EVPN VPWS as VRF Attachment Circuit

Abstract

When a VRF Attachment Cirucit (VRF-AC) is far away from its IP-VRF
instance, we can deploy an EVPN VPWS ([RFC8214]) between that VRF-AC
and its IP-VRF instance. From the viewpoint of the IP-VRF instance,
a local virtual interface takes the place of that remote "VRF-AC".
The intended IP address for that VRF-AC is now configured to the
virtual interface, in other words, the virtual interface is the
actual VRF-AC of the IP-VRF instance. The virtual interface is also
the AC of that VPWS instance, in other words, the virtual interface
is cross-connected to that remote "VRF-AC" by the VPWS instance.

This document proposes an extension to [RFC7432] to support this
scenario.

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 29 January 2022.

Copyright Notice

Table of Contents

1. Introduction

When a VRF Attachment Cirucit (VRF-AC) is far away from its IP-VRF instance, we can deploy an EVPN VPWS ([RFC8214]) between that VRF-AC and its IP-VRF instance. From the viewpoint of the IP-VRF instance, a local virtual interface takes the place of that remote "VRF-AC". The intended IP address for that VRF-AC is now configured to the virtual interface, in other words, the virtual interface is the actual VRF-AC of the IP-VRF instance. The virtual interface is also the AC of that VPWS instance, in other words, the virtual interface is cross-connected to that remote "VRF-AC" by the VPWS instance.

The requirements of this scenario is described in Section 1.1.

1.1. Integrated Routing and Cross-connecting

When an IP-VRF instance and an EVPN VPWS instance are connected by an virtual-interface, We call such scenarios as Integrated Routing and Cross-connecting (IRC) use-case, and the virtual-interface connecting EVPN VPWS and IP-VRF is called as IRC interface, because that the packets received from the virtual-interface is routed in the IP-VRF and the data packets sent to the virtual-interface is cross-connected to the remote AC of that EVPN VPWS.

The IRC use case are illustrated by the following figure:

```
                    PE1
              +---------------------+
              |      IRC1=10.1      |
              |   +-----+   +------+ |.
             .|   |VPWS1|---|IPVRF1| |   .
             .  |   +-----+   +------+ |    .
    PE4      .   |                    |      .   PE3
  +--------+.       +---------------------+      +---------+
  |        |        |                    |       |         |
  |+-----+ |        | RT-2               |+------+ |
  ||VPWS1| |        | <10.2, N1>         ||IPVRF1| |
  |+-----+ |        | label2=IPVRF1  |+------+ |
  |   |    |        | label1=VPWS1   |   |     |
  +---|----+.       | RT=VPWS1       .+---|-----+
      |     .     PE2          V          .     |
      |       .  +---------------------+  .      |
      |       .| |      IRC2=10.1      |.        |
    N1=10.2    |   +-----+   +------+ |         N2=20.2
              |   |VPWS1|---|IPVRF1| |
  Behind N1:   |   +-----+   +------+ |
  60.0/24      |                    |
  70.0/24      +---------------------+
```
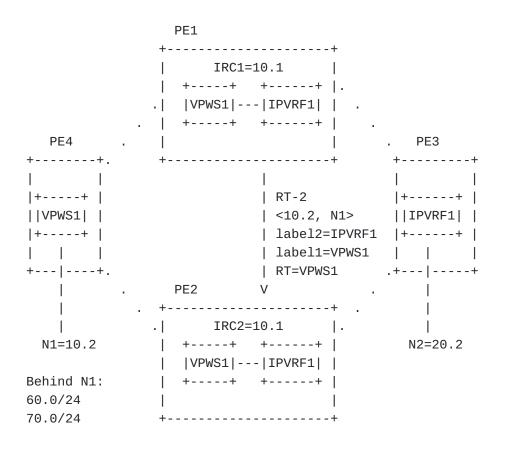
                 Figure 1: ARP/ND Synchronizing for IRC Interfaces

   There are four PE nodes named PE1/PE2/PE3/PE4 in the above network.
   PE4 is a pure EVPN VPWS PE, there may be no IP-VRFs on it. PE3 is a
   pure L3 EVPN PE, there may be no VPWSes or MAC-VRFs on it. PE1 and
   PE2 are the border of the EVPN VPWS domain and the L3 EVPN domain,
   so they are both EVPN VPWS PE and L3 EVPN PE, there will be both
   EVPN IP-VRFs and EVPN VPWSes on them.

   N1/N2/N3/N4 may be a host or an IP router. N1/N3/N4 is in the subnet
   10.0/24. N2 is in the subnet 20.0/24. When N1/N2/N3/N4 is a host, it
   is also called H1/H2/H3/H4 in this document. When N1/N2/N3/N4 is a
   router, it is also called R1/R2/R3/R4 in this document. N1/N2/N3/
   N4's MAC address is M1/M2/M3/M4 respectively.

   When N1 is a Router, there are two subnets behind N1, these subnets
   are 60.0/24 and 70.0/24.

   Note that there may be L2 switches between N1/N2/N3/N4 and their
   PEs. These switches are not shown in Figure 1.

   Note that the IRC interfaces are considered as AC interfaces in EVPN
   VPWS instances. At the same time, they are considered as VRF-ACs in
   IP-VRF instances.

When N1 sends an ARP Request REQ_P1, then REQ_P1 will be forwarded by PE4 to either PE1 or PE2, not to the both. Both the IRC1 on PE1 and IRC2 on PE2 are N1's subnet-gateway(SNGW). But when N2 send an ARP Reply REP_P2 to N1, then PE3 may load-balance REP_P2 to either PE1 or PE2, not to the both.

When REQ_P1 is load-balanced to PE1, not to PE2, but PE3 load-balance REP_P2 to PE2, The ARP entry of N1 will not be prepared on PE2 for REP_P2. So the fowarding of REP_P2 will be delayed due to ARP missing.

We use RT-2 routes to advertise the ARP entry of N1 from PE2 to PE3. But there SHOULD be no RT-2 advertisement in EVPN VPWS according to [RFC8214]. So the RT-2 routes from PE2 to PE3 SHOULD not carry any export-RTs of VPWS1, and the MPLS label1 field of these RT-2 routes should be set to NULL, not VPWS1.

Note that an ESI may be assigned to IRC1 and IRC2, But it is not necessary to advertise that ESI in the L3 EVPN domain. The ESI may be advertised in the EVPN VPWS domain only.

1.2. Terminology

Most of the terminology used in this documents comes from [RFC7432] and [I-D.ietf-bess-evpn-prefix-advertisement] except for the following:

VRF AC: VRF Attachment Circuit, An Attachment Circuit (AC) that attaches a CE to an IP-VRF. It is defined in [RFC4364].

IRC: Integrated Routing and Cross-connecting, thus a IRC interface is the virtual interface connecting an IP-VRF and an EVPN VPWS.

L3 EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN which contains VRF ACs and maybe contains IRB interfaces or IRC interfaces.

IP-AD/EVI: Ethernet Auto-Discovery route per EVI, and the EVI here is an IP-VRF.

IP-AD/ES: Ethernet Auto-Discovery route per ES, and the EVI for one of its route targets is an IP-VRF.

CE-BGP: The BGP session between PE and CE. Note that CE-BGP route doesn't have a RD or Route-Target.

RMAC: Router's MAC, which is signaled in the Router's MAC extended community.

RT-2E: A MAC/IP Advertisement Route with a non-reserved ESI.

RT-5E: An EVPN Prefix Advertisement Route with a non-reserved ESI.

RT-5G: An EVPN Prefix Advertisement Route with a zero ESI and a non-zero GW-IP.

RT-5L: An EVPN Prefix Advertisement Route with both zero ESI and zero GW-IP, but a valid MPLS label.

2. ARP/ND Synching and IP Prefix Synching

Host IP-MAC relations are learnt by PEs on the access side via a control plane protocol like ARP. In case where N1 is multihomed to multiple L3 EVPN PE nodes by an All-Active EVPN VPWS, N1's Host IP/ MAC will be learnt and advertised in the MAC/IP Advertisement only by the PE that receives the ARP packet. The MAC/ IP Advertisement with non-zero ESI will be received by the other multihomed PEs.

As a result, after PE2 receives the MAC/IP Advertisement and imports it to the VPWS Service Instance, PE2 installs an ARP entry to the VPWS Service instance's IRC interface. Such ARP entry is called remote synched ARP Entry in this document.

Note that the PE3 follows the DGW1 behavior of [I-D.ietf-bess-evpn-prefix-advertisement]'s section 4.1 to achieve the load balancing procedures based on the recursive route resolution by the GW-IP Overlay Index.

When PE3 load balance the traffic towards PE1/PE2, both PE1 and PE2 would have been prepared with corresponding ARP entry yet because of the following ARP synching procedures.

2.1. Constructing MAC/IP Advertisement Route

This draft introduces a new usage/construction of MAC/IP Advertisement route to enable ARP/ND synching for IP addresses in EVPN IRC use-cases. The usage/construction of this route remains similar to that described in RFC 7432 with a few notable exceptions as below.

  *The Route-Distinguisher should be set to the corresponding EVPN-VPWS context.

  *The Ethernet Tag should be set to the VPWS Service Instance Identifier of the IRC interface.

*The MAC/IP Advertisement SHOULD carry one EVI-RT (for the EVPN
    VPWS instance) and one ES-Import RT (for the ESI of the IRC
    interface).

   *The ESI can be set to the ESI of the IRC interface.

    Note that the receiver use the ESI and Ethernet Tag ID to
    determine the VPWS Service Instance whose IRC interface is the
    interface that the synced ARP entry will be installed to.

   *The MPLS Label1 should be set to the label of the <ESI,VPWS
    service instance identifier>.

   *The MPLS Label2 is optional. When it is used, it should be set to
    IPVRF1.

   *If the MPLS Label2 is used, the RMAC Extended Community attribute
    SHOULD be carried in VXLAN EVPN.

2.2. Constructing Ethernet A-D Route

   The ESI of the IRC interface is mainly used in the EVPN VPWS domain.
   That ESI typically has nothing to do with the fundamental function
   of the L3 EVPN domain.

   Note that PE3 or PE4 will not import the RT-2 route with an ES-
   import RT it doesn't recognize.

   Note that the Ethernet A-D route advertisement in the EVPN VPWS
   domain still follows [RFC8214]. The IRC interface is considered as
   an ordinary AC in the EVPN VPWS domain.

2.3. Constructing IP Prefix Advertisement Route

   There may be two types of IP prefixes on PE1/PE2. The first type is
   the prefix of the IRC interface itself. The second type is the
   prefixes behind N1 (especially when N1 is a router).

   Given that PE1/PE2 can install synced ARP entries to its proper IRC
   interface benefitting from the RT-2 route of Section 2. This ensures
   that both PE1 and PE2 will know all hosts of the IRC interface's own
   subnet. So it is not necessary for PE1/PE2 to advertise per-host IP
   prefixes of that subnet to PE3 by RT-2 routes. It is recommended
   that PE1/PE2 advertise a single RT-5 route of that subnet to PE3
   instead. The ESI of these RT-5 routes can be simply set to zero,
   because when PE3 receives such RT-5 routes from both PE1 and PE2,
   PE3 can consider them as ECMP or FRR even when their ESI is zero.

   Note that N1 may be a host or a router, when it is a router, there
   may be some prefixes behind N1 on PE1. Those prefixes will be learnt

via a PE-CE route protocol. N1's IP address may be considered as the overlay nexthop of those prefixes. The overlay nexthop of those prefixes will be carried in the RT-5 route's GW-IP field. Those RT-5 routes are called as RT-5G routes because their Overlay Indexes are their GW-IPs (and their ESI and label are zero).

Note that those RT-5G routes are advertised by PE1 to both PE2 and PE3. If the IRC1 interface fails, the prefixes of the second type will achieve more faster convergency on PE3 by the withdraw (from PE1) of the corresponding prefix of the first type.

3. Packet Walk Through

The procedures for local/remote host learning and MAC/IP Advertisement route constructing are described above.

When R2(N2) send a data packet P21 to a host 60.1 whose location is behind R1(N1), P21 will matches prefix 60.0/24 on PE3. The RT-5G route for 60.0/24 will be used. The GW-IP of that RT-5G route is 10.1 (R1). So PE3 use 10.1 to do recursive route resolution and matches the RT-5L route of 10.0/24.

Note that the recursive route resolution follows the DGW1 behavior of [I-D.ietf-bess-evpn-prefix-advertisement]'s section 4.1.

Both PE1 and PE2 have advertised the RT-5L route of 10.0/24 to PE3. PE3 may consider them as ECMP or FRR, depending on their route attributes. Then PE3 should forward P21 to PE1 or PE2, depending on the ECMP/FRR procedures.

We can assume that it is PE2 that will receive P21 from PE3. The destination IP of P21 is in prefix 60.0/24. That prefix has been installed into IPVRF1 on PE2. PE2 previously received that prefix either from a PE-CE route protocol or from a RT-5G route from PE1. The overlay nexthop or GW-IP of prefix 60.0/24 is 10.1 (R1). The outgoing interface for P21 is IRC2 interface.

The ARP entry for 10.1 is a synched ARP entry, because PE1 sent the ARP Request only to PE1. It is intalled to IRC2 interface just because the RT-2 route's route target mathes the EVPN VPWS instance and the RT-2 route's <ESI,Ethernet Tag ID> matches the IRC2 interfaces's ESI and VPWS Service Instance ID.

Then P21 is encapsulated with a ethernet header and becomes an ethernet packet P21E. The destination MAC address of P21E is N1's MAC address which is determined by that ARP entry. The source MAC address of P21E is IRC2's MAC address. Then P21E is sent over IRC2 interface.

After P21E is sent over IRC2 interface, it will be forwarded to PE4 in the EVPN VPWS instance according to [RFC8214]

4. Fast Convergence for Routed Traffic

When IRC1 interface goes down, PE1 will withdraw the RT-5L route of 10.0/24. And the RT-5G routes of 60.0/24 and 70.0/24 will be just changed to stale state. When PE3 receives the withdraw of that RT-5L route, it will stop to forward the data packets of those two subnets to PE1 again. But PE3 will continue to forward these data packets to PE2.

5. Considerations on ABRs and Route Reflectors

When an ABR or ASBR receives a MAC/IP Advertisement Route that contains both EVI-RT and ES-Import RT, It should re-advertise that route even if that route's MPLS label1 is null (It should not consider that route as malformed). When that route's nexthop are changed to itself, It don't have to allocate a new label for each RT-2 route's MPLS label1 field separately. That field can be rewritten to the same preconfigured MPLS label that will blackhole the data packets it received. But the MPLS label2 (if is not null) field should be rewritten normally along with the nexthop-rewritting.

6. Security Considerations

This document does not introduce any new security considerations other than already discussed in [RFC7432] and [I-D.ietf-bess-evpn-prefix-advertisement].

7. IANA Considerations

There is no IANA consideration needed.

8. Normative References

[I-D.ietf-bess-srv6-services]
          Dawra, G., Filsfils, C., Talaulikar, K., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "SRv6 BGP based Overlay Services", Work in Progress, Internet-Draft, draft-ietf-bess-srv6-services-07, 11 April 2021, <https://datatracker.ietf.org/doc/html/draft-ietf-bess-srv6-services-07>.

[I-D.ietf-bess-evpn-prefix-advertisement]
          Rabadan, J., Henderickx, W., Drake, J., Lin, W., and A. Sajassi, "IP Prefix Advertisement in EVPN", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-prefix-advertisement-11, 18 May 2018, <https://

datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-prefix-advertisement-11>.

[I-D.ietf-bess-evpn-inter-subnet-forwarding]
          Sajassi, A., Salam, S., Thoria, S., Drake, J., and J.
          Rabadan, "Integrated Routing and Bridging in EVPN", Work
          in Progress, Internet-Draft, draft-ietf-bess-evpn-inter-
          subnet-forwarding-15, 26 July 2021, <https://
          datatracker.ietf.org/doc/html/draft-ietf-bess-evpn-inter-
          subnet-forwarding-15>.

[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>.

[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>.

9. Informative References

[RFC8214] Boutros, S., Sajassi, A., Salam, S., Drake, J., and J.
          Rabadan, "Virtual Private Wire Service Support in
          Ethernet VPN", RFC 8214, DOI 10.17487/RFC8214, August
          2017, <https://www.rfc-editor.org/info/rfc8214>.

Authors' Addresses

Yubao Wang
ZTE Corporation
No. 68 of Zijinghua Road, Yuhuatai Distinct
Nanjing
China

Email: wang.yubao2@zte.com.cn

Zheng(Sandy) Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn