Multicast Source Redundancy for EVPN Non-MPLS data planes & DCI deployments

draft-nagaraj-bess-evpn-redundant-mcast-src-update-00

Vinod Kumar Nagaraj (Juniper)
Vikram Nagarajan (Juniper)
Jeffrey Zhang (Juniper)
Jorge Rabadan (Nokia)

IETF115, Nov 2022
London

Background

draft-ietf-bess-evpn-redundant-mcast-source-04

- Defines procedures for Multicast Source redundancy in EVPN MPLS deployments.

Terminology

- Upstream PEs, Downstream PEs, S-ES (Source Ethernet Segment), SFG (Single Flow Group), ESI (Ethernet Segment Identifier).
- OISM (Optimized Inter-subnet Multicast), DF (Designated Forwarder), SF (Single Forwarder), HS (Hot Standby), WS (Warm Standby).

Enhancements

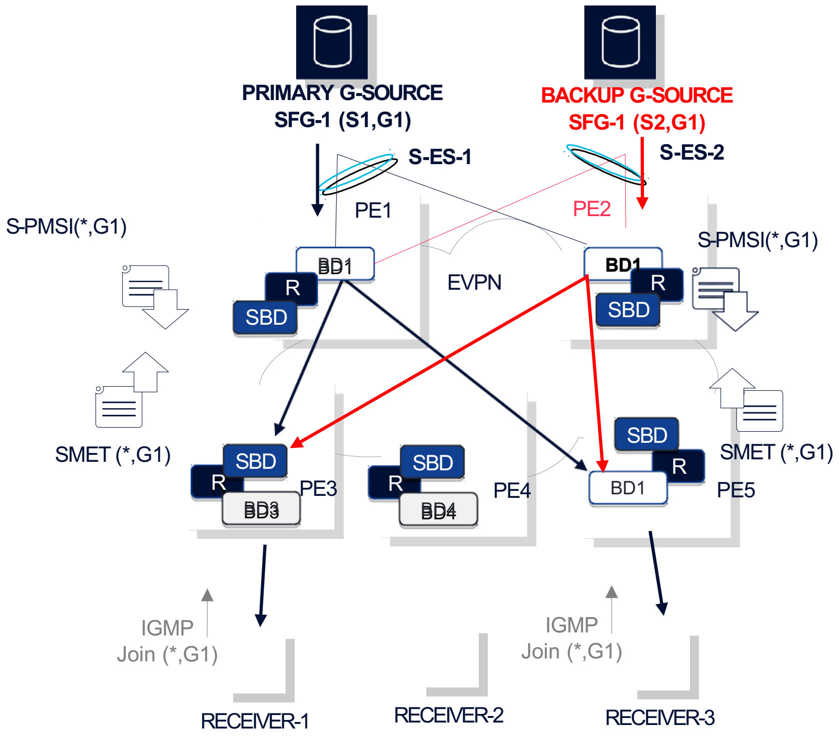
draft-nagaraj-bess-evpn-redundant-mcast-src-update-00

- Defines procedures for Multicast source redundancy in EVPN Non-MPLS (VXLAN/NVGRE/SRV6/GENEVE) deployments.
- Defines procedures for Multicast source redundancy in EVPN DCI MPLS/Non-MPLS deployments.

Terminology

- VTEP (VXLAN Tunnel Endpoint), S-VTEP IP (Source VTEP IP), I-ES (Interconnect Ethernet Segment).
- DCI (Data Center Interconnect), SBD (Supplementary Broadcast Domain), RPF (Reverse Path Forwarding)
- VNI (VXLAN Network Identifier)

Hot Standby (HS) Solution for Non-MPLS Deployments



S-ES – Ethernet Segment associated to a G-Source

Reuse all procedures of [I-D.ietf-bess-evpn-redundant-mcast-source] for hot standby redundancy with modifications as below for EVPN Non-MPLS Deployments

For VXLAN/NVGRE deployments (S-ES identification not carried)

- S-PMSI AD routes advertised by upstream PEs for each SFG MUST NOT carry ESI Label EC.
- Downstream PE elects Primary upstream PE from the list of S-PMSI route next-hops. Redundant flows are distinguished by source IP address (Source VTEP IP) in the outer IP header. RPF check enforced based on S-VTEP IP.

For SRV6/GENEVE deployments (S-ES identification carried)

- For GENEVE deployments, Ethernet option TLV must encode ESI Label Value. This ESI label value is signaled by AD/ES route & advertised for SFG sources in S-PMSI AD routes. Redundant flows are distinguished based on Source Identifier in Ethernet Option TLV. RPF check enforced based on Source-ID of Ethernet Option TLV.
- For SRV6 deployments, the upstream PEs send multicast packets encapsulated in SRv6 tunnels that use End.DT2M as function and Arg.FE2 as argument. Arg.FE2 argument is signaled by AD/ES route & advertised for SFG sources in S-PMSI AD routes. RPF check enforced based on Arg.FE2 argument.

EVPN DCI Redundancy - Overview

DCI redundancy procedures

- ❑ Source redundancy exists in ingress/source DC
- ❑ No source redundancy in ingress/source DC

True Source redundancy with EVPN DCI

- Source redundancy exists in ingress/source DC - sources of the same flow are attached to different Ethernet Segments.
- 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 GWs forward accepted redundant flows regardless of DF status.
- The GWs remove ESI Label ECs when they re-originate the S-PMSI A-D routes into the next DC/DCI.
- In case of IP data plane, Egress PE chooses traffic from an upstream PE/GW based on procedures described in slide 3.
- In case of MPLS data plane, Egress PE chooses traffic from an upstream PE/GW based on PE distinguisher label.

GW introduced Flow redundancy with EVPN DCI

- No source redundancy in ingress/source DC
- The GWs may also forward all multicast traffic regardless of DF status – GW-DF & GW-BDF may be forwarding multicast traffic from a DC to DCI & vice-versa. This creates a similar scenario of source redundancy, though it is introduced by the GWs.
- 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.
- Similarly, GW3a/GE3b may need to accept and forward (*, G) traffic from GW1a/ GW2a but not from GW1b/GW2b.
- For (S,G) case, I-ES A-D per ES routes are used to choose primary upstream, and for (*,g) flows, reverse logic (of specifying PEs/GWs from which traffic should not be accepted) is needed.

Next steps

❑ Update EVPN DCI redundancy procedures in the next revision

❑ Authors would like to request Feedback from WG

Thank you