L2VPN Workgroup
Internet Draft
Intended status: Standards Track

J. Rabadan
S. Sathappan
W. Henderickx
Alcatel-Lucent

R. Shekhar
N. Sheth
W. Lin
Juniper

M. Katiyar
Nuage Networks

Expires: January 5, 2015

July 4, 2014

Optimized Ingress Replication solution for EVPN
draft-rabadan-l2vpn-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.

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 January 5, 2015.

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 large amounts 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 section 4 describes the Assisted-Replication (AR) solution and section 5 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:

- o All-active multi-homing, including the split-horizon and Designated Forwarder (DF) functions.
- o Single-active multi-homing, including the DF function.
- o 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 routes and attributes so that an NVE/PE can signal its optimized-IR capabilities.

The Inclusive Multicast Ethernet Tag route and its PMSI Tunnel attribute's format used in EVPN are shown below:

```
+---------------------------------+
|     RD (8 octets)               |
+---------------------------------+
|  Ethernet Tag ID (4 octets)     |
+---------------------------------+
|  IP Address Length (1 octet)    |
+---------------------------------+
|  Originating Router's IP Addr   |
|        (4 or 16 octets)         |
+---------------------------------+


+---------------------------------+
|  Flags (1 octet)                |
+---------------------------------+
|  Tunnel Type (1 octets)         |
+---------------------------------+
|  MPLS Label (3 octets)          |
+---------------------------------+
|  Tunnel Identifier (variable)   |
+---------------------------------+
```

Where:

- o Originating Router's IP Address, Tunnel Type (0x06), MPLS Label and Tunnel Identifier MUST be used as described in [EVPN] for non-optimized-IR behavior.

- o A different Originating Router's IP Address, a new Tunnel Type (TBD), MPLS Label and Tunnel Identifier may be used for Assisted-Replication (AR).

- o The Flags field is defined as follows:

```
 0 1 2 3 4 5  6 7
+-+-+-+-+-+--+-+-+
|rsved| T |BM|U|L|
+-+-+-+-+-+--+-+-+
```

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

  - T is the AR Type field (2 bits):

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

    + 01 (decimal 1) = AR REPLICATOR

+ 10 (decimal 2) = AR LEAF

- New PFL (Pruned-Flood-Lists) flags:

  + 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 has no use in this solution.

Each AR-enabled EVI node MUST understand and process the AR type field in the PMSI attribute (Flags field) and MUST signal the corresponding type (1 or 2) according to its administrative choice.

Each EVI node 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, regardless of the AR settings.

The T field and BM/U flags MAY be used individually or together, i.e. a given PMSI attribute may only convey the AR type information, or only the BM/U flags, or both pieces of information at the same time.

Non-optimized-IR nodes will be unaware of the new PMSI attribute flag definition, i.e. they will ignore the information contained in the flags field.

4. Assisted-Replication (AR) Solution Description

The following figure illustrates an example NVO network where the AR function is enabled. This scenario will be used to describe the solution throughout the rest of the document.

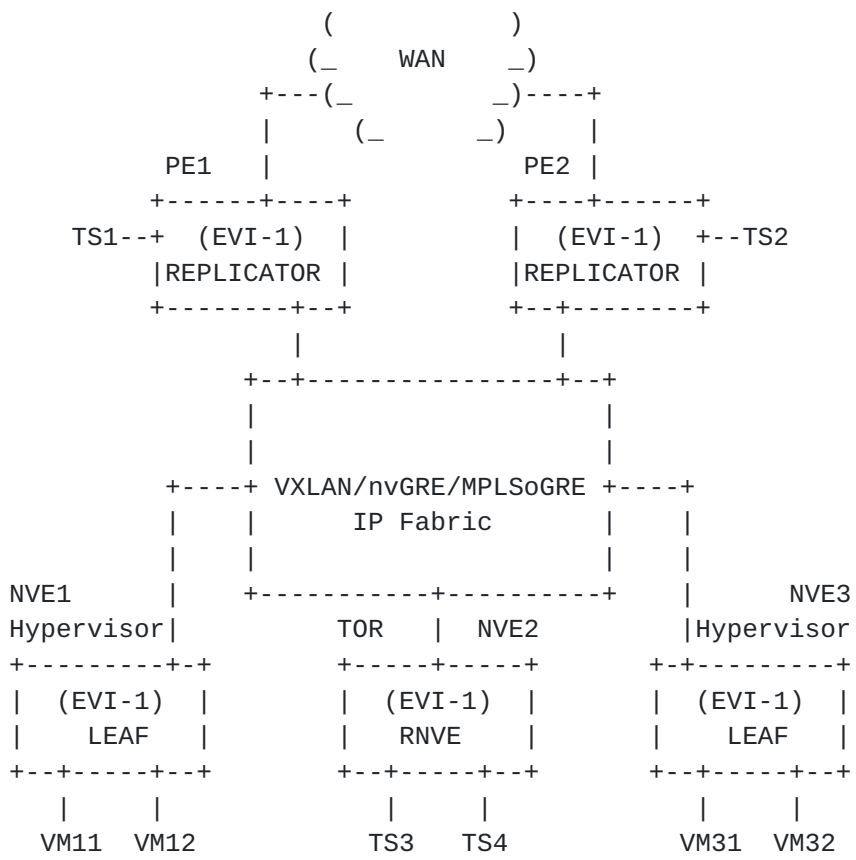

Figure 1 Optimized-IR scenario

4.1. AR roles and control plane

The solution defines three different roles in an AR EVI service:

a) AR-REPLICATOR (REPLICATOR)
b) AR-LEAF (LEAF)
c) Regular NVE (RNVE)

4.1.1. AR-REPLICATOR procedures

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 REPLICATOR signals its REPLICATOR role in the control plane and understands where the other roles (LEAF nodes, RNVEs and other REPLICATORs) are located. A given AR EVI service may have zero, one or more REPLICATORs. In our example in figure 1, PE1 and PE2 are defined as REPLICATORs. The following considerations apply to the REPLICATOR role:

a) The AR-REPLICATOR role SHOULD be an administrative choice in any

NVE/PE that is part of an AR EVI. This administrative option to enable REPLICATOR capabilities MAY be implemented as a system level option as opposed to as per-EVI option.

b) An AR-REPLICATOR MUST advertise an AR inclusive multicast route and MAY advertise an IR inclusive multicast route.

c) An IR Inclusive Multicast Route is an Inclusive Multicast Route as defined in [EVPN] and MUST NOT be generated by the AR REPLICATOR if it does not have local attachment circuits (AC).

d) An AR Inclusive Multicast Route MUST be generated by the AR REPLICATOR and it is comprised of:

- AR Originating Router's IP Address, which is different from the IR IP address used in the IR Inclusive Multicast Route.
- T = 1 (AR REPLICATOR)
- Tunnel type = TBD (AR tunnel)
- Tunnel Identifier MUST contain the same value as the AR Originating Router's IP Address.
- The rest of the route fields are used as per [EVPN].

e) When a node defined as REPLICATOR receives a packet from an overlay tunnel, it will do a tunnel destination IP lookup and follow the following procedures:

- If the destination IP is the IR Originating Router's IP Address the node will process the packet normally as in [EVPN].
- If the destination IP is the AR Originating Router's IP Address, the node MUST replicate the packet to local ACs and overlay tunnels (excluding the overlay tunnel to the source of the packet). Selective replication to only interested AR-LEAF nodes will be added in a future revision of this document.

4.1.2. AR-LEAF procedures

LEAF is defined as an NVE/PE that - given its poor replication performance - sends all the BM traffic to a REPLICATOR that can replicate the traffic further on its behalf. It signals its LEAF capability in the control plane and understands where the other roles are located (REPLICATOR and RNVEs). A given service can have zero, one or more LEAF nodes. Figure 1 shows NVE1 and NVE2 (both residing

in hypervisors) acting as LEAF. The following considerations apply to the LEAF role:

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

b) An AR-LEAF MUST advertise a single inclusive multicast route where the AR type is set to T = 2 (AR LEAF) and the rest of fields follow [EVPN].

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

d) In a service where there is more than one or more REPLICATORs, the LEAF can locally select which REPLICATOR it sends the BM traffic to:

   - A single REPLICATOR may be selected for all the BM packets received on LEAF attachment circuits (ACs). This selection is a local decision and it does not have to match other LEAF's selection within the same service.

   - A LEAF may select more than one REPLICATOR and do either per-flow or per-service load balancing.

   - In case of a failure on the selected REPLICATOR, another REPLICATOR will be selected.

   - When a REPLICATOR is selected, the LEAF MUST send all the BM packets to that REPLICATOR using the forwarding information given by the AR Inclusive Multicast Route previously sent by the REPLICATOR, with tunnel type = TBD (AR tunnel). The underlay destination IP address MUST be the AR Originating Router's IP Address signaled by the REPLICATOR for the AR tunnel type.

   - 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 REPLICATORs MUST not replicate these control plane packets to other overlay tunnels since

they will use the regular Originating Router's IP Address.

4.1.3. RNVE procedures

RNVE (Regular Network Virtualization Edge node) is defined as an NVE/PE without REPLICATOR or LEAF capabilities that does IR as described in [EVPN]. The RNVE does not signal any special role and is unaware of the REPLICATOR/LEAF roles in the EVI. The RNVE will ignore AR Inclusive Multicast Routes (due to an unknown tunnel type in the PMSI attribute).

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

4.2. Multi-destination traffic forwarding behavior in 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 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 4.2.1 describes the expected forwarding behavior for BM traffic in nodes acting as REPLICATOR, LEAF and RNVE. Section 4.2.2 describes the forwarding behavior for unknown unicast traffic.

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

4.2.1. Broadcast and Multicast forwarding behavior

The expected behavior per role is described in this section.

4.2.1.1. REPLICATOR BM forwarding

The 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. The REPLICATOR will also build a list of remote REPLICATORs, LEAF nodes and RNVEs for the EVI.

- When a 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 a 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 Originating Router IP, the REPLICATOR will forward the BM packet to its flooding list (ACs and overlay tunnels) excluding the non-BM overlay tunnels. The REPLICATOR will do source squelching to ensure the traffic is not sent back to the originating LEAF. If the overlay encapsulation is MPLS and the EVI label is not the bottom of the stack, the REPLICATOR MUST copy the rest of the labels and forward them to the egress overlay tunnels.

- If the destination IP matches its IR Originating Router IP, the 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].

4.2.1.2. LEAF BM forwarding

The LEAF nodes will build two flood-lists:

1) Flood-list #1 - composed of ACs and a REPLICATOR-set of overlay tunnels. The REPLICATOR-set is defined as one or more overlay tunnels to the AR Originating Router's IP Addresses of the remote REPLICATOR(s) in the EVI. The selection of more than one REPLICATOR is described in section 4.1.2 and it is a local LEAF decision.

2) Flood-list #2 - composed of ACs and overlay tunnels to the remote IR Originating Router's IP Addresses.

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

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

o If the REPLICATOR-set is NOT empty, the LEAF will send the packet to flood-list #1.

When a 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].

4.2.1.3. RNVE BM forwarding

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]. Any regular non-AR node is fully compatible with the RNVE role described in this document.

4.2.2. Unknown unicast forwarding behavior

The expected behavior is described in this section.

4.2.2.1. REPLICATOR/LEAF Unknown unicast forwarding

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

The REPLICATOR/LEAF nodes will build a flood-list composed of ACs and overlay tunnels to the IR Originating Router's 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 a 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 a 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. Pruned-Flood-Lists (PFL)

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. An example use-case

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 REPLICATORs, due to their high-performance replication capabilities. PE1 and PE2 will signal AR type = 1 and BM/U flags = 00.

- o NVE1 and NVE3 are administratively configured as LEAF nodes, due to their low-performance software-based replication capabilities. They will signal AR type = 2. 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.

5. 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 (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.

6. 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.

7. Security Considerations

This section will be added in future versions.

8. IANA Considerations

8. References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-07.txt, work in progress, May, 2014

[EVPN-OVERLAY] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", draft-sd-l2vpn-evpn-overlay-02.txt, work in progress, October, 2013

9. Acknowledgments

The authors would like to thank Neil Hart and David Motz for their valuable feedback and contributions.

10. Authors' Addresses

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA 94043 USA
Email: jorge.rabadan@alcatel-lucent.com

Senthil Sathappan
Alcatel-Lucent
Email: senthil.sathappan@alcatel-lucent.com

Mukul Katiyar
Nuage Networks
Email:

Wim Henderickx
Alcatel-Lucent
Email: wim.henderickx@alcatel-lucent.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