EVPN Anycast Aliasing For Multi-Homing

draft-rabnag-bess-evpn-anycast-aliasing-00

Jorge Rabadan (Nokia)
Kiran Nagaraj (Nokia)
Alex Nichol (Arista)
Nick Morris (Verizon)

IETF117, Jul 2023
San Francisco
Agenda

- Problem Statement
- BGP EVPN Extensions
- Anycast Aliasing Solution
- Next Steps
Problem Statement
draft-rabnag-bess-evpn-anycast-aliasing-00

EVPN Multi-Homing Aliasing

- Aliasing is an all-active multi-homing procedure that makes sure unicast traffic from TS3 to TS1 is per-flow load balanced to Leaf-1/Leaf-2 irrespective of the MAC/IP Advertisement route for TS1 not being advertised from both Leaf nodes
- Aliasing on Leaf-3 creates an “overlay ECMP-set” for ESI 1 in addition to the “underlay ECMP-set” to each VTEP of the ES

Challenges in very Large DCs

- Control Plane Scale
- Convergence and Processing overhead
- Hardware Resource consumption
- Inefficient forwarding during a failure
Problem Statement: Inefficient forwarding during failures
draft-rabnag-bess-evpn-anycast-aliasing-00

Failure Example
- Leaf1-to-Spine1 fails
- In-flight VXLAN packets from Leaf3 with destination VTEP Leaf1 arrive at Spine1 and are rerouted via e.g., Leaf2->Spine2->Leaf1->TS1, while they could go directly via Leaf2->TS1.
- After the underlay routing protocol converges, all VXLAN packets with destination VTEP Leaf1 are correctly sent to Spine2 and Leaf3 removes Spine1 from the underlay ECMP-set for Leaf L1.
Flags in the ESI label extended community
- RED: Multihomed redundancy mode [I-D.ietf-bess-rfc7432bis]
- SHT: Split Horizon type [I-D.ietf-bess-evpn-mh-split-horizon]
- A: Anycast Aliasing mode (This document)
  - Only set in all-active mode

Anycast VTEP address
- Loopback shared among NVEs attached to the same ES, vs Unicast VTEP which is the unique loopback associated to each NVE (advertised as next hop in MAC/IP routes)
- Encoded in the BGP Tunnel Encapsulation Attribute, Tunnel Egress Endpoint Sub-TLV (code point 6) [RFC9012] (never encoded as NLRI next hop)
- Advertised with the EVPN A-D per ES route for the ES
Anycast Aliasing Solution Procedures

draft-rabnag-bess-evpn-anycast-aliasing-00

What procedures are modified
- Only the Aliasing procedure on a per ES basis, based on the A flag
- Only if the ES NVEs use the same VNI/label

Egress NVEs (e.g. Leaf1 and Leaf2)
- ES NVEs configured for anycast aliasing mode, and with anycast VTEP
- Anycast VTEP reachability advertised in the underlay
- A-D per EVI routes are suppressed for the ES
- In case of ES link failure, anycast VTEP removed from underlay, and A-D per ES / ES route withdrawn

Ingress NVEs (e.g. Leaf3)
- Upon receiving MAC/IP route for TS1 with ESI1, programs TS1 with destination ESI1 (resolved to ESI1 anycast VTEP IP12)
- Incoming frames to TS1 encapsulated using IP DA = IP12 and use the underlay ECMP-set to IP12
- Mass withdraw triggered upon reception of last A-D per ES MP_UNREACH_NLRI
Multi-ES Anycast Aliasing Solution
draft-rabnag-bess-evpn-anycast-aliasing-00

Egress NVEs (e.g. Leaf1 and Leaf2)
- Anycast VTEP allocated per group of NVEs that share the same Ethernet Segments, e.g. IP12 used for ESI1 and ESI2
- IP12 is removed from underlay only if all ESes of the anycast VTEP are down (or the entire NVE is down)
- If traffic for TS1 is received on Leaf1 with IP12 and link is down, FRR procedures are applied (packets re-encapsulated with IP SA = IP12):
  - local-bias is bypassed if received traffic comes with anycast VTEP
  - local anycast VTEP as IP SA indicates FRR (no re-encap)

Ingress NVEs (e.g. Leaf3)
- anycast-aliasing-threshold = 2 (number of egress NVEs per ES under which the ingress PE uses the unicast VTEP instead of the anycast VTEP)
- collect-timer = t (triggered on Rx of MAC/IP route, upon expiration, the number of NVEs per ES is computed and compared against aa-threshold)
- If number of egress NVEs ≥ threshold → anycast VTEP is used
Anycast Aliasing benefits
draft-rabnag-bess-evpn-anycast-aliasing-00

• Suppression of A-D per EVI routes reduces the control plane pressure
• More efficient forwarding / better convergence in case of failures
• More efficient use of the hardware resources
Anycast Aliasing MUST NOT be used if
draft-rabnag-bess-evpn-anycast-aliasing-00

- Multihoming redundancy mode different from all-active
- ES used for VPWS ACs
- AC-DF [RFC8584] is required
- Unequal load balancing is needed [I-D.ietf-bess-evpn-unequal-lb]
- Transport is not NVO3 tunnels (i.e. IP tunnels)
- NVEs attached to the same ES do not use the same unicast VNI/label
Next Steps

draft-sr-bess-evpn-vpws-gateway-03

• Alternative solutions described in the draft
• Possible application to SRv6
• Feedback from WG
Thank you