BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: July 29, 2022

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

January 25, 2022

Optimized Ingress Replication Solution for Ethernet VPN (EVPN) draft-ietf-bess-evpn-optimized-ir-12

Abstract

Network Virtualization Overlay networks using Ethernet VPN (EVPN) as their control plane may use Ingress Replication or PIM (Protocol Independent Multicast)-based trees to convey the overlay Broadcast, Unknown unicast and Multicast (BUM) traffic. PIM provides an efficient solution to avoid sending multiple copies of the same packet over the same physical link, however it may not always be deployed in the Network Virtualization Overlay core network. Ingress Replication avoids the dependency on PIM in the Network Virtualization Overlay network core. While Ingress Replication provides a simple multicast transport, some Network Virtualization Overlay 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.

Table of Contents

1. Introduction

Ethernet Virtual Private Networks (EVPN) may be used as the control plane for a Network Virtualization Overlay network [RFC8365]. Network Virtualization Edge (NVE) and Provider Edge (PE) devices that

are part of the same EVPN Broadcast Domain (BD) use Ingress Replication or PIM-based trees to transport the tenant's Broadcast, Unknown unicast and Multicast (BUM) traffic.

In the Ingress Replication approach, the ingress NVE receving a BUM frame from the Tenant System will create as many copies of the frame as remote NVEs/PEs 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, 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, NV3, PE1. On the right-hand side of the diagram, 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 muticast state, e.g., the Spine keeps a PIM multicast routing entry for (S1,G1) with an Incoming Interface (IIF) and three Outgoing Interfaces (OIFs).

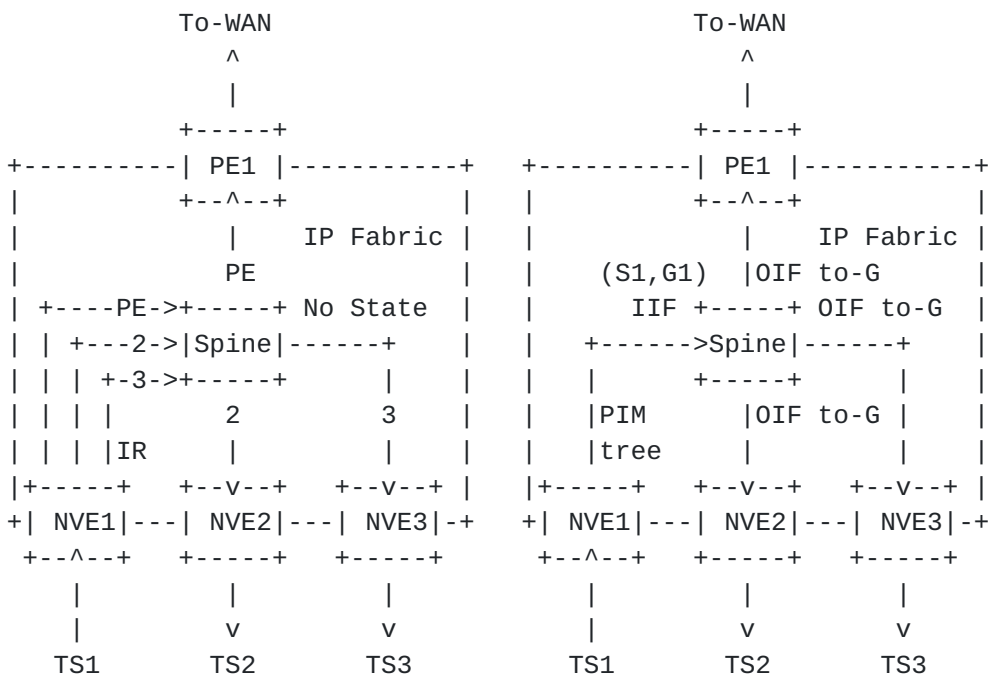

Figure 1: Ingress Replication vs PIM-based trees in NVO networks

In Network Virtualization Overlay networks where PIM-based trees cannot be used, Ingress Replication is the only option. Examples of these situations are Network Virtualization Overlay 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 is greatly reduced due to the proxy ARP (Address Resolution Protocol) and proxy ND (Neighbor Discovery) capabilities supported by EVPN on the NVEs [I-D.ietf-bess-evpn-proxy-arp-nd]. Some NVEs can even provide Dynamic Host Configuration Protocol (DHCP) server functions for the attached Tenant Systems, reducing the broadcast even further.

b. Unknown unicast traffic is greatly reduced in Network Virtualization Overlay networks where all the MAC and IP addresses from the Tenant Systems are learned in the control plane.

c. Multicast applications are not used.

If the above assumptions are true for a given Network Virtualization Overlay network, then Ingress Replication provides a simple solution

for multi-destination traffic. However, the 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 TORs (Top Of Rack 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-Flood-Lists (PFL)

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 Tenant System to the Broadcast Domain, 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 Tenant System. The AR-LEAF sends a single copy of a Broadcast or Multicast packet to a selected AR-REPLICATOR that replicates the packet mutiple times to remote AR-LEAF or AR-REPLICATOR routers, and therefore "assisting" the ingress AR-LEAF in delivering the Broadcast or Multicast traffic to the remote NVEs/PEs attached to the same Broadcast Domain. 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 NVE/PEs. The procedures that use a single AR-REPLICATOR (Non-Selective Assisted-Replication Solution) are specified in Section 5, whereas Section 6 describes how multi-staged replication, i.e., two AR-REPLICATOR routers in the path between the ingress AR-LEAF and destination NVEs/PEs, is accomplished (Selective Assisted-Replication Solution). The Assisted-Replication procedures do not impact unknown unicast traffic, which follows the same forwarding procedures as known unicast traffic so that packet re-ordering does not occur.

Pruned-Flood-Lists is a method for the ingress NVE/PE to prune or remove certain destination NVEs/PEs from a flood-list, depending on the interest of those NVEs/PEs in receiving Broadcast, Multicast or Unknown unicast. As specified in [RFC8365], an NVE/PE builds a flood-list for BUM traffic based on the Next-Hops of the received EVPN Inclusive Multicast Ethernet Tag routes for the Broadcast Domain. While [RFC8365] states that the flood-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 flood-list

with Next-Hops PE1, PE2 and PE3. If PE2 and PE3 signaled no-interest in receiving Unknown Unicast in their Inclusive Multicast Ethernet Tag routes, when the ingress NVE receives an Unknown Unicast frame from a Tenant System it will replicate it only to PE1. That is, PE2 and PE3 are "pruned" from the NVE's flood-list for Unknown Unicast traffic. Pruned-Flood-Lists can be used with Ingress Replication or Assisted-Replication, and it is described in Section 7.

Both optimizations, Assisted-Replication and Pruned-Flood-Lists, 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], and they are described in Section 4.

The Assisted-Replication solution described in this document is focused on Network Virtualization Overlay networks (hence it uses IP tunnels) and MPLS transport networks are out of scope. The Pruned-Flood-Lists solution MAY be used in Network Virtualization Overlay and MPLS transport networks.

Section 3 lists the requirements of the combined optimized Ingress Replication solution, whereas Section 5 and Section 6 describe the Assisted-Replication solution (for Non-Selective and Selective procedures, respectively), and Section 7 the Pruned-Flood-Lists 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 the document:

- Asisted Replication forwarding mode: for an AR-LEAF, it means sending an Attachment Circuit BM packet to a single AR-REPLICATOR with tunnel destination IP AR-IP. For an AR-REPLICATOR, it means sending a BM packet to a selected number or all the overlay tunnels when the packet was previously received from an overlay tunnel.

- AR-LEAF: Assisted Replication - LEAF, refers to an NVE/PE that sends all the Broadcast and Multicast 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. This document defines the control and data plane procedures that an AR-REPLICATOR needs to follow.

- AR-IP: 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-VNI: VNI advertised by the AR-REPLICATOR along with the Replicator-AR route. It is used to identify the incoming packets that must follow AR procedures ONLY in the Single-IP AR-REPLICATOR case Section 8.

- BM traffic: Refers to Broadcast and Multicast frames (excluding unknown unicast frames).

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

- BD label: defined as the MPLS label that identifies the Broadcast Domain and is advertised in Regular-IR or Replicator-AR routes, when the encapsulation is MPLSoGRE or MPLSoUDP.

- DF and NDF: Designated Forwarder and Non-Designated Forwarder, are roles defined in NVE/PEs attached to Multi-Homed Tenant Systems, as per [RFC7432] and [RFC8365].

- ES and ESI: Ethernet Segment and Ethernet Segment Identifier, as EVPN Multi-Homing concepts 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: it refers to the Ingress Replication behavior explained in [RFC7432]. It means sending an Attachment Circuit BM packet copy to each remote PE/NVE in the BD and sending an overlay BM packet only to the Attachment Circuits and not other overlay tunnels.

- IR-IP: local IP address of an NVE/PE that is used for the Ingress Replication signaling and procedures in [RFC7432]. Encapsulated incoming traffic with outer destination IP matching the IR-IP will follow the Ingress Replication procedures and not the Assisted-

Replication procedures. The IR-IP is also used in the Tunnel Identifier and Next-hop fields of the Regular-IR route.

- IR-VNI: VNI advertised along with the Inclusive Multicast Ethernet Tag route for Ingress Replication Tunnel Type.

- MPLS: Multi-Protocol Label Switching.

- NVE: Network Virtualization Edge router, used in this document as in [RFC8365].

- NVGRE: Network Virtualization using Generic Routing Encapsulation, as in [RFC7637].

- PE: Provider Edge router.

- 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 Ingress Replication Tunnel Type.

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

- 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.

- TOR: Top Of Rack switch.

- TS and VM: Tenant System and Virtual Machine. In this document Tenant Systems and Virtual Machiness are the devices connected to the Attachment Circuits of the PEs and NVEs.

- VNI: VXLAN Network Identifier, used in VXLAN tunnels.

- VSID: Virtual Segment Identifier, used in NVGRE tunnels.

- VXLAN: Virtual Extensible LAN [RFC7348].

3. Solution Requirements

The Ingress Replication optimization solution specified in this document meets the following requirements:

a. It provides an Ingress Replication optimization for Broadcast and Multicast 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. It reduces the flooded traffic in Network Virtualization Overlay 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 CE procedures for BM traffic. In particular, the solution supports the following EVPN functions:

   - All-active multi-homing, including the split-horizon and Designated Forwarder (DF) functions.

   - Single-active multi-homing, including the DF function.

   - Handling of multi-destination traffic and processing of broadcast and multicast as per [RFC7432].

d. The solution is backwards 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 Network Virtualization Overlay 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

This solution extends the [RFC7432] Inclusive Multicast Ethernet Tag routes and attributes so that an NVE/PE can signal its optimized Ingress Replication capabilities.

The NLRI of the Inclusive Multicast Ethernet Tag route as in [RFC7432] is shown in Figure 2 and it is used in this document without any modifications to its format. The PMSI Tunnel Attribute's general format as 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:

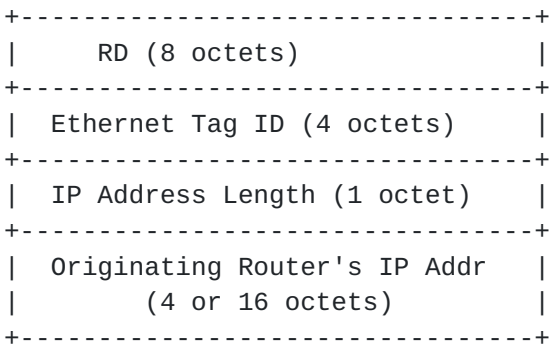

Figure 2: EVPN Inclusive Multicast Tag route's NLRI

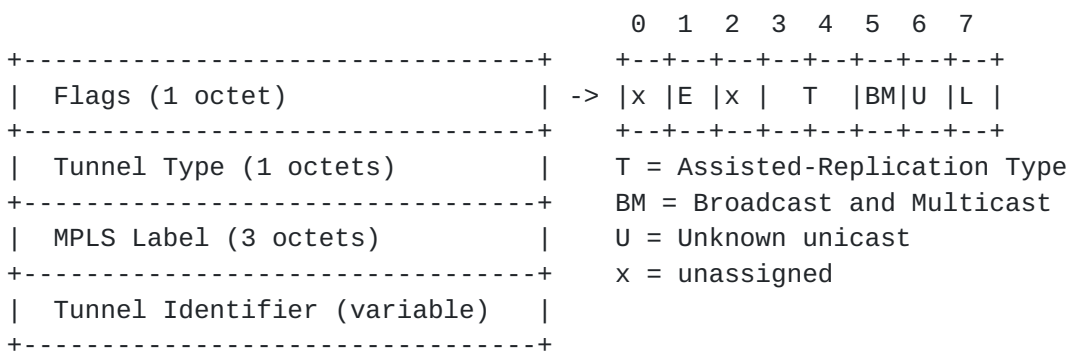

Figure 3: PMSI Tunnel Attribute

The Flags field in Figure 3 is 8 bits long as per [RFC7902], where the Extension flag (E) and the Leaf Information Required (L) Flag are already allocated. This document defines the use of 4 bits of this Flags field, and suggests the following allocation to IANA:

- bits 3 and 4, forming together 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-Flood Lists (PFL) flags.

The T field and Pruned-Flood-Lists flags are defined as follows:

- T is the Assisted-Replication Type field (2 bits) that 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 Pruned-Flood-Lists 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 means regular behavior.

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

- Flag L is an existing flag defined in [RFC6514] (L=Leaf Information Required, bit 7) and it 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 type T 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 SHOULD be set to the same IP address as the Originating Router's IP address when the NVE/PE originates the route, that is, when the NVE/PE is not an ASBR as in 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 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 in Section 8 are followed.

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

o T (AR role type) MUST be set to 01 (AR-REPLICATOR).

o 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 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 the AR-LEAFs what outer IP DA, i.e., the 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 [I-D.ietf-bess-evpn-bum-procedure-updates] in case 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 Auto-Discovery route is only used for selective AR and the fields of such route are set as follows:

o Originating Router's IP Address is set to the advertising router's IP address (same IP 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 as in section 10.2 of [RFC8365].

o Route Key is the "Route Type Specific" NLRI of the Replicator-AR route for which this Leaf Auto-Discovery route is generated.

o The AR-LEAF constructs an IP-address-specific route-target, analogously to [I-D.ietf-bess-evpn-bum-procedure-updates], by

placing the IP address carried in the Next-Hop field of the received Replicator-AR route in the Global Administrator field of the Community, with the Local Administrator field of this Community set to 0, and setting the Extended Communities attribute of the Leaf Auto-Discovery route to that 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 Auto-Discovery routes.

- The Leaf Auto-Discovery route MUST include the PMSI Tunnel attribute with the Tunnel Type set to AR (Section 11), T (AR role type) set to AR-LEAF and the 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 process 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 AR 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 (Pruned-Flood-Lists flags). Note that these BM/U flags may be used to optimize the delivery of multi-destination traffic and their use SHOULD be an administrative choice, and independent of the AR role. When the Pruned-Flood-List capability is enabled, the BM/U flags can be used with the Regular-IR, Replicator-AR and Leaf Auto-Discovery 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 Network Virtualization Overlay network where the non-selective AR function is enabled. Three different roles are defined for a given BD: AR-REPLICATOR, AR-LEAF and RNVE (Regular NVE). The solution is called "non-selective"

because the chosen AR-REPLICATOR for a given flow MUST replicate the BM traffic to all the NVE/PEs in the BD except for the source NVE/PE. Network Virtualization Overlay tunnels, i.e., IP tunnels, exist among all the PEs and NVEs in the diagram. The PEs and NVEs in the diagram have Tenant Systems or Virtual Machines connected to their Attachment Circuits.

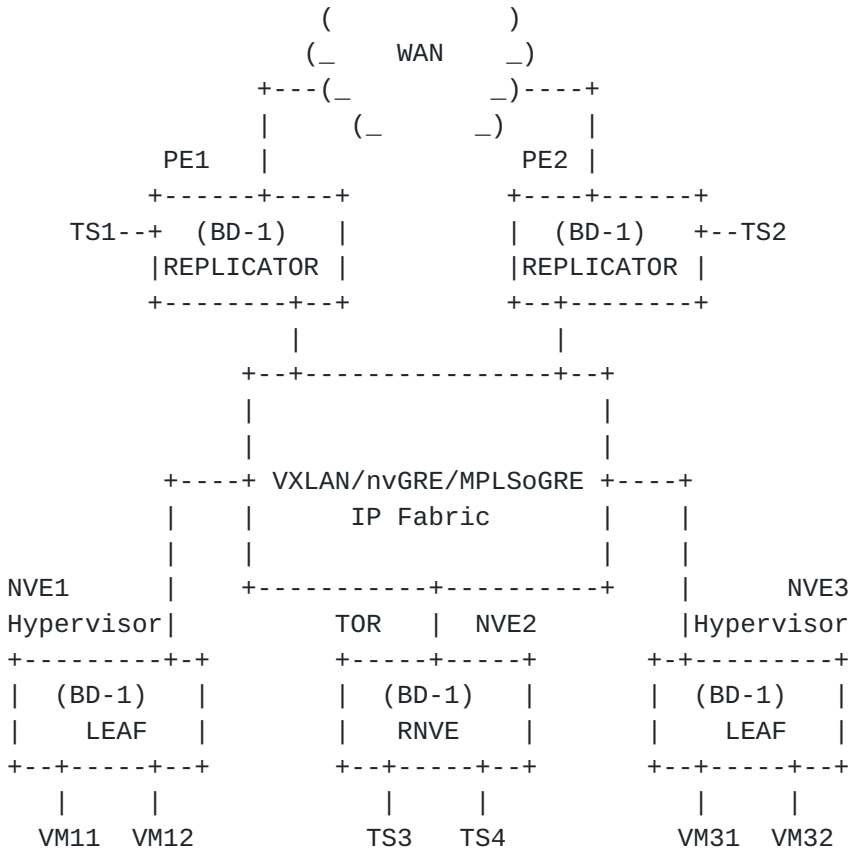

Figure 4: Non-Selective AR scenario

In AR BDs such as BD-1 in the example, BM (Broadcast and Multicast) traffic between two NVEs may follow a different path than unicast traffic. This solution recommends the replication of BM through the AR-REPLICATOR node, whereas unknown/known unicast 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 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 Attachment Circuits. 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 as 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 attachment circuits (AC). If the Regular-IR route is advertised, the Assisted-Replication Type field of the Regular-IR route MUST be set to zero.

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 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 in [RFC7432].

   - If the destination IP address is the AR-REPLICATOR AR-IP Address the node MUST replicate the packet to local Attachment Circuits 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. That will be an indication for 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 Attachment Circuits 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 Attachment Circuit, it will forward the BM packet to its flooding list (including local Attachment Circuits and remote NVE/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:

  o 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 Attachment Circuits. This is the regular Ingress Replication behavior described in [RFC7432].

  o 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 the traffic is not sent back to the originating AR-LEAF.

  o If the encapsulation is MPLSoGRE or MPLSoUDP and the received BD label that the AR-REPLICATOR advertised in the Replicator-AR route is not the bottom of the stack, the AR-REPLICATOR MUST copy the 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 flood-list composed of Attachment Circuits 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.

  o When an AR-REPLICATOR/LEAF receives an unknown unicast packet on an Attachment Circuit, it will forward the unknown unicast packet to its flood-list, skipping the non-U overlay tunnels.

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

5.2. Non-Selective AR-LEAF Procedures

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-REPLICATOR and RNVEs). A given service can have zero, one or more AR-LEAF nodes. Figure 4 shows NVE1 and NVE3 (both residing in hypervisors) acting as AR-LEAF. 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 as per-BD option.

b. In this non-selective AR solution, the AR-LEAF MUST advertise a single Regular-IR inclusive multicast route as in [RFC7432]. The AR-LEAF SHOULD set the Assisted-Replication Type field to AR-LEAF. Note that although this field does not make any difference for the remote nodes when creating an EVPN destination to the AR-LEAF, this field is useful for an easy 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 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 any other detection mechanism).

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

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

d. In a service where there is 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 attachment circuits (ACs) for a given

BD. This selection is a local decision and it 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 case of a 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 AR-REPLICATOR(s) selection dynamically, due to an administrative or policy configuration change.

- AR-LEAF nodes SHALL send service-level BM control plane packets following regular Ingress Replication procedures. An example would be IGMP, MLD or PIM multicast 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 regular IR-IP Address.

e. The use of an AR-REPLICATOR-activation-timer (in seconds, default value is 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 case of a new added AR-REPLICATOR, or in case the 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 account for the time it takes for the AR-LEAF Regular-IR inclusive multicast route to get to the AR-REPLICATOR and be programmed. While the AR-REPLICATOR-activation-time is running, the AR-LEAF node will use regular ingress replication.

f. If the AR-LEAF has selected an AR-REPLICATOR, it is a matter of local policy to change to a new preferred AR-REPLICATOR for the existing BM traffic flows.

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

- The AR-LEAF nodes will build two flood-lists:

  1. Flood-list #1 - composed of Attachment Circuits 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 point d) above and it is a local AR-LEAF decision.

  2. Flood-list #2 - composed of Attachment Circuits and overlay tunnels to the remote IR-IP Addresses.

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

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

  o If the AR-REPLICATOR-set is NOT empty, the AR-LEAF MUST send the packet to flood-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 Attachment Circuits 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

RNVE (Regular Network Virtualization Edge node) 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 Auto-Discovery routes (due to the IP-address-specific route-target).

This role provides EVPN with the backwards compatibility required in optimized Ingress Replication BDs. Figure 4 shows NVE2 as RNVE.

6. Selective Assisted-Replication (AR) Solution Description

Figure 5 is used to describe the selective AR solution.

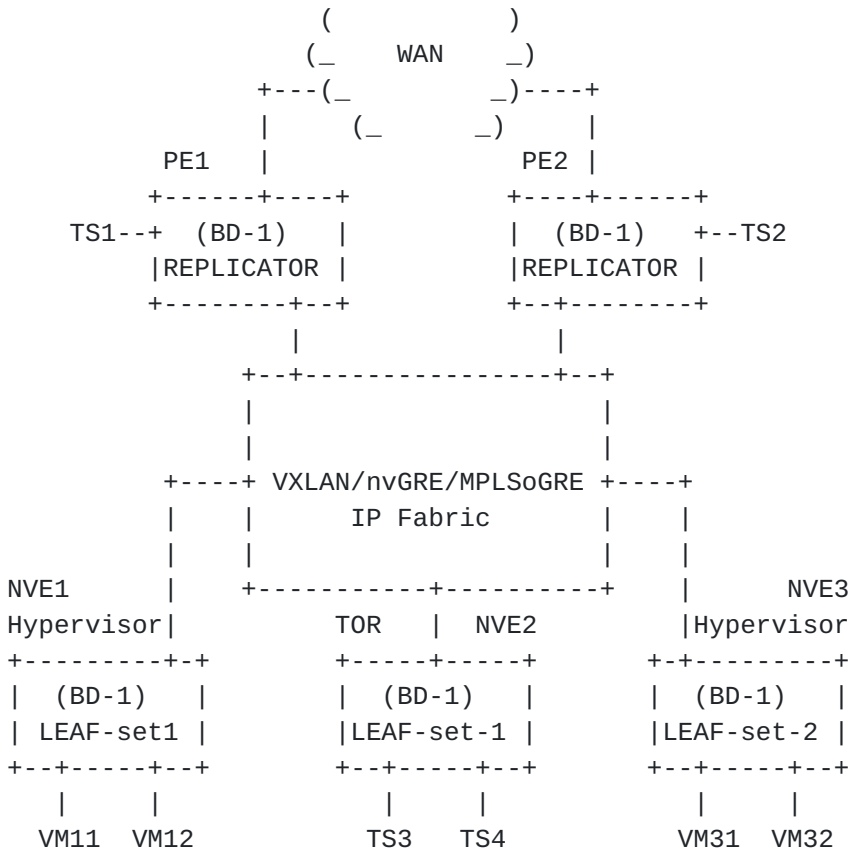

Figure 5: Selective AR scenario

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 defined in Section 4 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 two-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 sub-sections describe the differences in the procedures of AR-REPLICATOR/LEAFs compared to the non-selective AR solution. There is no change on the 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 capability SHOULD be an administrative choice in any NVE/PE that is part of an Assisted-Replication-enabled BD, as the AR role itself. This administrative option MAY be implemented as a system level option as opposed to as 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 Replicator-AR route MUST include L=1 (Leaf Information Required) in 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 Auto-Discovery route 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 NVE1 and NVE2 advertise a Leaf Auto-Discovery route with PE1's IP-address-specific route-target and NVE3 advertises a Leaf Auto-Discovery 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 Attachment Circuits

+ 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 IP AR-IP will be an indication for 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 flood-lists:

  1. Flood-list #1 - composed of Attachment Circuits and overlay tunnels to the remote nodes in the BD, always using the IR-IPs in the tunnel destination IP addresses.

  2. Flood-list #2 - composed of Attachment Circuits, 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 Auto-Discovery route for the local AR-REPLICATOR. This set is updated with every Leaf Auto-Discovery 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.

- Some of the overlay tunnels in the flood-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 Attachment Circuit, it MUST forward the BM packet to its flood-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:

  o 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 flood-list #2, unless some AR-REPLICATOR within the BD has advertised L=0. In the latter case, the node reverts back to non-selective mode and flood-list #1 MUST be used. Non-BM overlay tunnels are skipped when sending BM packets.

  o 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 flood-list #2 but skipping the AR-REPLICATOR-set. Non-BM overlay tunnels are skipped when sending BM packets.

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

- In any case, the AR-REPLICATOR ensures 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 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 of 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-LEAF/REPLICATORs send BM traffic, NVE1 will receive that traffic from PE1. These are the differences in the behavior of a Selective AR-LEAF compared to a non-selective AR-LEAF:

a. The AR-LEAF role selective capability SHOULD be an administrative choice in any NVE/PE that is part of an Assisted-Replication-enabled BD. This administrative option to enable AR-LEAF capabilities MAY be implemented as a system level option as opposed to as 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 Auto-Discovery route after receiving a Replicator-AR route with L=1. It is RECOMMENDED that the Selective AR-LEAF waits for an AR-LEAF-join-wait-timer (in seconds, default value is 3) before sending the Leaf Auto-Discovery route, so that the AR-LEAF can collect all the Replicator-AR routes for the BD before advertising the Leaf Auto-Discovery route. If the Replicator-AR route with L=1 is withdrawn, the corresponding Leaf Auto-Discovery 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:

   o The Selective AR-LEAF MUST send a Leaf Auto-Discovery route including the Route-key and IP-address-specific route-target of the selected AR-REPLICATOR.

   o The Selective AR-LEAF MUST send all the BM packets received on the attachment circuits (ACs) for a given BD to that AR-REPLICATOR.

   o In case of a failure on the selected AR-REPLICATOR (detected when the Replicator-AR route becomes infeasible as the result of any of the underlying BGP mechanisms), another AR-REPLICATOR will be selected and a new Leaf Auto-Discovery update will be issued for the new AR-REPLICATOR. This new route will update the selective list in the new Selective AR-REPLICATOR. In case of failure of the active Selective AR-REPLICATOR, it is RECOMMENDED for the Selective AR-LEAF to

revert to Ingress Replication behavior for a timer AR-REPLICATOR-activation-timer (in seconds, default value is 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 in Section 5.2.

- o A Selective AR-LEAF MAY change the AR-REPLICATOR(s) selection 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. In case there is a non-selective AR-LEAF, 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 flood-lists:

  1. Flood-list #1 - composed of Attachment Circuits and the overlay tunnel to the selected AR-REPLICATOR (using the AR-IP as the tunnel destination IP address).

  2. Flood-list #2 - composed of Attachment Circuits and overlay tunnels to the remote IR-IP addresses.

- Some of the overlay tunnels in the flood-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 Attachment Circuit, it will check if there is any selected AR-REPLICATOR. If there is, flood-list #1 MUST be used. Otherwise, flood-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 Attachment Circuits and never to an overlay tunnel. This is the regular Ingress Replication behavior described in [RFC7432].

7. Pruned-Flood-Lists (PFL)

In addition to AR, the second optimization supported by this solution is the ability for the all the BD nodes to signal Pruned-Flood-Lists (PFL). As described in Section 4, an EVPN node can signal a given value for the BM and U Pruned-Food-Lists flags in the Regular-IR, Replicator-AR or Leaf Auto-Discovery routes, where:

- BM is the Broadcast and Multicast flag. BM=1 means "prune-me" from the BM flood-list. BM=0 means regular behavior.

- U is the Unknown flag. U=1 means "prune-me" from the Unknown flood-list. U=0 means regular behavior.

The ability to signal and process these Pruned-Flood-Lists flags SHOULD be an administrative choice. If a node is configured to process the Pruned-Flood-Lists flags, upon receiving a non-zero Pruned-Flood-Lists flag for a route, the NVE/PE will add the corresponding flag to the created overlay tunnel in the flood-list. When replicating a BM packet in the context of a flood-list, the NVE/PE will skip the overlay tunnels marked with the flag BM=1, since the NVE/PE 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 Pruned-Flood-Lists flags. An AR-LEAF or RNVE receiving BUM traffic on an overlay tunnel MUST replicate the traffic to its local Attachment Circuits, regardless of the BM/U flags on the overlay tunnels.

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

7.1. A Pruned-Flood-List Example

In order to illustrate the use of the solution described in this document, 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 both NVEs advertise all the attached Virtual

Machines MAC and IP addresses in EVPN as soon as they come up, and these NVEs do not have any Virtual Machines interested in multicast applications, they will be configured to signal BM/U flags = 11 for BD-1. That is, neither NVE1 nor NVE3 are interested in receiving BM or Unknown Unicast traffic since:

- o Their attached VMs (VM11, VM12, VM31, VM32) do not support multicast applications.
- o Their attached VMs will not receive ARP Requests. Proxy-ARP [I-D.ietf-bess-evpn-proxy-arp-nd] on the remote NVE/PEs will reply ARP Requests locally, and no other Broadcast is expected.
- o Their attached VMs will not receive unknown unicast traffic, since the VMs' MAC and IP addresses are always advertised by EVPN 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 forward further the BM packets to TS1, WAN link, PE2 and NVE2, but not to NVE3. PE2 and NVE2 will replicate the BM packets to their local Attachment Circuits but we will avoid NVE3 having to replicate unnecessarily those BM packets to VM31 and VM32.

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 Virtual Machines. PE1 and NVE2 will replicate to their local Attachment Circuits only.

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

8. AR Procedures for Single-IP AR-REPLICATORS

The procedures explained in sections Section 5 and Section 6 assume that the AR-REPLICATOR can use two local routable IP addresses to terminate and originate Network Virtualization Overlay 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 Network Virtualization Overlay 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 in sections Section 5 and Section 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 AR-VNI along with the Replicator-AR route and 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 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 the Ingress Replication or Assisted Replication forwarding mode at the subsequent AR-REPLICATOR.

The rest of the procedures will follow what is described in sections Section 5 and Section 6.

9. AR Procedures and EVPN All-Active Multi-homing Split-Horizon

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

9.1. Ethernet Segments on AR-LEAF Nodes

If VXLAN or NVGRE are used, and if the Split-horizon is based on the tunnel IP Source Address and "Local-Bias" as described in [RFC8365], the Split-horizon check will not work if there is an Ethernet-Segment shared between two AR-LEAF nodes, and the AR-REPLICATOR replaces the tunnel IP Source Address of the packets with its own AR-IP.

In order to be compatible with the IP Source Address split-horizon check, the AR-REPLICATOR MAY keep the original received tunnel IP Source 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 Ethernet-Segment. Even if the AR-LEAF's IP Source Address is preserved when replicating to AR-LEAFs or RNVEs, the AR-REPLICATOR MUST always use its IR-IP as the IP Source Address when replicating to other AR-REPLICATORs.

When EVPN is used for MPLS over GRE (or UDP), the ESI-label based split-horizon procedure as in [RFC7432] will not work for multi-homed Ethernet-Segments defined on AR-LEAF nodes. "Local-Bias" is recommended in this case, as in the case of VXLAN or NVGRE explained above. The "Local-Bias" and tunnel IP Source 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 IP Source Address, the use of uRPF (unicast Reverse Path Forwarding) checks in the IP fabric based on the tunnel IP Source Address MUST be disabled.

9.2. Ethernet Segments on AR-REPLICATOR nodes

AR-REPLICATOR nodes attached to the same all-active Ethernet Segment will follow "Local-Bias" procedures [RFC8365], as follows:

a. For BUM traffic received on a local AR-REPLICATOR's Attachment Circuit, "Local-Bias" procedures as in [RFC8365] MUST be followed.

b. For BUM traffic received on an AR-REPLICATOR overlay tunnel with AR-IP as the IP Destination Address, "Local-Bias" MUST also be followed. That is, traffic received with AR-IP as IP Destination Address will be treated as though it had been received on a local Attachment Circuit that is part of the Ethernet Segment and will be forwarded to all local Ethernet Segments, irrespective of their DF or NDF state.

c. BUM traffic received on an AR-REPLICATOR overlay tunnel with IR-IP as the IP Destination Address, will follow regular [RFC8365] "Local-Bias" rules and will not be forwarded to local Ethernet Segments 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 in 'b' above if the received VNI is the AR-VNI, and as in '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 Auto-Discovery route in [I-D.ietf-bess-evpn-bum-procedure-updates] apply too.

In addition, the Assisted-Replication method introduced by this document may bring some new risks for the successful delivery of BM traffic. Unicast traffic is not affected by Assisted-Replication (although Unknown unicast traffic is affected by the Pruned-Flood-Lists procedures). The forwarding of Broadcast and Multicast (BM) traffic is modified, and BM traffic from the AR-LEAF nodes will be attracted by the existence of 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 change of the advertised AR type will modify the selection on 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 on their performance, since the AR-LEAF nodes are usually NVEs/PEs with poor replication performance.

This document introduces the ability for the AR-REPLICATOR to forward traffic received on an overlay tunnel to another overlay tunnel. The reader may interpret 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. The procedures in this document prevent these BM loops, since the AR-REPLICATOR will always forward the BM traffic using the correct tunnel IP Destination Address (or correct VNI in case of single-IP AR-REPLICATORs) that instructs the remote nodes how to forward the traffic. This is true in both the Non-Selective and Selective modes defined in this document. However, a wrong implementation of the procedures in this document may lead to those unexpected BM loops.

The Selective mode provides a multi-staged replication solution, where a proper configuration of all the AR-REPLICATORs will avoid 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 fall back solution to Non-Selective mode in case the AR-REPLICATORs advertised an inconsistent AR Replication mode.

This document allows the AR-REPLICATOR to preserve the tunnel IP Source Address of the AR-LEAF (as an option) when forwarding BM packets from an overlay tunnel to another overlay tunnel. Preserving the AR-LEAF IP Source Address makes the "Local Bias" filtering procedures possible for AR-LEAF nodes that are attached to the same Ethernet Segment. If the AR-REPLICATOR does not preserve the AR-LEAF IP Source Address, AR-LEAF nodes attached to all-active Ethernet Segments will cause packet duplication on the multi-homed CE.

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

Finally, the use of PFL as in Section 7, should be handled with care. An 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	[This document]

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

Value	Name	Reference
3-4	Assisted-Replication Type (T)	[This document]
5	Broadcast and Multicast (BM)	[This document]
6	Unknown (U)	[This document]

12. Contributors

In addition to the names in the front page, the following co-authors also contributed to this document:

Wim Henderickx
Nokia

Kiran Nagaraj
Nokia

Ravi Shekhar
Juniper Networks

Nischal Sheth
Juniper Networks

Aldrin Isaac
Juniper

Mudassir Tufail
Citibank

13. Acknowledgments

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 that improved the quality of the document significantly.

14. References

14.1. Normative References

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[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>.

[I-D.ietf-bess-evpn-bum-procedure-updates] Zhang, Z., Lin, W., Rabadan, J., Patel, K., and A. Sajassi, "Updates on EVPN BUM Procedures", draft-ietf-bess-evpn-bum-procedure-updates-14 (work in progress), November 2021.

[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>.

[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>.

[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>.

14.2. Informative References

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

[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>.

[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>.

[I-D.ietf-bess-evpn-proxy-arp-nd]
Rabadan, J., Sathappan, S., Nagaraj, K., Hankins, G., and T. King, "Operational Aspects of Proxy ARP/ND in Ethernet Virtual Private Networks", draft-ietf-bess-evpn-proxy-arp-nd-16 (work in progress), October 2021.

Authors' Addresses

J. Rabadan (editor)
Nokia
777 Middlefield Road
Mountain View, CA 94043
USA

Email: jorge.rabadan@nokia.com

S. Sathappan
Nokia

Email: senthil.sathappan@nokia.com

W. Lin
Juniper Networks

Email: wlin@juniper.net

M. Katiyar
Versa Networks

Email: mukul@versa-networks.com

A. Sajassi
Cisco Systems

Email: sajassi@cisco.com