BESS Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan, Ed.
S. Sathappan
W. Henderickx
Nokia

R. Shekhar
N. Sheth
W. Lin
M. Katiyar
Juniper

A. Sajassi
Cisco

A. Isaac
Juniper

M. Tufail
Citibank

Expires: September 1, 2016

February 29, 2016

Optimized Ingress Replication solution for EVPN
draft-ietf-bess-evpn-optimized-ir-00

Abstract

Network Virtualization Overlay (NVO) networks using EVPN as control plane may use ingress replication (IR) or PIM-based trees to convey the overlay multicast 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.

material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at
http://www.ietf.org/shadow.html

This Internet-Draft will expire on September 1, 2016.

Table of Contents

1. Problem Statement

EVPN may be used as the control plane for a Network Virtualization Overlay (NVO) network. Network Virtualization Edge (NVE) devices and PEs that are part of the same EVI use Ingress Replication (IR) or PIM-based trees to transport the tenant's multicast traffic. In NVO networks where PIM-based trees cannot be used, IR is the only alternative. 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 (Broadcast, Unknown unicast and Multicast 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 and proxy-ND capabilities supported by EVPN on the NVEs. Some NVEs can even provide 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 learnt 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, 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:

i) Assisted-Replication (AR)
ii) 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 [EVPN] that are described in section 3.

Section 2 lists the requirements of the combined optimized-IR solution, whereas sections 4 and 5 describe the Assisted-Replication (AR) solution, and section 6 the Pruned-Flood-Lists (PFL) solution.

2. Solution requirements

The IR optimization solution (optimized-IR hereafter) MUST meet the following requirements:

a) The solution MUST provide an IR optimization for BM (Broadcast and Multicast) traffic, while preserving the packet order for unicast applications, i.e. known and unknown unicast traffic SHALL follow the same path.

b) The solution MUST be compatible with [EVPN] and [EVPN-OVERLAY] and not have any impact on the EVPN procedures for BM traffic. In particular, the solution MUST support 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 [EVPN].

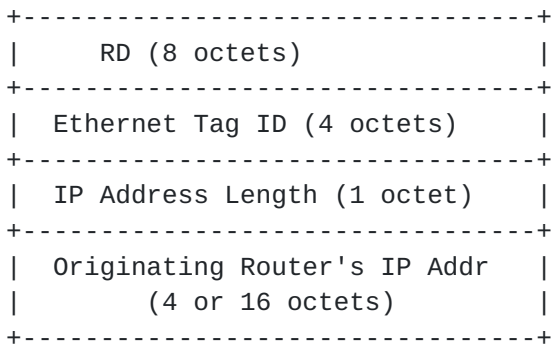
c) The solution MUST be backwards compatible with existing NVEs using a non-optimized version of IR. A given EVI can have NVEs/PEs supporting regular-IR and optimized-IR.

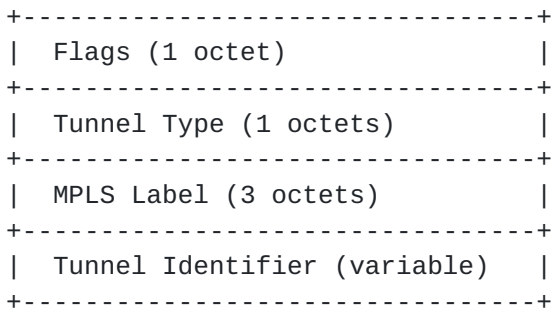
d) The solution MUST be independent of the NVO specific data plane encapsulation and the virtual identifiers being used, e.g.: VXLAN VNIs, NVGRE VSIDs or MPLS labels.

3. EVPN BGP Attributes for optimized-IR

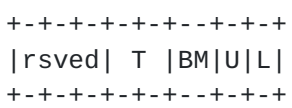
This solution proposes some changes to the [EVPN] Inclusive Multicast Ethernet Tag routes and attributes so that an NVE/PE can signal its optimized-IR capabilities.

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





The Flags field is defined as follows:



Where a new type field (for AR) and two new flags (for PFL signaling) are defined:

- T is the AR 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

- The PFL (Pruned-Flood-Lists) flags defined the desired behavior of the advertising router for the different types of traffic:

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

  + U= Unknown 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 10 for the IANA considerations related to the PTA flags.

In this document, the above RT-3 and PTA can be used in three different modes for the same EVI/Ethernet Tag:

o Regular-IR route: in this route, Originating Router's IP Address, Tunnel Type (0x06), MPLS Label, Tunnel Identifier and Flags MUST be used as described in [EVPN]. The Originating Router's IP Address and Tunnel Identifier are set to an IP address that we denominate IR-IP in this document.

o 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 as well as the Tunnel Identifier are set to the same routable IP address that we denominate AR-IP and SHOULD be different than the IR-IP for a given PE/NVE.

  + Tunnel Type = Assisted-Replication (AR). 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).

- o Leaf-AR route: this route MAY be used by the AR-LEAF to advertise its desire to receive the multicast 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 IR-IP (same IP used by the AR-LEAF in regular-IR routes).

  + Tunnel Identifier is set to the AR-IP of the AR-REPLICATOR from which the multicast traffic is requested.

  + Tunnel Type = Assisted-Replication (AR). Section 11 provides the allocated type value.

  + T (AR role type) = 02 (AR-LEAF).

Each AR-enabled node MUST understand and process the AR type field in the PTA (Flags field) of replicator-AR and leaf-AR routes, and MUST signal the corresponding type (1 or 2) according to its administrative choice for replicator-AR and leaf-AR routes.

Each node, part of the EVI, 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).

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

The following figure illustrates an example NVO network where the non-selective AR function is enabled. Three different roles are defined for a given EVI: 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 multicast traffic to 'all' the NVE/PEs in the EVI except for the source NVE/PE.

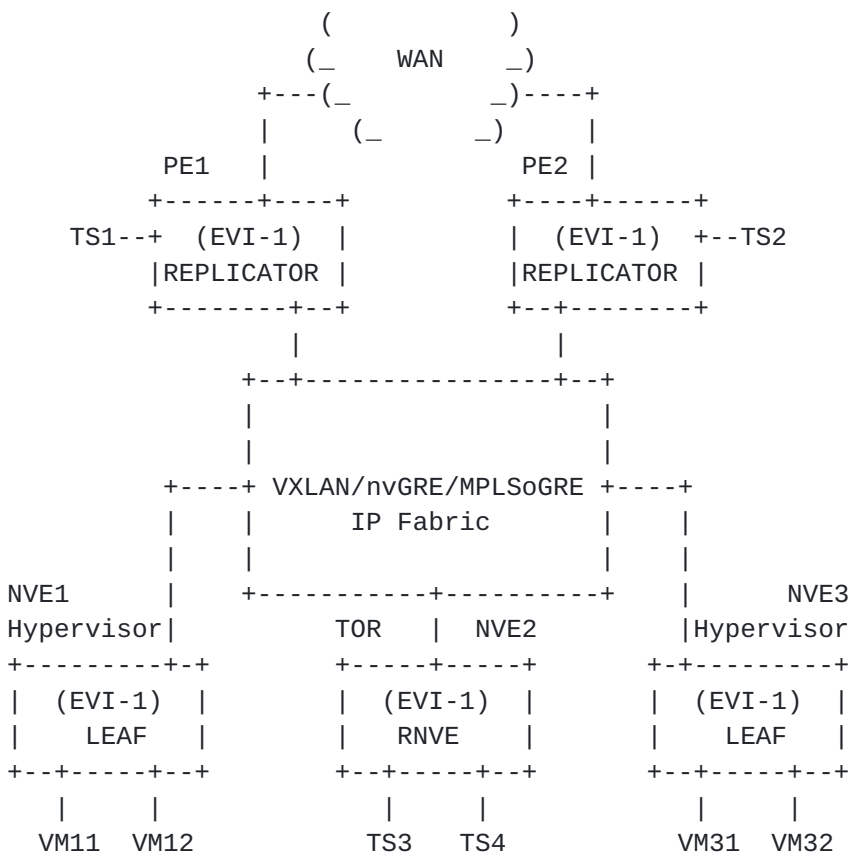


Figure 1 Optimized-IR scenario

4.1. 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 EVI 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 EVI. This administrative option to enable AR-REPLICATOR capabilities MAY be implemented as a system level option as opposed to as a per-EVI 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).

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

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

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

- o 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 will be the AR-IP of the AR-REPLICATOR.

4.2. 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 NVE2 (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 EVI. This administrative option to enable AR-LEAF capabilities MAY be implemented as a system level option as opposed to as per-EVI option.

b) In this non-selective AR solution, the AR-LEAF MUST advertise a single Regular-IR inclusive multicast route as in [EVPN]. 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 the AR-LEAF, it is RECOMMENDED the use of this flag for an easy operation and troubleshooting of the EVI.

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 (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 [EVPN].

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

- o A single AR-REPLICATOR MAY be selected for all the BM packets received on the AR-LEAF attachment circuits (ACs) for a given EVI. This selection is a local decision and it does not have to match other AR-LEAF's selection within the same EVI.
- o An AR-LEAF MAY select more than one AR-REPLICATOR and do either per-flow or per-EVI load balancing.
- o In case of a failure on the selected AR-REPLICATOR, another AR-REPLICATOR will be selected.
- o 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 = TBD (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) 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. This will give the AR-REPLICATOR some time to program the AR-LEF nodes before the AR-LEAF sends BM traffic.

4.3. 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 [EVPN]. The RNVE does not signal any AR role and is unaware of the AR-REPLICATOR/LEAF roles in the EVI. The RNVE will ignore the Flags in the Regular-IR routes and will ignore the Replicator-AR and Leaf-AR routes entirely (due to an unknown tunnel type in the PTA).

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

4.4. Forwarding behavior in non-selective AR EVIs

In AR EVIs, BM (Broadcast and Multicast) traffic between two NVEs may follow a different path than unicast traffic. This solution proposes 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. Section 4.4.1. describes the expected forwarding behavior for BM traffic in nodes acting as AR-REPLICATOR, AR-LEAF and RNVE. Section 4.4.2. describes the forwarding behavior for unknown unicast traffic.

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

4.4.1. Broadcast and Multicast forwarding behavior

The expected behavior per role is described in this section.

4.4.1.1. Non-selective AR-REPLICATOR BM forwarding

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

- 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:

  - 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. If the overlay encapsulation is MPLS

and the EVI label is not the bottom of the stack, the AR-REPLICATOR MUST copy the rest of the labels and forward them to the egress overlay tunnels.

- 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 [EVPN].

4.4.1.2. Non-selective AR-LEAF BM forwarding

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 EVI. The selection of more than one AR-REPLICATOR is described in section 4.2. 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 [EVPN].

4.4.1.3. RNVE BM forwarding

The RNVE is completely unaware of the AR-REPLICATORs, AR-LEAF nodes and BM/U flags (that information is ignored). Its forwarding behavior is the regular IR behavior described in [EVPN]. Any regular non-AR node is fully compatible with the RNVE role described in this document.

4.4.2. Unknown unicast forwarding behavior

The expected behavior is described in this section.

4.4.2.1. Non-selective AR-REPLICATOR/LEAF Unknown unicast forwarding

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 a flood-list composed of ACs and overlay tunnels to the IR-IP Addresses of the remote nodes in the EVI. 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 EVI.

- 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.

- 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 [EVPN].

4.4.2.2. RNVE Unknown unicast forwarding

As described for BM traffic, the RNVE is completely unaware of the REPLICATORs, LEAF nodes and BM/U flags (that information is ignored). Its forwarding behavior is the regular IR behavior described in [EVPN], also for Unknown unicast traffic. Any regular non-AR node is fully compatible with the RNVE role described in this document.

5. 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 EVI-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 slightly 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.

5.1. 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 EVI, as the AR role itself. This administrative option MAY be implemented as a system level option as opposed to as a per-EVI option.

b) Each AR-REPLICATOR will build a list of AR-REPLICATOR, AR-LEAF and RNVE nodes (AR-LEAF nodes that sent only a regular-IR route are accounted as RNVEs by the AR-REPLICATOR). 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 EVI context), the rest of the AR-REPLICATORs will fall back to non-selective AR mode.

b) The Selective AR-REPLICATOR MUST follow the procedures described in section 4.1, 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.

- 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-AR route with PE1's AR-IP (as Tunnel Identifier) and NVE3 advertises a Leaf-AR route with PE2's AR-IP, PE1 MUST only add NVE1/NVE2 in its selective AR-LEAF-set for EVI-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, 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 the 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.

5.2. Selective AR-LEAF procedures

A Selective AR-LEAF chooses a single Selective AR-REPLICATOR per EVI and:

- Sends all the EVI BM traffic to that AR-REPLICATOR and
- Expects to receive the BM traffic for a given EVI from the same AR-REPLICATOR.

In the example of Figure 1, we consider that 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 EVI-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 EVI. This administrative option to enable AR-LEAF capabilities MAY be implemented as a system level option as opposed to as per-EVI option.

b) The AR-LEAF MAY advertise a Regular-IR route if there are RNVEs or non-selective AR-LEAFs in the EVI. The Selective AR-LEAF MUST advertise a Leaf-AR route after receiving a Replicator-AR route with L=1. It is recommended that the Selective AR-LEAF waits for a timer t before sending the Leaf-AR route, so that the AR-LEAF receives all the Replicator-AR routes for the EVI.

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 EVI. Once selected:

   - The Selective AR-LEAF will send a Leaf-AR route including the AR-IP of the selected AR-REPLICATOR.

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

- In case of a failure on the selected AR-REPLICATOR, another AR-REPLICATOR will be selected and a new Leaf-AR update will be issued, including the new AR-IP. 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 t to speed up the convergence. When the timer expires, the Selective AR-LEAF will resume its AR mode with the new Selective AR-REPLICATOR.

5.3. Forwarding behavior in selective AR EVIs

This section describes the differences of the selective AR forwarding mode compared to the non-selective mode. Compared to section 4.4, there are no changes for the forwarding behavior in RNVEs or for unknown unicast traffic.

5.3.1. Selective AR-REPLICATOR BM forwarding

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 EVI, 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 EVI.

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-AR route with the AR-IP of the local AR-REPLICATOR. This set is updated with every Leaf-AR route received with a change in the AR-IP included in the PTA's Tunnel Identifier.

   - 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:

- 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 EVI 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.

- 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.

- 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 [EVPN].

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 overlay encapsulation is MPLS and the EVI 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.

5.3.2. Selective AR-LEAF BM forwarding

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 [EVPN].

6. Pruned-Flood-Lists (PFL)

In addition to AR, the second optimization supported by this solution is the ability for the all the EVI 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= Broadcast and Multicast (BM) flag. BM=1 means "prune-me" from the BM flood-list. BM=0 means regular behavior.

+ U= 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 EVI 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.

6.1. A PFL example

In order to illustrate the use of the solution described in this document, we will assume that EVI-1 in figure 1 is optimized-IR enabled and:

o 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.

o NVE1 and NVE3 are administratively configured as AR-LEAF nodes, due to their low-performance software-based replication capabilities. They will advertise a Leaf-AR route. 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 EVI-1.

o NVE2 is optimized-IR unaware; therefore it takes on the RNVE role in EVI-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.

7. AR Procedures for single-IP AR-REPLICATORS

The procedures explained in sections 4 (Non-selective AR) and 5 (Selective AR) assume that the AR-REPLICATOR can use two local routable IP addresses to terminate and initiate 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 initiate 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 4 and 5 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 for both routes.

- 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 described in sections 4 and 5. 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 4 and 5.

8. AR Procedures and EVPN Multi-homing Split-Horizon

When EVPN is used for MPLS over GRE, all the multi-homing procedures are compatible with sections 4 and 5 of this document.

If VXLAN or NVGRE are used, and if the Split-horizon is based on the tunnel IP SA and "Local-Bias" as described in [EVPN-OVERLAY], 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 AR-REPLICATOR. This will allow DF (Designated Forwarder) AR-LEAF nodes to apply Split-horizon check procedures for BM packets, before sending them to the local Ethernet-Segment.

Note that if the AR-REPLICATOR implementation keeps the received tunnel IP SA, the use of uRPF in the IP fabric based on the tunnel IP SA MUST be disabled.

9. Out-of-band distribution of Broadcast/Multicast traffic

The use of out-of-band mechanisms to distribute BM traffic between AR-REPLICATORS MAY be used. Details will be provided in future versions of this document.

10. Benefits of the optimized-IR solution

A solution for the optimization of Ingress Replication in EVPN is described in this document (optimized-IR). The solution brings the following benefits:

- Optimizes the multicast forwarding in low-performance NVEs, by relaying the replication to high-performance NVEs (AR-REPLICATORs) and while preserving the packet ordering for unicast applications.

- Reduces the flooded traffic in NVO networks where some NVEs do not need broadcast/multicast and/or unknown unicast traffic.

- It is fully compatible with existing EVPN implementations and EVPN functions for NVO overlay tunnels. Optimized-IR NVEs and regular NVEs can be even part of the same EVI.

- It does not require any PIM-based tree in the NVO core of the network.

11. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

12. Security Considerations

This section will be added in future versions.

13. IANA Considerations

A new Tunnel-Type (AR) must be requested and allocated by IANA for the PTA (PMSI Tunnel Attribute) used in this document.

In addition to the new Tunnel-Type, this document requests the allocation of the PTA flags as in section 3. A registry is created as per [PTA-FLAGS].

14. Terminology

Regular-IR: Refers to Regular Ingress Replication, where the source NVE/PE sends a copy to each remote NVE/PE part of the EVI.

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 [EVPN].

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-LEF, 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 selective 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 [EVPN]. It means sending an AC BM packet copy to each remote PE/NVE in the EVI 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

15. References

15.1 Normative References

[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, <http://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, <http://www.rfc-editor.org/info/rfc7432>.

15.2 Informative References

[EVPN-OVERLAY] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", draft-ietf-bess-evpn-overlay-02.txt, work in progress, October 2015

[PTA-FLAGS] Rosen, E., "IANA Registry for P-Multicast Service Interface Tunnel Attribute Flags", draft-ietf-bess-pta-flags-01.txt, work in progress, August 2015

16. Acknowledgments

The authors would like to thank Neil Hart, David Motz, Kiran Nagaraj, Dai Truong, Thomas Morin and Jeffrey Zhang for their valuable feedback and contributions.

17. Authors' Addresses

Jorge Rabadan (Editor)
Nokia
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@nokia.com

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

Mukul Katiyar
Juniper Networks
Email: mkatiyar@juniper.net

Wim Henderickx
Nokia
Email: wim.henderickx@nokia.com

Ravi Shekhar
Juniper Networks
Email: rshekhar@juniper.net

Nischal Sheth
Juniper Networks
Email: nsheth@juniper.net

Wen Lin
Juniper Networks
Email: wlin@juniper.net

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Aldrin Isaac
Juniper
Email: aisaac@juniper.net

Mudassir Tufail
Citibank
mudassir.tufail@citi.com