.

* Some of the overlay tunnels in the flooding lists MAY be flagged as non-BM receivers based on the BM flag received from the remote nodes in the routes.

* When an AR-LEAF receives a BM packet on an AC, it will check to see if an AR-REPLICATOR was selected; if one is found, flooding list #1 MUST be used. Otherwise, flooding list #2 MUST be used. Non-BM overlay tunnels are skipped when sending BM packets.

* When an AR-LEAF receives a BM packet on an overlay tunnel, it MUST forward the BM packet to its local ACs and never to an overlay tunnel. This is the regular ingress replication behavior described in [RFC7432].

7. Pruned Flooding Lists (PFLs)

In addition to AR, the second optimization supported by the ingress replication optimization solution specified in this document is the ability of all the BD nodes to signal PFLs. As described in Section 4, an EVPN node can signal a given value for the BM and U PFLs flags in the Regular-IR, Replicator-AR, or Leaf A-D routes, where:

* BM is the Broadcast and Multicast flag. BM = 1 means "prune me from the BM flooding list". BM = 0 indicates regular behavior.

* U is the Unknown flag. U = 1 means "prune me from the Unknown flooding list". U = 0 indicates regular behavior.

The ability to signal and process these PFLs flags SHOULD be an administrative choice. If a node is configured to process the PFLs flags, upon receiving a non-zero PFLs flag for a route, an NVE/PE will add the corresponding flag to the created overlay tunnel in the flooding list. When replicating a BM packet in the context of a flooding list, the NVE/PE will skip the overlay tunnels marked with the flag BM = 1, since the NVEs/PEs at the end of those tunnels are not expecting BM packets. Similarly, when replicating unknown unicast packets, the NVE/PE will skip the overlay tunnels marked with U = 1.

An NVE/PE not following this document or not configured for this optimization will ignore any of the received PFLs flags. An AR-LEAF or RNVE receiving BUM traffic on an overlay tunnel MUST replicate the traffic to its local ACs, regardless of the BM/U flags on the overlay tunnels.

This optimization MAY be used along with the Assisted Replication solution.

7.1. Example of a Pruned Flooding List

In order to illustrate the use of the PFLs solution, we will assume that BD-1 in Figure 4 is optimized ingress replication enabled and:

* PE1 and PE2 are administratively configured as AR-REPLICATORs due to their high-performance replication capabilities. PE1 and PE2 will send a Replicator-AR route with BM/U flags = 00.

* NVE1 and NVE3 are administratively configured as AR-LEAF nodes due to their low-performance software-based replication capabilities. They will advertise a Regular-IR route with type AR-LEAF. Assuming that both NVEs advertise all of the attached VMs' MAC and IP addresses in EVPNs as soon as they come up and these NVEs do not have any VMs interested in multicast applications, they will be configured to signal BM/U flags = 11 for BD-1. That is, neither NVE1 nor NVE3 is interested in receiving BM or unknown unicast traffic, since:

  - Their attached VMs (VM11, VM12, VM31, VM32) do not support multicast applications.

  - Their attached VMs will not receive ARP Requests. Proxy ARP [RFC9161] on the remote NVEs/PEs will reply to ARP Requests locally, and no other broadcast traffic is expected.

  - Their attached VMs will not receive unknown unicast traffic, since the VMs' MAC and IP addresses are always advertised by EVPNs as long as the VMs are active.

* NVE2 is optimized ingress replication unaware; therefore, it takes on the RNVE role in BD-1.

Based on the above assumptions, the following forwarding behavior will take place:

1. Any BM packets sent from VM11 will be sent to VM12 and PE1. PE1 will then forward the BM packets on to TS1, the WAN link, PE2, and NVE2 but not to NVE3. PE2 and NVE2 will replicate the BM packets to their local ACs, but NVE3 will be prevented from having to replicate those BM packets to VM31 and VM32 unnecessarily.

2. Any BM packets received on PE2 from the WAN will be sent to PE1 and NVE2 but not to NVE1 and NVE3, sparing the two hypervisors from replicating unnecessarily to their local VMs. PE1 and NVE2 will replicate to their local ACs only.

3. Any unknown unicast packet sent from VM31 will be forwarded by NVE3 to NVE2, PE1, and PE2 but not to NVE1. The solution prevents unnecessary replication to NVE1, since the destination of the unknown traffic cannot be NVE1.

4. Any unknown unicast packet sent from TS1 will be forwarded by PE1 to the WAN link, PE2, and NVE2 but not to NVE1 and NVE3, since the target of the unknown traffic cannot be NVE1 or NVE3.

8. AR Procedures for Single-IP AR-REPLICATORS

The procedures explained in Sections 5 and 6 assume that the AR-REPLICATOR can use two local routable IP addresses to terminate and originate NVO tunnels, i.e., IR-IP and AR-IP addresses. This is usually the case for PE-based AR-REPLICATOR nodes.

In some cases, the AR-REPLICATOR node does not support more than one IP address to terminate and originate NVO tunnels, i.e., the IR-IP and AR-IP are the same IP addresses. This may be the case in some software-based or low-end AR-REPLICATOR nodes. If this is the case, the procedures provided in Sections 5 and 6 MUST be modified in the following way:

* The Replicator-AR routes generated by the AR-REPLICATOR use an AR-IP that will match its IR-IP. In order to differentiate the data plane packets that need to use ingress replication from the packets that must use Assisted Replication forwarding mode, the Replicator-AR route MUST advertise a different VNI/VSID than the one used by the Regular-IR route. For instance, the AR-REPLICATOR will advertise an AR-VNI along with the Replicator-AR route and an IR-VNI along with the Regular-IR route. Since both routes have the same key, different Route Distinguishers are needed in each route.

* An AR-REPLICATOR will perform Ingress Replication forwarding mode or Assisted Replication forwarding mode for the incoming overlay packets based on an ingress VNI lookup as opposed to the tunnel IP DA lookup. Note that when replicating to remote AR-REPLICATOR nodes, the use of the IR-VNI or AR-VNI advertised by the egress node will determine whether Ingress Replication forwarding mode or Assisted Replication forwarding mode is used at the subsequent AR-REPLICATOR.

The rest of the procedures will follow those described in Sections 5 and 6.

9. AR Procedures and EVPN All-Active Multihoming Split-Horizon

This section extends the procedures for the cases where two or more AR-LEAF nodes are attached to the same ES and two or more AR-REPLICATOR nodes are attached to the same ES in the BD. The mixed case -- where an AR-LEAF node and an AR-REPLICATOR node are attached to the same ES -- would require extended procedures that are out of scope for this document.

9.1. Ethernet Segments on AR-LEAF Nodes

If a VXLAN or NVGRE is used and if the split-horizon is based on the tunnel source IP address and "local bias" as described in [RFC8365], the split-horizon check will not work if an ES is shared between two AR-LEAF nodes, and the AR-REPLICATOR replaces the tunnel source IP address of the packets with its own AR-IP.

In order to be compatible with the source IP address split-horizon check, the AR-REPLICATOR MAY keep the original received tunnel source IP address when replicating packets to a remote AR-LEAF or RNVE. This will allow AR-LEAF nodes to apply split-horizon check procedures for BM packets before sending them to the local ES. Even if the AR-LEAF's source IP address is preserved when replicating to AR-LEAFs or RNVEs, the AR-REPLICATOR MUST always use its IR-IP as the source IP address when replicating to other AR-REPLICATORs.

When EVPNs are used for MPLSoGRE or MPLSoUDP, the ESI-label-based split-horizon procedure provided in [RFC7432] will not work for multihomed ESs defined on AR-LEAF nodes. Local bias is recommended in this case, as it is in the case of a VXLAN or NVGRE as explained above. The local-bias and tunnel source IP address preservation mechanisms provide the required split-horizon behavior in non-selective or selective AR.

Note that if the AR-REPLICATOR implementation keeps the received

tunnel source IP address, the use of unicast Reverse Path Forwarding (uRPF) checks in the IP fabric based on the tunnel source IP address MUST be disabled.

9.2. Ethernet Segments on AR-REPLICATOR Nodes

AR-REPLICATOR nodes attached to the same all-active ES will follow local-bias procedures [RFC8365] as follows:

a. For BUM traffic received on a local AR-REPLICATOR's AC, local-bias procedures as provided in [RFC8365] MUST be followed.

b. For BUM traffic received on an AR-REPLICATOR overlay tunnel with AR-IP as the IP DA, local bias MUST also be followed. That is, traffic received with AR-IP as the IP DA will be treated as though it had been received on a local AC that is part of the ES and will be forwarded to all local ESs, irrespective of their DF or NDF state.

c. BUM traffic received on an AR-REPLICATOR overlay tunnel with IR-IP as the IP DA will follow regular local-bias rules [RFC8365] and will not be forwarded to local ESs that are shared with the AR-LEAF or AR-REPLICATOR originating the traffic.

d. In cases where the AR-REPLICATOR supports a single IP address, the IR-IP and the AR-IP are the same IP address, as discussed in Section 8. The received BUM traffic will be treated as specified in item b above if the received VNI is the AR-VNI and as specified in item c if the VNI is the IR-VNI.

10. Security Considerations

The security considerations in [RFC7432] and [RFC8365] apply to this document. The security considerations related to the Leaf A-D route in [RFC9572] apply too.

In addition, the Assisted Replication method introduced by this document may introduce some new risks that could affect the successful delivery of BM traffic. Unicast traffic is not affected by Assisted Replication (although unknown unicast traffic is affected by the procedures for PFLs). The forwarding of BM traffic is modified, and BM traffic from the AR-LEAF nodes will be drawn toward AR-REPLICATORs in the BD. An AR-LEAF will forward BM traffic to its selected AR-REPLICATOR; therefore, an attack on the AR-REPLICATOR could impact the delivery of the BM traffic using that node. Also, an attack on the AR-REPLICATOR and any change to the advertised AR type will modify the selections made by the AR-LEAF nodes. If no other AR-REPLICATOR is selected, the AR-LEAF nodes will be forced to use Ingress Replication forwarding mode, which will impact their performance, since the AR-LEAF nodes are usually NVEs/PEs with poor replication performance.

This document introduces the ability of the AR-REPLICATOR to forward traffic received on an overlay tunnel to another overlay tunnel. The reader may determine that this introduces the risk of BM loops -- that is, an AR-LEAF receiving a BM-encapsulated packet that the AR-LEAF originated in the first place due to one or two AR-REPLICATORs "looping" the BM traffic back to the AR-LEAF. Following the procedures provided in this document will prevent these BM loops, since the AR-REPLICATOR will always forward the BM traffic using the correct tunnel IP DA (or the correct VNI in the case of single-IP AR-REPLICATORs), which instructs the remote nodes regarding how to forward the traffic. This is true for both the Non-selective and Selective modes defined in this document. However, incorrect implementation of the procedures provided in this document may lead to those unexpected BM loops.

The Selective mode provides a multi-stage replication solution, where proper configuration of all the AR-REPLICATORs will prevent any issues. A mix of mistakenly configured selective and non-selective AR-REPLICATORs in the same BD could theoretically create packet duplication in some AR-LEAFs; however, this document specifies a fallback solution -- falling back to Non-selective mode in cases where the AR-REPLICATORs advertised an inconsistent AR mode.

This document allows the AR-REPLICATOR to preserve the tunnel source IP address of the AR-LEAF (as an option) when forwarding BM packets from an overlay tunnel to another overlay tunnel. Preserving the AR-LEAF source IP address makes the local-bias filtering procedures possible for AR-LEAF nodes that are attached to the same ES. If the AR-REPLICATOR does not preserve the AR-LEAF source IP address, AR-LEAF nodes attached to all-active ESs will cause packet duplication on the multihomed CE.

The AR-REPLICATOR nodes are, by design, using more bandwidth than PEs [RFC7432] or NVEs [RFC8365] would use. Certain network events or unexpected low performance may exceed the AR-REPLICATOR’s local bandwidth and cause service disruption.

Finally, PFLs (Section 7) should be used with care. Intentional or unintentional misconfiguration of the BDs on a given leaf node may result in the leaf not receiving the required BM or unknown unicast traffic.

11. IANA Considerations

IANA has allocated the following Border Gateway Protocol (BGP) parameters:

* Allocation in the "P-Multicast Service Interface Tunnel (PMSI Tunnel) Tunnel Types" registry:

Value	Meaning	Reference
0x0A	Assisted Replication Tunnel	RFC 9574

Table 1

* Allocations in the "P-Multicast Service Interface (PMSI) Tunnel Attribute Flags" registry:

Value	Name	Reference
3-4	Assisted Replication Type (T)	RFC 9574
5	Broadcast and Multicast (BM)	RFC 9574
6	Unknown (U)	RFC 9574

Table 2

12. References

12.1. 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>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/ BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[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>.

[RFC7902] Rosen, E. and T. Morin, "Registry and Extensions for P-Multicast Service Interface Tunnel Attribute Flags", RFC 7902, DOI 10.17487/RFC7902, June 2016, <https://www.rfc-editor.org/info/rfc7902>.

[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>.

[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>.

[RFC9572] Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates to EVPN Broadcast, Unknown Unicast, or Multicast (BUM) Procedures", RFC 9572, DOI 10.17487/RFC9572, May 2024, <https://www.rfc-editor.org/info/rfc9572>.

12.2. Informative References

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, Ed., "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, DOI 10.17487/RFC4023, March 2005, <https://www.rfc-editor.org/info/rfc4023>.

[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>.

[RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", RFC 7637, DOI 10.17487/RFC7637, September 2015, <https://www.rfc-editor.org/info/rfc7637>.

[RFC9161] Rabadan, J., Ed., Sathappan, S., Nagaraj, K., Hankins, G., and T. King, "Operational Aspects of Proxy ARP/ND in Ethernet Virtual Private Networks", RFC 9161, DOI 10.17487/RFC9161, January 2022, <https://www.rfc-editor.org/info/rfc9161>.

Acknowledgements

The authors would like to thank Neil Hart, David Motz, Dai Truong, Thomas Morin, Jeffrey Zhang, Shankar Murthy, and Krzysztof Szarkowicz

for their valuable feedback and contributions. Also, thanks to John Scudder for his thorough review, which improved the quality of the document significantly.

Contributors

In addition to the authors listed on the front page, the following people also contributed to this document and should be considered coauthors:

Wim Henderickx
Nokia

Kiran Nagaraj
Nokia

Ravi Shekhar
Juniper Networks

Nischal Sheth
Juniper Networks

Aldrin Isaac
Juniper

Mudassir Tufail
Citibank

Authors' Addresses

Jorge Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
United States of America
Email: jorge.rabadan@nokia.com

Senthil Sathappan
Nokia
Email: senthil.sathappan@nokia.com

Wen Lin
Juniper Networks
Email: wlin@juniper.net

Mukul Katiyar
Versa Networks
Email: mukul@versa-networks.com

Ali Sajassi
Cisco Systems
Email: sajassi@cisco.com