Workgroup: bess
Internet-Draft:
draft-nagaraj-bess-evpn-redundant-mcast-src-update-00
Published: 24 October 2022
Intended Status: Standards Track
Expires: 27 April 2023
A
u
t
h
o
r
s
:
V. Nagaraj
Juniper Networks
V. Nagarajan
Juniper Networks
Z. Zhang
Juniper Networks
J. Rabadan
Nokia

Enhancements to Multicast Source Redundancy in EVPN Networks

Abstract

draft-ietf-bess-evpn-redundant-mcast-source specifies Warm Standby (WS) and Hot Standby (HS) procedures for handling redundant multicast traffic into an EVPN tenant domain. With the Hot Standby procedure, multiple ingress PEs may inject traffic and an egress PE will decide from which ingress PE traffic will be accepted and forwarded. The decision is based on certain signaling messages and/or BFD status of provider tunnels from the ingress PEs, and the traffic is associated with ingress PEs based on Ethernet Segment Identifier (ESI) labels. As a result, the procedures in that document only apply to MPLS data plane. This document extends the Hot Standby procedures to non-MPLS data planes and EVPN Data Center Interconnect scenarios.

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 27 April 2023.

Table of Contents

1. Background

[I-D.ietf-bess-evpn-redundant-mcast-source] specifies Warm Standby and Hot Standby procedures for handling redundant multicast traffic into an EVPN tenant domain. With the Hot Standby procedure, multiple ingress PEs will inject traffic and an egress PE will decide from which ingress PE traffic will be accepted and forwarded.

The PEs that inject redundant traffic advertises Selective Provider Multicast Service Interface (S-PMSI) A-D routes. The routes carry an EVPN Multicast Flags Extended Community with a bit to indicate that matching traffic is from redundant sources. With MPLS data plane, the routes also carry an Ethernet Segment Identifier (ESI) label, indicating the Ethernet Segment on which the traffic is received.

When an egress PE receives S-PMSI A-D routes, it decides from which ingress PE it should accept the traffic. The decision could be based on the following factors:

*The presence/lack of S-PMSI A-D routes from ingress PEs of the redundant traffic

*The presence/lack of Ethernet A-D per EVI routes from ingress PEs for the Ethernet Segment that redundant traffic arrives on

*BFD status of provider tunnels for redundant traffic

All the above options are local behaviors on individual egress PEs.

1.1. Hot Standby mode in non-MPLS IP tunnels

With MPLS data plane, the Hot Standby redundant flows from different PEs are distinguished via ESI labels. With non-MPLS IP encapsulations like VXLAN/NVGRE, this document specifies that the redundant flows are distinguished by the source IP address (Source VTEP IP) in the outer IP header.

This document also makes explicit that non-MPLS IP tunnels that carry an identifier of the source Ethernet Segment reuse all the procedures of [I-D.ietf-bess-evpn-redundant-mcast-source] for Hot Standby redundancy. Examples of these tunnels used by EVPN are GENEVE [I-D.ietf-bess-evpn-geneve] and SRv6 [RFC9252].

1.2. EVPN DCI Use Case

When an EVPN network is used as Data Center Interconnect (DCI) for DCs (e.g., VxLAN or EVPN), multiple gateways (GWs) are placed between a DC and DCI, as described in [RFC9014]. A virtual Ethernet Segment is defined for each EVPN (the DC and/or DCI) and multi-homed to the GWs. A Designated Forwarder (DF) is elected for each virtual ES (ethernet segment). Each GW can receive the same BUM traffic from a DC/DCI EVPN but only the DF will forward traffic to the next DCI/ DC (corresponding to the virtual ES).

This section discusses how source redundancy works with DCI, and how DCI GWs can optionally introduce redundant flows even when there is no source redundancy at source DC.

1.2.1. True Source Redundancy

[I-D.ietf-bess-evpn-redundant-mcast-source] is MPLS-based. It is "true" source redundancy in that multiple of sources of the same flow are attached to different Ethernet Segments. S-PMSI A-D routes announce the redundant flows and carry ESI Label Extended Communities (ECs) for the ESes so that an egress PE can choose from which source ES the packets will be accepted.

With DCI, the source ESes are hidden outside the source DC, and different DC/DCI may use different data planes. Additionally, currently only the GW that is the DF for the Interconnect Ethernet Segment (I-ES) will forward BUM traffic to the downstream DC/DCI, so the benefit of HS is lost once the first DC boundary is crossed.

The above issues are solved as following:

* The GWs forward accepted redundant flows regardless of DF status. Note that, a GW will only accept one of the redundant flows from its redundant upstream PEs/GWs.

* The GWs remove ESI Label ECs when they re-originate the S-PMSI A-D routes into the next DC/DCI. Note that, Even if the downstream DC/DCI is MPLS, the re-originated S-PMSI A-D routes do not carry the ESI Label EC for the I-ES. This is because the GWs use the same ESI label for the I-ES, so the ESI label cannot be used to distinguish the flows.

*When the S-PMSI A-D routes do not carry ESI Label ECs, an egress PE chooses from which PE/GW (vs. ES) to accept traffic from.

- In case of IP based data plane, this is the same as non-DCI case.

- In case of MPLS data plane, a PE needs to be able to distinguish from which node the traffic is. In some cases, the PE Distinguisher Label concept [RFC6513] need to be used.

1.2.2. GW-introduced Flow Redundancy

In the "true source redundancy" case, S-PMSI A-D routes announce the redundancy and the DCI GWs always forward accepted flows regardless of the DF status.

The GWs may also forward all BUM traffic regardless of DF status - not just those redundant flows announced by S-PMSI A-D routes. This creates a similar scenario of source redundancy, though it is introduced by the GWs. A downstream GW/PE can choose which redundant flows need to be accepted/discarded based on the A-D per ES routes for the I-ES instead of S-PMSI A-D routes.

This requires that all downstream PEs/GWs behave consistently. That is ensured either based on provisioning or based on signaling (details to be added in a future revision).

In the "true source redundancy" case, all flows covered by the *(, g) or (s-prefix, g) in the S-PMSI A-D routes are treated as redundant flows. In the GW-introduced redundancy, (*, g) flows are treated as distinct flows that have redundant copies. They may be from different PEs in the local DC and all must be accepted, or they may be from different DCs in which case only traffic from one GW for each upstream DC can be accepted, as explained below.

Consider that a DCI interconnects three DCs. GW1a/GW1b connect DC1 and the DCI, GW2a/GW2b connect DC2 and the DCI, and GW3a/GW3b connect DC3 and the DCI.

An egress PE1 in DC1 may need to accept and forward *(, G) traffic from all local PEs in DC1 and GW1a but not from the GW1b. To do so, it installs a (*, G) forwarding state in a BD (Broadcast Domain) with indication that traffic from GW1b must be discarded. Similarly, GW3a/GE3b may need to accept and forward *(, G) traffic from GW1a/ GW2a but not from GW1b/GW2b. To do so, it installs a (*, G) forwarding state with indication that traffic from GW1b/GW2b must be discarded.

The reverse logic (of specifying PEs/GWs from which traffic should not be accepted) is only needed for (*, G) entries in the DCI case. For (S, G) case, the reverse logic is not needed because an egress PE should be able to decide from which PE/GW the traffic should be accepted.

1.2.3. Co-existence of Source Redundancy and GW-introduced Redundancy

Both flavors of redundancy can co-exist. For redundant flows announced by S-PMSI A-D routes, the method described in

Section 1.2.1 is used. For GW-introduced redundancy, the method described in Section 1.2.2 is used. The difference between the two on downstream PEs/GWs is that one uses S-PMSI A-D routes while the other uses I-ES A-D per ES routes to choose which flow to accept, and for (*,g) flows in the latter case, reverse logic is needed.

2. Specifications

2.1. Hot Standby Procedures for non-MPLS IP tunnels

In case the EVPN network uses non-MPLS IP tunnels without source Ethernet Segment identification, e.g., VXLAN/NVGRE, the procedures in [I-D.ietf-bess-evpn-redundant-mcast-source] for Hot Standby redundancy are modified as follows:

* The S-PMSI A-D routes advertised for each SFG (Single Flow Group) by the upstream PEs MUST NOT carry any ESI Label Extended Communities. The rest of the procedures on the upstream PEs remain the same.

* Upon receiving the S-PMSI A-D routes, the downstream PEs select a primary upstream PE out of the list of (S-PMSI A-D route) next hops and add an RPF check to the *(,G)/(S,G) state in the BD or SBD (Supplementary Broadcast Domain). This RPF check discards all ingress packets to* (,G)/(S,G) that are not received from the selected primary Source VTEP. The selection of the primary upstream PE is a matter of local policy, for instance, an egress PE could keep track of traffic statistics of redundant flows and dynamically decide which flow is accepted based on traffic threshold information.

  The selection of the upstream PE for non-MPLS IP tunnels, instead of the primary Source Ethernet Segment, provides a solution for redundant sources connected to different upstream PEs, however it MUST NOT be used when the redundant sources are connected to the same upstream PE, or multi-homed to the same set of upstream PEs.

  In case the EVPN network uses non-MPLS IP tunnels that can carry a source Ethernet Segment identification, e.g., GENEVE or SRv6, all the procedures in [I-D.ietf-bess-evpn-redundant-mcast-source] for Hot Standby redundancy are followed. The following considerations apply:

* In case of GENEVE [I-D.ietf-bess-evpn-geneve], an Ethernet option TLV MUST encode the ESI (Ethernet Segment Identifier) label value. This ESI label value is signaled by the EVPN A-D per ES routes, and advertised for SFG sources in S-PMSI A-D routes in the ESI Label Extended Communities as described in [I-D.ietf-bess-evpn-redundant-mcast-source]. The downstream PE can identify the packets coming from a selected primary Source Ethernet Segment based on a lookup on the Source Identifier of the Ethernet option TLV.

* In case of SRv6 [RFC9252], the upstream PEs send multicast packets encapsulated in SRv6 tunnels that use End.DT2M as function and Arg.FE2 as argument. The Arg.FE2 argument in the packets identify the Source Ethernet Segment. The argument is signaled by the EVPN A-D per ES routes as specified in [RFC9252],

and this document uses the same encoding of the argument also for the S-PMSI A-D routes that signal the Source Ethernet Segments for SFG sources, with the consideration that there may be multiple arguments signaled and that the arguments for the same Ethernet Segment in different upstream PEs MUST match. The downstream PE can then identify the packets coming from a selected primary Source Ethernet Segment based on the received argument.

2.2. Procedures for EVPN DCI Use Case

To be added.

3. Security Considerations

No additional security considerations are needed besides what are in [I-D.ietf-bess-evpn-redundant-mcast-source].

4. Acknowledgements

5. References

5.1. Normative References

[I-D.ietf-bess-evpn-geneve] Boutros, S., Sajassi, A., Drake, J., Rabadan, J., and S. Aldrin, "EVPN control plane for Geneve", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-geneve-04, 23 May 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-geneve-04.txt>.

[I-D.ietf-bess-evpn-redundant-mcast-source] Rabadan, J., Kotalwar, J., Sathappan, S., Zhang, Z. J., Lin, W., and E. C. Rosen, "Multicast Source Redundancy in EVPN Networks", Work in Progress, Internet-Draft, draft-ietf-bess-evpn-redundant-mcast-source-04, 7 October 2022, <https://www.ietf.org/archive/id/draft-ietf-bess-evpn-redundant-mcast-source-04.txt>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <https://www.rfc-editor.org/info/rfc6513>.

[RFC9252] Dawra, G., Ed., Talaulikar, K., Ed., Raszuk, R., Decraene, B., Zhuang, S., and J. Rabadan, "BGP Overlay Services Based on Segment Routing over IPv6 (SRv6)", RFC 9252, DOI 10.17487/RFC9252, July 2022, <https://www.rfc-editor.org/info/rfc9252>.

5.2. Informative References

[RFC9014] Rabadan, J., Ed., Sathappan, S., Henderickx, W., Sajassi, A., and J. Drake, "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, DOI 10.17487/RFC9014, May 2021, <https://www.rfc-editor.org/info/rfc9014>.

Authors' Addresses

Vinod Kumar Nagaraj
Juniper Networks

Email: vinkumar@juniper.net

Vikram Nagarajan
Juniper Networks

Email: vikramna@juniper.net

Zhaohui Zhang
Juniper Networks

Email: zzhang@juniper.net

Jorge Rabadan
Nokia

Email: jorge.rabadan@nokia.com