           Extended Procedures for EVPN Optimized Ingress Replication
              draft-wsv-bess-extended-evpn-optimized-ir-01

Abstract

   [EVPN-AR] specifies an optimized ingress replication solution for
   more efficient multicast and broadcast delivery in a Network
   Virtualization Overlay (NVO) network for EVPN.

   This document extends the optimized ingress replication procedures
   specified in [EVPN-AR] to overcome the limitation that an AR-
   REPLICATOR may have. An AR-REPLICATOR may be unable to retain the
   source IP address or include the expected ESI label that is required
   for EVPN split horizon filtering when replicating the packet on
   behalf of its multihomed AR-LEAF. Under this circumstance, the
   extended procedures specified in this document allows the support of
   EVPN multihoming on the AR-LEAFs as well as optimized ingress
   replication for the rest of the EVPN overlay network.

Requirements Language

   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.

   This Internet-Draft will expire on September 7, 2019.

Copyright Notice

Table of Contents

1. Terminology

AR-IP Tunnel

An overlay tunnel with a destination IP address of AR-IP that an AR-REPLICATOR advertises in its REPLICATE-AR route.

This document heavily uses the terminology specified in [EVPN-AR].
It also uses the terminology specified in [RFC7432] and [RFC8365].

2. Introduction

2.1. Background

2.1.1. EVPN Multihoming and Split Horizon Filtering Rule

This section gives a brief overview of the existing split horizon
filtering rules used for EVPN multihoming.

[RFC7432] defines the split-horizon filtering rule based on ESI label
for EVPN multihoming with MPLS encapsulation, and this filtering rule
also applies for EVPN with IP-based encapsulation for MPLS, such as
MPLS over GRE or MPLS over UDP. [RFC8365] defines the split horizon
filtering rule based on "Local-Bias" for EVPN multihoming with VXLAN
encapsulation.

When EVPN is used in an NVO network, a Tenant System (TS) may connect
to a set of Network Virtualization Edge (NVE) devices through a
multihomed Ethernet segment (ES). The split-horizon filtering rule
for EVPN all-active multihoming ensures that a Broadcast, Unknown
unicast or Multicast (BUM) packet received from an ES that is a part
of a multihomed ES is not looped back to the multihomed TS through an
egress NVE connected to the same multihomed ES. For EVPN with VXLAN
encapsulation, the split-horizon filtering rule is based on the
egress NVE examining the source IP address of the BUM packet received
from an overlay tunnel. The egress PE identifies the ingress NVE
through the source IP address. The egress NVE does not forward the
BUM packet received from an overlay tunnel to the multihomed Ethernet
segment that it has in common with the ingress NVE.

For EVPN with MPLS over IP tunnel, the split-horizon filtering rule
is based on the ESI label. For ingress replication, an ESI label is
downstream assigned per multihomed ES. The ingress NVE MUST include
the ESI label, assigned by the egress PE, when it forwards a BUM

packet to the egress NVE if the BUM traffic is form the AC that is
part of the multihomed ES associated with that ESI label. The egress
NVE does not forward the BUM packet it received from an overlay
tunnel to the multihomed ES if the ESI label is allocated by the
egress NVE for that multihomed ES.

2.2. Optimized-IR and the Need to Maintain the Original Source IP address or Include the ESI Label

[EVPN-AR] specifies an optimized ingress replication procedures for
the delivery of Multicast and Broadcast (BM) traffic within a bridge
domain. It defines the control plane and forwarding plane procedures
for AR-REPLICATOR, AR-LEAF and RNVE. To support EVPN AR-LEAF
multihoming, [EVPN-AR] recommends that split horizon filtering rule
based on "Local-Bias" procedures is used for EVPN NVO network using
either 24-bit VNI or MPLS label.

To support EVPN all-active multihoming based on "Local-Bias"
procedures, when an AR-REPLICATOR performs assisted replication on
behalf of a multihomed AR-LEAF, the AR-REPLICATOR shall use the
source IP address of the ingress AR-LEAF for packet received on the
AR-IP tunnel. This ensures that other remote NVEs, when receiving a
packet from its AR-REPLICATOR, can perform the regular split horizon
filtering based on the source IP address.

To support EVPN all-active multihoming with MPLSoGRE or MPLSoUDP,
sometimes it is desirable to continue using the existing split
horizon filtering rule based on [RFC7432] procedures. In this case,
when performing assisted replication on behalf of a multihomed AR-
LEAF, an AR-REPLICATOR shall include the ESI label advertised by a
remote NVE for that multihomed ES.

Due to either implementation complexity or hardware limitation, an
AR-REPLICATOR may be unable to retain the source IP address or
include the ESI label when replicating the packet to the remote NVEs
on behalf of a multihomed AR-LEAF. Under this circumstance, when
receiving the packet, a remote NVE is unable to use the existing
split horizon filtering rules to prevent the looping of BM traffic
required for all-active multihoming.

For example, with VXLAN encapsulation, consider a case where TS1 is
multihomed to AR-LEAF1 and AR-LEAF2 through a multihomed ES. When
AR-LEAF1 receives an IP multicast packet from TS1, AR-LEAF1 sends the
packet to its AR-REPLICATOR with the source IP address set to AR-
LEAF1's IR-IP and the destination IP address set to the AR-IP of the
AR-REPLICATOR. Since the AR-REPLICATOR is unable to retain the
source IP address for the packet it received on the AR-IP tunnel, the
AR-REPLICATOR uses one of its own IP addresses as the source IP

address when it replicates the packet to other NVEs. When AR-LEAF2
receives the packet from the AR-REPLICATOR, it checks for the source
IP address. AR-LEAF2 is unable to detect that this packet was
originally sent by AR-LEAF1. If AR-LEAF2 is the DF for the
multihomed ES connected to TS1, AR-LEAF2 forwards the packet to TS1.
This causes the same IP multicast packet to be looped back to TS1.

The same problem can also happen to EVPN with MPLS over IP network if
an AR-REPLICATOR cannot include the ESI label to the remote NVE for
the multihomed ES when the split horizon filtering rule based on
[RFC7432] is used.

3. Solution

This document extends the procedures defined in the [EVPN-AR] to
support EVPN multihoming on AR-LEAFs when an NVE acts as an AR-
REPLICATOR is incapable of retaining the source IP address or
including an ESI label for its AR-LEAF either due to its hardware
limitation or implementation complexity. The solution specified in
this document is intended to work for EVPN over IP-based network with
NVO tunnel using either 24-bit VNI or MPLS label. The solution
relies on either [RFC7432] or "Local-Bias" split-horizon filtering
rules to prevent the looping of BUM traffic. We refer to the
procedures specified in this document as the extended Optimized-IR
procedures. The extended Optimized-IR procedures also work with
RNVE. The extended Optimized-IR procedures do not apply to EVPN with
MPLS encapsulation.

3.1. AR-REPLICATOR Announcing Multihoming Assistant Capability for
         Optimized-IR

An AR-REPLICATOR announces its AR-REPLICATOR role through the control
plane. A REPLICATOR-AR route, as it is specified in the [EVPN-AR],
is an Inclusive Multicast Ethernet Tag (IMET) route that an AR-
REPLICATOR originates for its AR-IP and corresponding AR-replication
tunnel.

If an AR-REPLICATOR cannot or chose not to retain the source IP
address or include the expected ESI label for its multihomed AR-
LEAFs, it MUST informs other NVEs in the control plane through the
use of EVPN Multicast Flags Extended Community as follow: a) the AR-
REPLICATOR MUST set the "Extended-MH-AR" flag, as it is specified in
the section 6, in the multicast flags extended community, and b) it
MUST attach this community to the REPLICATOR-AR route it originates.
We call such an AR-REPLICATOR an Extended-MH AR-REPLICATOR.

An Extended-MH AR-REPLICATOR supports extended Optimized-IR
procedures defined in this document for its multihomed AR-LEAFs. An

Extended-MH AR-REPLICATOR keeps track of its AR-LEAF's multihomed
peer. An Extended-MH AR-REPLICATOR can perform assisted replication
for an AF-LEAF to other NVEs that are not attached to the same
multihomed ES as the AR-LEAF. An Extended-MH AR-REPLICATOR does not
perform assisted replication for its AR-LEAF to other NVEs that have
a multihomed ES in common with the AR-LEAF. The changes in the
control plane and forwarding plan procedures for an Extended-MH AR-
REPLICATOR is further explained in detail in section 5.2.

An AR-REPLICATOR originating a REPLICATOR-AR route without a
multicast flags extended community or with the Extended-MH-AR flag
unset is considered to be an MH-capable-assistant AR-REPLICATOR. An
MH-capable-assistant AR-REPLICATOR can perform assisted replication
for its single-homed AR-LEAF as well as multihomed AR-LEAF.

3.2. Multihomed AR-LEAF and Extended-MH AR-REPLICATOR

An AR-LEAF follows the control plane and forwarding plane procedures
specified in [EVPN-AR]. In addition, if a multihomed AR-LEAF detects
that one of its AR-REPLICATORs is Extended-MH AR-REPLICATOR based on
the processing of its REPLICATOR-AR route, the multihomed AR-LEAF
follows the extended Optimized-IR procedures specified in this
document. With the extended Optimized-IR procedures, within the same
BD, the multihomed AR-LEAF will use the regular ingress replication
procedure to deliver a copy of a BUM packet received from its local
AC to each of the remote NVEs that has a multihomed ES in common with
it. In this way, the egress NVE can use the regular split horizon
filtering rule defined in [RFC7432] or [RFC8365] to prevent the BUM
traffic to be looped through the egress NVE to the source of origin.
The extended procedures required for an AR-LEAF is further specified
in detail in section 5.

For an AR-LEAF, please note that the additional forwarding procedures
specified above apply to BM packets coming from any of its ACs in the
same BD, whether that AC is a single homed ES or a part of a
multihomed ES. It may also applies to Unknown unicast traffic. This
is to further alleviate the burden of an Extended-MH AR-REPLICATOR as
it may be unable to detect whether a packet received on its AR-IP
tunnel was originally received from a single-homed or multihomed ES.

Consider an EVPN NVO network with a tenant domain consists of a set
of m AR-LEAFs in BD X: AR-LEAF1, AR-LEAF2, AR-LEAF3, ..., AR-LEAFm.
TS1 is multihomed to AR-LEAF1 and AR-LEAF2 in BD X through a
multihomed ES ES1. TS2 is multihomed to AR-LEAF1 and AR-LEAF3 in BD
X through another multihomed ES ES2. Also, suppose that there are
two Extended-MH AR-REPLICATORs in the same tenant domain: AR-
REPLICATOR1 and AR-REPLICATOR2. AR-LEAF1 will detect that its AR-
REPLICATORs are Extended-MH AR-REPLICATORs. AR-LEAF1 will also

detect that both AR-LEAF2 and AR-LEAF3 have a multihomed ES in common
with it. AR-LEAF1 will use regular ingress replication to send the
BUM traffic it receives from its access to both AR-LEAF2 and AR-
LEAF3. AR-LEAF1 will rely on one of its AR-REPLICATORs to send the
BM traffic to AR-LEAF4, AR-LEAF5, ..., and AR-LEAFm.

3.3. The Benefit of the Extended Optimized-IR Procedure

The extended Optimized-IR procedures specified in this document
greatly reduces the implementation complexity of an AR-REPLICATOR or
helps to overcome the limitation of an AR-REPLICATOR. It frees all
AR-REPLICATORs from performing multihoming assisted replication while
at the same time, it allows the support of EVPN multihoming on the
AR-LEAFs with the existing multihoming procedures and split horizon
filtering rules. For EVPN with MPLS over IP-based encapsulation, an
NVE can continue to use the split horizon filtering rule based on the
ESI label. Furthermore, it still allows the support of efficient
Optimized-IR for the rest of an EVPN NVO network.

For example, in a typical NVO network, a TS is most likely multihomed
to two or a small set of NVEs for redundancy. In an NVO network
consisting of many NVEs, the AR-REPLICATOR is still responsible for
replicating the BM packet to the most of NVEs for its AR-LEAF and
thus it inherits the benefit of optimized ingress replication for the
most of its NVO network.

3.4. Support for Mixed AR-REPLICATORs

When there are mixed MH-capable-assistant AR-REPLICATORs and
Extended-MH AR-REPLICATORs in the same tenant domain, all AR capable
NVEs MUST follow the extended Optimized-IR procedures as long as one
of the AR-REPLICATORs is an Extended-MH AR-REPLICATOR.

When there are mixed AR-REPLICATORs, this document recommends that
all MH-capable-assistant AR-REPLICATORs to be administratively
provisioned to behave as Extended-MH AR-REPLICATORs. In this case,
each AR-REPLICATOR originates its REPLICATOR-AR route with the
Extended-MH-AR flag set in the multicast flags extended community.

The procedure for using mixed AR-REPLICATORs is beyond the scope of
this document.

4. Extended Optimized-IR Procedure for Supporting Extended-MH AR-REPLICATOR

4.1. **AR-LEAF Procedure**

   This section covers the extended Optimized-IR procedures required for
   an AR-LEAF in further detail when at least one of the AR-REPLICATORs
   is an Extended-MH AR-REPLICATOR. It is assumed that an AR-LEAF
   follows the procedures defined in [EVPN-AR] unless it is specified
   otherwise.

4.1.1. **Control Plane Procedure for AR-LEAF**

   An AR-LEAF detects whether an AR-REPLICATOR is capable of performing
   multihoming assisted replication through the Extended-MH-AR flag in
   the multicast flags extended community carried in the REPLICATOR-AR
   route. An AR-REPLICATOR originating a REPLICATOR-AR route without a
   multicast flags extended community or with the Extended-MH-AR flag
   unset is considered to be multihoming assistant capable.

   If an AR-LEAF does not have any locally attached segment that is a
   part of a multihomed ES, then there is no additional extended
   Optimized-IR procedure for an AR-LEAF to follow and we can go
   directly to section 4.2.

   If selective assistant-replication is used for the EVI, selective AR-
   LEAFs that share the same multihomed ES MUST select the same primary
   AR-REPLICATOR and the same backup AR-REPLICATOR, if there is one.
   This can be achieved through either manual configuration on each
   multihomed selective AR-LEAF or by other methods that are beyond the
   scope of this document. Each selective AR-LEAF follows the
   procedures defined in the [EVPN-AR] to send its corresponding leaf-AD
   routes to its AR-REPLICATOR.

   An AR-LEAF follows the normal procedures defined in [RFC7432] when it
   originates a type-4 ES route and type-1 Ethernet A-D routes for its
   locally attached segment that is a part of a multihomed ES.

   In addition, an AR-LEAF builds a peer-multihomed-flood-list for each
   BD it attaches. Per normal EVPN procedures defined in [RFC7432], an
   AR-LEAF discovers the ESI of each multihomed ES that every remote NVE
   connects to. For a given BD, an AR-LEAF constructs a peer-
   multihomed-flood-list that consists of its peer multihomed NVEs in
   that BD that have at least one multihomed ES in common with it. An
   AR-LEAF may consider a common multihomed ES that it shares with a
   remote NVE in a BD specific scope or an EVI scope. Please section 5
   for detail.

[4.1.2](). Forwarding Procedure for AR-LEAF

Suppose that a multihomed AR-LEAF detects through the control plane procedure that at least one of its AR-REPLICATORs is an Extended-MH AR-REPLICATOR, then in addition to follow the forwarding procedures defined in [[EVPN-AR]()], the AR-LEAF will use regular ingress replication to send the BUM packet, received from one of its ACs, to each NVE in that BD's peer-multihomed-flood-list.

In the case that there are no more AR-REPLICATORs in the tenant domain, the AR-LEAF reverts back to the regular IR behavior as it is defined in [[RFC7432]()].

An AR-LEAF will follow the regular EVPN procedures when it receives a packet from an overlay tunnel and it will never send the packet back to the core.

[4.2](). AR-REPLICATOR Procedure

This section describes the additional procedures for an AR-REPLICATOR when there is at least one AR-REPLICATOR in the same tenant domain that is an Extended-MH AR-REPLICATOR.

It is also assumed that an AR-REPLICATOR follows the procedures defined in [[EVPN-AR]()] unless specified otherwise.

[4.2.1](). Control Plane Procedure for AR-REPLICATOR

An NVE that performs an AR-REPLICATOR role follows the control plane procedures for AR-REPLICATOR defined in the [[EVPN-AR]()].

In addition, if an AR-REPLICATOR is an Extended-MH AR-REPLICATOR or if it is administratively provisioned to behave as an Extended-MH AR-REPLICATOR, it SHALL attach a multicast flags extended community to its REPLICATOR-AR route with the Extended-MH-AR flag set.

An AR-REPLICATOR also discovers whether another AR-REPLICATOR is an Extended-MH AR-REPLICATOR based on the multicast flags extended community. If at least one AR-REPLICATOR is an Extended-MH AR replicator, then the rest of AR-REPLICATORs SHALL fall back to support the extended procedures specified in this document.

When there are mixed AR-REPLICATORs, this document recommends that all MH-capable-assistant AR-REPLICATORs SHOULD fall back to behave as Extended-MH AR-REPLICATOTRs through administrative provisioning.

An Extended-MH AR-REPLICATOR builds a multihomed list for each BD that its AR-LEAF attaches to. We refer to such a multihomed list as

an AR-LEAF's multihomed-list. Per normal EVPN procedures defined in
[RFC7432], an AR-REPLICATOR imports the Ethernet A-D per EVI route,
the alias route, originated by each remote NVE in the same tenant
domain. For a given BD that an AR-LEAF belongs to, an AR-LEAF's
multihomed-list consists of all the NVEs in that BD that have at
least one multihomed ES in common with the said AR-LEAF. Please also
refer to section 5 for the common multihomed ES an AR-LEAF shares
with its remote NVE.

Consider an EVPN NVO network specified in the section 3.2. Both AR-
LEAF1 and AR-LEAF2 originate its Ethernet A-D per EVI route for ES1
respectively. Both AR-LEAF1 and AR-LEAF3 originate its Ethernet A-D
per EVI route for ES2 respectively. Per normal EVPN procedures, each
AR-REPLICATOR imports and processes Ethernet A-D per EVI routes.
Each AR-REPLICATOR builds an AR-LEAF1's multihomed-list for BD X that
consists of AR-LEAF2 and AR-LEAF3. Each AR-REPLICATOR also builds
AR-LEAF's multihomed-lists for other AR-LEAFs.

4.2.2. Forwarding Procedure for AR-REPLICATOR

When an AR-REPLICTOR determines that it is an Extended-MH AR-
REPLICATOR or determines that it SHALL fall back to become an
Extended-MH AR_REPLICATOR, it MUST follow the forwarding procedures
described in this section.

For a given BD, when an AR-REPLICATOR replicates the packet, received
from its AR-IP tunnel, to other overlay tunnels on behalf of its
ingress AR-LEAF, the AR-REPLICATOR MUST skip any NVE that is in that
ingress AR-LEAF's multihomed-list built for that said BD.

When replicating the traffic to other AR-REPLICATORs or other AR-
LEAFs over an overlay tunnel, an AR-REPLICATOR does not set the
source IP address to its ingress AR-LEAF's IR-IP. It is assumed
under the scope of this document that an AR-LEAF does not share any
common multihoming ES with any AR-REPLICATOR.

When replicating the traffic to other RNVEs, an AR-REPLICATOR should
set the source IP address to its own IR-IP. This is because an RNVE
does not recognize the AR-IP.

4.3. RNVE Procedure

There is no change to the RNVE control and forwarding procedures.
RNVE follows the regular ingress replication procedure defined in
[RFC7432].

5. AR-LEAF's Peer multihomed NVE in the Extended Optimized-IR Procedure

   For the extended Optimized-IR procedures specified in this document,
   a multihomed AR-LEAF may keep track of the common multihomed ES it
   shares with other remote NVEs in a BD specific scope or in an EVI
   scope. Correspondingly, an Extended-MH AR-REPLICATOR MUST also use
   the same scheme to keep track of the common multihomed ES that its
   AR-LEAF shares with other remote NVEs. All multihomed AR-LEAFs and
   all AR-REPLICATORs within the same EVI MUST use the same scheme to
   keep track of the common multihomed ES that an AR-LEAF shares with
   other remote NVEs. This consistency can be enforced through a manual
   configuration.

   A multihomed AR-LEAF maintains a peer-multihomed-flood-list for each
   BD it attaches. If the common multihomed ES is tracked in a per EVI
   scope, an AR-LEAF's peer-multihomed-flood-list for a given BD X
   contains all the NVEs in BD X that have at least one multihomed ES in
   common with it, regardless whether each common multihomed ES contains
   BD X or not. If the common multihomed ES is tracked in a BD specific
   scope, for a given BD X, each common multihomed ES must contain BD X.
   The same MUST be applied to the AR-LEAF's multihomed-list for BD X an
   AR-REPLICATOR maintains for its AR-LEAF.

   When the Ethernet A-D per EVI route is advertised at the granularity
   of per ES, the common multihomed ES is tracked in a per EVI scope.

6. Multicast Flags Extended Community

   The EVPN Multicast Flags Extended Community is defined in the
   [EVPN-IGMP-PROXY]. This transitive extended community can carry many
   flags in its Flags field. This document proposes one new flag in the
   Flags bit vector.

   o Extended-MH-AR

   The Extended-MH-AR flag, M flag for short, takes the next available
   low-order bit from the Flags field.

   The Extended-MH-AR flag is used by the AR-REPLICATOR. When this flag
   is set, the AR-REPLICATOR indicates to other NVEs that it will not
   retain the source IP address or include the ESI label for an ingress
   NVE when replicating the packet over an NVO tunnels on behalf of the
   ingress NVE. Such an AR-REPLICATOR supports the extended optimized-
   IR procedures defined in this document.

7. IANA Considerations

   A request for a new flag named Extended-MH-AR flag in the Flags field
   of the multicast flags extended community will be submitted to IANA.

8. Security Considerations

   This document inherits the same securities as they are defined in the
   [RFC7432], [RFC8365] and [EVPN-AR].

9. Acknowledgements

   The authors would like to thank Eric Rosen and Jeffrey Zhang for
   their valuable comments and feedbacks. The authors would also like
   to thank Aldrin Isaac for his useful discussion, insight on this
   subject.

10. Normative References

   [EVPN-AR] Rabadan, J., Ed., "Optimized Ingress Replication solution
              for EVPN", internet-draft ietf-bess-evpn-optimized-ir-
              06.txt, October 2018.

   [EVPN-IGMP-PROXY]
              Sajassi, A., Ed., "IGMP and MLD Proxy for EVPN", internet-
              draft ietf-bess-evpn-igmp-mld-proxy-02.txt, June 2018.

   [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>.

   [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>.

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

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

Authors' Addresses

   Wen Lin (editor)
   Juniper Networks, Inc.

   EMail: wlin@juniper.net

   Selvakumar Sivaraj
   Juniper Networks, Inc.

   EMail: ssivaraj@juniper.net

   Vishal Garg
   Juniper Networks, Inc.

   EMail: vishalg@juniper.net

   Jorge Rabadan
   Nokia

   EMail: jorge.rabadan@nokia.com