BESS Workgroup
Internet-Draft
Intended status: Standards Track
Expires: March 27, 2022

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

September 23, 2021

Optimized Ingress Replication solution for EVPN
draft-ietf-bess-evpn-optimized-ir-09

Abstract

Network Virtualization Overlay (NVO) networks using EVPN as control plane may use Ingress Replication (IR) 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 NVO core network. IR avoids the dependency on PIM in the NVO network core. While IR 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 IR in NVO networks.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on March 27, 2022.

Table of Contents

1. Introduction

Ethernet Virtual Private Networks (EVPN) may be used as the control plane for a Network Virtualization Overlay (NVO) network. Network Virtualization Edge (NVE) devices and Provider Edges (PEs) that are

part of the same EVPN Instance (EVI) use Ingress Replication (IR) or PIM-based trees to transport the tenant's Broadcast, Unknown unicast and Multicast (BUM) traffic. In NVO networks where PIM-based trees cannot be used, IR is the only option. Examples of these situations are NVO networks where the core nodes don't 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. Some NVEs can even provide Dynamic Host Configuration Protocol (DHCP) server functions for the attached Tenant Systems (TS) reducing the broadcast even further.

b. Unknown unicast traffic is greatly reduced in virtualized NVO networks where all the MAC and IP addresses are learned in the control plane.

c. Multicast applications are not used.

If the above assumptions are true for a given NVO network, then IR 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 IR optimizations:

1. Assisted-Replication (AR)

2. Pruned-Flood-Lists (PFL)

Both optimizations 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 [RFC7432] that are described in Section 4.

Section 3 lists the requirements of the combined optimized-IR solution, whereas Section 5 and Section 6 describe the Assisted-Replication (AR) solution, and Section 7 the Pruned-Flood-Lists (PFL) 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:

- AC: Attachment Circuit
- BM traffic: Refers to Broadcast and Multicast frames (excluding unknown unicast frames)
- NVO: Network Virtualization Overlay
- NVE: Network Virtualization Edge router
- PE: Provider Edge router
- 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. This document defines the control and data plane procedures that an AR-REPLICATOR needs to follow.
- AR-LEAF: Assisted Replication - LEAF, refers to an NVE/PE that - given its poor replication performance - sends all the Broadcast and Multicast traffic to an AR-REPLICATOR that can replicate the traffic further on its behalf.
- 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 RT-3 (route type 3) that is advertised by an AR-REPLICATOR to signal its capabilities.
- Regular-IR: Refers to Regular Ingress Replication, where the source NVE/PE sends a copy to each remote NVE/PE part of the BD.

- AR-IP: IP address owned by the AR-REPLICATOR and used to differentiate the ingress traffic that must follow the AR procedures.

- IR-IP: IP address used for Ingress Replication as in [RFC7432].

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

- IR-VNI: VNI advertised along with the RT-3 for IR.

- AR forwarding mode: for an AR-LEAF, it means sending an AC 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.

- IR forwarding mode: it refers to the Ingress Replication behavior explained in [RFC7432]. It means sending an AC BM packet copy to each remote PE/NVE in the BD and sending an overlay BM packet only to the ACs and not other overlay tunnels.

- PTA: PMSI Tunnel Attribute

- RT-3: EVPN Route Type 3, Inclusive Multicast Ethernet Tag route

- RT-11: EVPN Route Type 11, Leaf Auto-Discovery (A-D) route

- VXLAN: Virtual Extensible LAN

- GRE: Generic Routing Encapsulation

- NVGRE: Network Virtualization using Generic Routing Encapsulation

- GENEVE: Generic Network Virtualization Encapsulation

- VNI: VXLAN Network Identifier

- EVI: EVPN Instance. An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN

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

- TOR: Top Of Rack switch

3. Solution requirements

The IR optimization solution specified in this document (optimized-IR hereafter) meets the following requirements:

a. It provides an IR optimization for BM (Broadcast and Multicast) traffic without the need for PIM, while preserving the packet order for unicast applications, i.e., known and unknown unicast traffic should follow the same path. This optimization is required in low-performance NVEs.

b. It 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 EVPN 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].

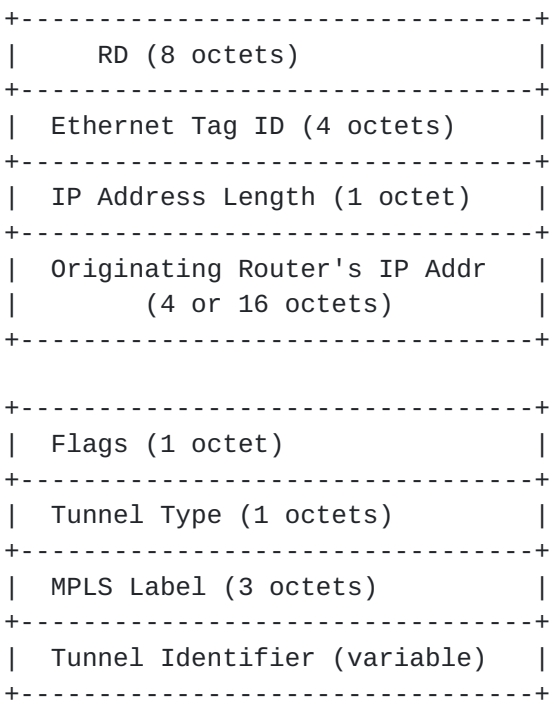
d. The solution is backwards compatible with existing NVEs using a non-optimized version of IR. A given BD can have NVEs/PEs supporting regular-IR and optimized-IR.

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-IR

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

The Inclusive Multicast Ethernet Tag route (RT-3) and its PMSI Tunnel Attribute's (PTA) general format used in [RFC7432] are shown below:



The Flags field is 8 bits long. This document defines the use of 4 bits of this Flags field:

- 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 PFL (Pruned-Flood Lists) flags.

The T field and PFL flags are defined as follows:

- T is the AR Type field (2 bits) that defines the AR role of the advertising router:

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

  o 01 (decimal 1) = AR-REPLICATOR

  o 10 (decimal 2) = AR-LEAF

  o 11 (decimal 3) = RESERVED

- The PFL (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) and it will be used only in the Selective AR Solution.

Please refer to Section 11 for the IANA considerations related to the PTA flags.

In this document, the above RT-3 and PTA 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 SHOULD be advertised with type T= 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 PE that should be common to all the EVIs on the PE (usually this is the PE's loopback address). 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. The Next-Hop address is referred to as the AR-IP and SHOULD be different than the IR-IP for a given PE/NVE.

  - Tunnel Type = Assisted-Replication Tunnel. Section 11 provides the allocated type value.

  - T (AR role type) = 01 (AR-REPLICATOR).

  - L (Leaf Information Required) = 0 (for non-selective AR) or 1 (for selective AR).

In addition, this document also uses the Leaf A-D route (RT-11) defined in [I-D.ietf-bess-evpn-bum-procedure-updates] in case the

selective AR mode is used. The Leaf A-D route MAY be used by the AR-LEAF in response to a Replicator-AR route (with the L flag set) to advertise its desire to receive the BM traffic from a specific AR-REPLICATOR. It is only used for selective AR and its fields are set as follows:

- Originating Router's IP Address is set to the advertising PE's IP address (same IP used by the AR-LEAF in regular-IR routes). The Next-Hop address is set to the IR-IP.

- Route Key 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 as indicated in [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. Note that the same IP-address-specific import route-target is auto-configured by the AR-REPLICATOR that sent the Replicator-AR, in order to control the acceptance of the Leaf A-D routes.

- The Leaf A-D route MUST include the PMSI Tunnel attribute with the Tunnel Type set to AR, type set to AR-LEAF and the Tunnel Identifier set to the IP 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 MUST understand and process the AR type field in the PTA (Flags field) of the routes, and MUST signal the corresponding type (1 or 2) according to its administrative choice.

Each node attached to the BD may understand and process the BM/U flags. Note that these BM/U flags may be used to optimize the delivery of multi-destination traffic and its use SHOULD be an administrative choice, and independent of the AR role.

Non-optimized-IR nodes will be unaware of the new PMSI attribute flag definition as well as the new Tunnel Type (AR), i.e. they will ignore the information contained in the flags field for any RT-3 and will ignore the RT-3 routes with an unknown Tunnel Type (type AR in this case).

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

Figure 1 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 (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.

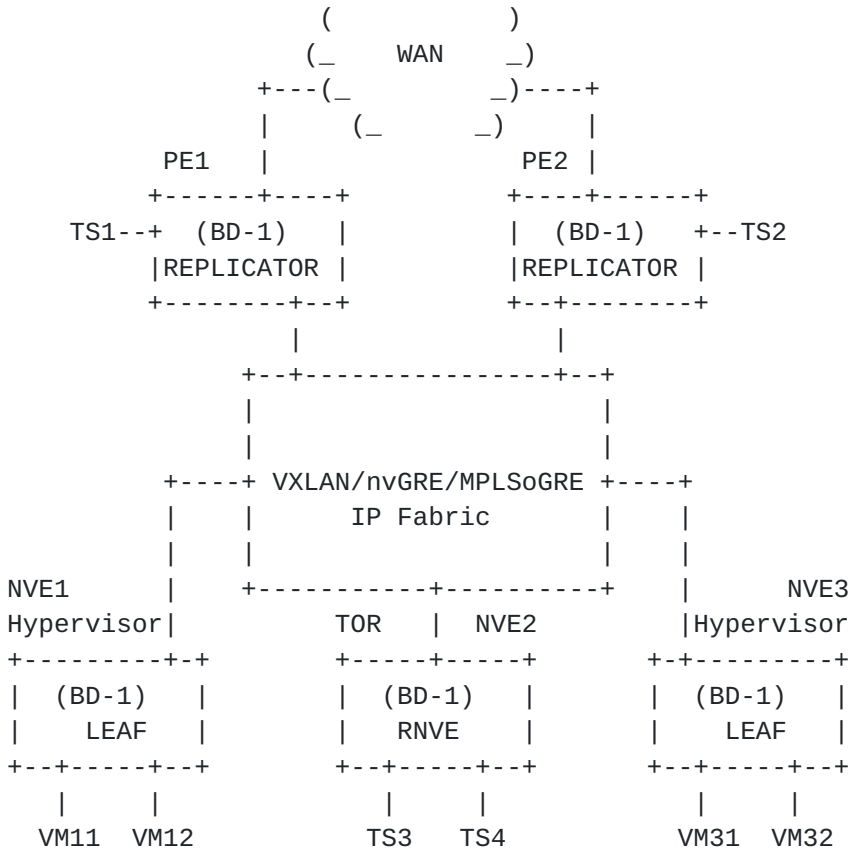


Figure 1: Optimized-IR scenario

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. Unknown unicast SHALL follow the same path as known unicast traffic in order to avoid packet reordering for unicast applications and simplify the control and data plane procedures.

Note that known unicast forwarding is not impacted by this solution.

5.1. Non-selective AR-REPLICATOR procedures

An AR-REPLICATOR is defined as an NVE/PE capable of replicating ingress BM (Broadcast and Multicast) traffic received on an overlay tunnel to other overlay tunnels and local Attachment Circuits (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 1, 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 AR Type field is set to zero.

c. The Replicator-AR and Regular-IR routes are generated according to section 3. The AR-IP and IR-IP used by the AR-REPLICATOR are different routable IP addresses.

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

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

    - If the destination IP 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 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 used by the AR-REPLICATOR MUST be its IR-IP when replicating to either AR-REPLICATOR or AR-LEAF nodes.

An AR-REPLICATOR will 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 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 of the underlay IP header and:

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

  o If the destination IP 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 IR behavior described in [RFC7432].

- While the forwarding behavior in AR-REPLICATORs and AR-LEAF nodes is different for BM traffic, as far as Unknown unicast traffic forwarding is concerned, AR-LEAF nodes behave exactly in the same way as AR-REPLICATORs do.

- The AR-REPLICATOR/LEAF nodes will build an Unknown unicast flood-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.

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

  o When an AR-REPLICATOR/LEAF receives an unknown packet on an overlay tunnel will forward the unknown packet to its local ACs and never to an overlay tunnel. This is the regular IR 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 1 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 AR Type field to AR-LEAF. Note that although this flag does not make any difference for the egress nodes when creating an EVPN destination to the AR-LEAF, it is RECOMMENDED to use this flag for an easy operation and troubleshooting of the BD.

c. In a service where there are no AR-REPLICATORs, the AR-LEAF MUST use regular ingress replication. This will happen when a new update from the last former AR-REPLICATOR is received and contains a non-REPLICATOR AR type, or when the AR-LEAF detects that the last AR-REPLICATOR is down (via next-hop tracking in the IGP or any other detection mechanism). Ingress replication MUST use the forwarding information given by the remote Regular-IR Inclusive Multicast Routes as described in [RFC7432].

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-LEAF's selection 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 on the selected AR-REPLICATOR, another AR-REPLICATOR will be selected.

   - When an AR-REPLICATOR is selected, the AR-LEAF MUST send all the BM packets 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.

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

An AR-LEAF will 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 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 point d) above and it is a local AR-LEAF decision.

  2. Flood-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:

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

  - o If the AR-REPLICATOR-set is NOT empty, the AR-LEAF will 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, will forward the BM packet to its local ACs and never to an overlay tunnel. This is the regular IR behavior described in [RFC7432].

- AR-LEAF nodes process Unknown unicast traffic in the same way AR-REPLICATORS do, as described in section 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 IR 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 PTA) and the Leaf A-D routes (due to the IP-address-specific route-target).

This role provides EVPN with the backwards compatibility required in optimized-IR BDs. Figure 1 shows NVE2 as RNVE.

6. Selective Assisted-Replication (AR) Solution Description

Figure 1 is also used to describe the selective AR solution, however in this section we consider NVE2 as one more AR-LEAF for BD-1. The solution is called "selective" because a given AR-REPLICATOR MUST replicate the BM traffic to only the AR-LEAF that requested the replication (as opposed to all the AR-LEAF nodes) and MAY replicate the BM traffic to the RNVEs. The same AR roles defined in Section 4 are used here, however the procedures are different.

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 1, 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 AR-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 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 section 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 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 MUST only add NVE1/NVE2 to its selective AR-LEAF-set for BD-1, and exclude NVE3.

- When a node defined and operating as Selective AR-REPLICATOR receives a packet on an overlay tunnel, it will do a tunnel destination IP lookup and if the destination IP 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 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 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), where the tunnel destination IP 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 will be compatible with the following rules:

- The Selective AR-REPLICATORs will build two flood-lists:

  1. Flood-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. 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.

  2. Flood-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.

- When a Selective AR-REPLICATOR receives a BM packet on an AC, it will 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 matches its AR-IP and the source IP matches an IP of its own Selective AR-LEAF-set, the AR-REPLICATOR will forward the BM packet to its flood-list #2, as long as the list of AR-REPLICATORs for the BD matches the Selective AR-REPLICATOR-set. If the Selective AR-REPLICATOR-set does not match the list of AR-REPLICATORs, the node reverts back to non-selective mode and flood-list #1 is used.

  o If the destination IP matches its AR-IP and the source IP does not match any IP of its Selective AR-LEAF-set, the AR-REPLICATOR will forward the BM packet to flood-list #2 but skipping the AR-REPLICATOR-set.

  o If the destination IP matches its IR-IP, the AR-REPLICATOR will use flood-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-IR behavior described in [RFC7432].

- In any case, non-BM overlay tunnels are excluded from flood-lists and, also, source squelching is always done in order to ensure the traffic is not sent back to the originating source. If the encapsulation is MPLSoGRE (or MPLSoUDP) and the BD label 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 BM traffic to that AR-REPLICATOR and
- Expects to receive the BM traffic for a given BD from the same AR-REPLICATOR.

In the example of Figure 1, 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 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. 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 waits for a AR-LEAF-join-wait-timer (in seconds, default value is 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.

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

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

- In case of a failure on the selected AR-REPLICATOR, 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 case of failure on the active Selective AR-REPLICATOR, it is RECOMMENDED for the Selective AR-LEAF to revert to IR behavior for a timer AR-REPLICATOR-activation-timer (in seconds, default value is 3) to speed up the convergence. 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.

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 will 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 ACs and the overlay tunnel to the selected AR-REPLICATOR (using the AR-IP as the tunnel destination IP).

  2. Flood-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 if there is any selected AR-REPLICATOR. If there is, flood-list #1 will be used. Otherwise, flood-list #2 will.

- When an AR-LEAF receives a BM packet on an overlay tunnel, will forward the BM packet to its local ACs and never to an overlay tunnel. This is the regular IR 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 3, an EVPN node can signal a given value for the BM and U PFL flags in the IR Inclusive Multicast 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 these PFL flags is an administrative choice. Upon receiving a non-zero PFL flag, a node MAY decide to honor the PFL flag and remove the sender from the corresponding flood-list. A given BD node receiving BUM traffic on an overlay tunnel MUST replicate the traffic normally, regardless of the signaled PFL flags.

This optimization MAY be used along with the AR solution.

7.1. A PFL example

In order to illustrate the use of the solution described in this document, we will assume that BD-1 in figure 1 is optimized-IR 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 VMs in EVPN as soon as they come up and don't have any VMs interested in multicast applications, they will be configured to signal BM/U flags = 11 for BD-1.

- NVE2 is optimized-IR 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 ACs 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 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 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 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 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 IR from the packets that must use AR 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 RDs are needed in each route.

- An AR-REPLICATOR will perform IR or AR 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 IR or AR 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 AR-LEAF nodes or AR-REPLICATOR nodes are attached to the the same Ethernet Segment in the BD. The case where one (or more) AR-LEAF node(s) and one (or more) AR-REPLICATOR node(s) are attached to the same Ethernet Segment 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 SA 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 changes the tunnel IP SA of the packets with its own AR-IP.

In order to be compatible with the IP SA split-horizon check, the AR-REPLICATOR MAY keep the original received tunnel IP SA 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 SA is preserved when replicating to AR-LEAFs or RNVEs, the AR-REPLICATOR MUST always use its IR-IP as IP SA 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 SA 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 SA, the use of uRPF (unicast Reverse Path Forwarding) checks in the IP fabric based on the tunnel IP SA MUST be disabled.

9.2. Ethernet Segments on AR-REPLICATOR nodes

Ethernet Segments associated to one or more AR-REPLICATOR nodes SHOULD follow "Local-Bias" procedures for EVPN all-active multi-homing, as follows:

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

- For BUM traffic received on an AR-REPLICATOR overlay tunnel with AR-IP as the IP DA, "Local-Bias" SHOULD also be followed. That is, traffic received with AR-IP as 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 ES, irrespective of their DF or NDF state.

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

10. Security Considerations

The Security Considerations in [RFC7432] and [RFC8365] apply to this document.

In addition, the procedures introduced by this document may bring some new risks for the successful delivery of BM traffic. Unicast traffic is not affected by this document. The forwarding of Broadcast and Multicast (BM) traffic is modified though, and BM traffic from the AR-LEAF nodes will be attracted by the existance 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.

An implementation following the procedures in this document should not create BM loops, since the AR-REPLICATOR will always forward the BM traffic using the correct tunnel IP Destination Address that indicates the remote nodes how to forward the traffic. This is true in both, the Non-Selective and Selective modes defined in this document.

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 provides a fall back solution to Non-Selective mode in case the AR-REPLICATORs advertised an inconsistent AR Replication mode.

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.

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-10 (work in progress), September 2021.

14.2. Informative References

[RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <https://www.rfc-editor.org/info/rfc8365>.

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