Optimized Inter-Subnet Multicast for EVPN

draft-lin-bess-evpn-irb-mcast-04

W. Lin, Z. Zhang, J. Drake, E. Rosen

J. Rabadan, A. Sajassi

Optimized Inter-Subnet Multicast (OISM)

- Draft-lin-bess-evpn-irb-mcast-04
    - Authors: Lin, Zhang, Drake, Rosen, Rabadan, Sajassi
- Highlights of -04 revision
    - Revamped and expanded with many additional details and more explanatory material
    - Enhanced MVPN integration
    - Enhanced interworking with legacy PEs
- Highlights of this presentation:
    - Characteristics of OISM
    - MVPN integration

OISM Key Concepts

When IP Multicast Frame or Packet Received by EVPN-PE:

- Frame from ***local AC:***
  - Switch it to ***local*** receivers in source BD
  - Switch it in ***Layer 2 tunnel*** to egress PEs
    - In the Source BD or a Supplemental BD (SBD – present on all PEs of a tenant)
  - Route it via IRB interfaces to ***local*** receivers on non-source BDs
- Frame from ***Layer 2 tunnel***
  - Switch it to ***local*** receivers in receiving BD
    - Receiving BD is either the source BD or SBD
  - Route it via IRB interfaces to ***local*** receivers on other BDs
- Packet from ***external source*** (outside EVPN)
  - Route it via IRB interfaces to ***local*** receivers
  - Route it down IRB interface to "Supplementary" BD -- causes ***Layer 2 tunneling*** to egress PEs that do not get it from the external source

Characteristics I

- Maintains correct Ethernet emulation:
  - Receivers on source BD see unaltered frame
  - Receivers on other BDs see TTL decremented by 1
  - Operator's EVPN infrastructure remains invisible to tenants and to tenant applications
- If no sources or receivers for group G are external, *Join(*,G)* does not require RPs or Register messages
- Does not require PEs to run PIM unless needed for interworking with external sources or receivers

Characteristics II

- Works with all tunnel types: IR/AR/P2MP/BIER
  - MPLS or VxLAN
- Optimal routing/replication within Tenant Domain
- No change to EVPN multi-homing procedures
- Builds upon already-deployed features: SMET routes, IRB interfaces
- Maintains clear distinction between L2 and L3 multicast states

Interworking with Legacy EVPN PEs I

- Legacy PEs:
  - Don’t support OISM
  - May not even support IRB interfaces or SMET routes
- Easy scenario:
  - Legacy Ingress PE, OISM egress PE attached to source BD
  - Ingress PE sends unaltered frame as BUM traffic to egress PE (legacy procedures)
    - Egress PE may need to send back to ingress PE for other BDs

Interworking with Legacy EVPN PEs II

- Trickier Scenarios:
  - Ingress PE is legacy, egress PE (OISM or not) not attached to source BD
  - Inter-BD multicast between legacy PEs
- These scenarios require a gateway to relay traffic between ingress and egress PEs
  - For each BD, a dynamically selected gateway node relays traffic as needed
  - Gateway procedure simple if IR used by legacy PEs
    - More complicated if legacy PEs use P2MP, but that’s probably not a practical scenario

Interworking with External Multicast Infrastructure

- Two cases:
  - EVPN receivers, external sources
  - EVPN sources, external receivers
- *External* could be PIM/IP, MVPN, GTM
  - We will focus now on MVPN interworking
- MVPN interworking: some of the EVPN PEs become MVPN Points of Attachment *(MEGs)*
  - MEGs run MVPN and EVPN procedures
  - MEGs move traffic between MVPN and EVPN
  - Vanilla (non-MEG) MVPN and EVPN PEs do not see each other
- EVPN appears as a stub LAN to the external network
  - MEGs act as FHRs/LHRs as sources/receivers in EVPN

MVPN Interworking Principles

- Basic principle: **No entanglement**
  - MVPN and EVPN domains operate independently
  - Clean, clear interfaces between domains
  - Each domain has its native control/data plane
    - Interaction between multicast control planes achieved through the creation/deletion/modification of the L3 multicast states that each control plane recognizes
  - Procedures internal to MVPN or EVPN do not require modification to accommodate other domain:
    - E.g., no modification of EVPN multi-homing procedures
    - Intra-subnet multicast correctness not compromised

MVPN Interworking

- Operators choose which PEs become MEGs
  - Choice of deployment scenarios:
    - Make every EVPN PE a MEG, or
    - Put a few MEGs at natural entry/exit points to a DC
    - Or anything in between
  - To get optimal routing/replication between domains, put MEGs along the best path between domains
  - Why not always make all EVPN PEs MEGs?
    - Adding MVPN procedures to EVPN PE brings added complexity in config, procedures, provisioning, etc.
    - Operators need to be able to evaluate the trade-offs and do what is best for their deployment

What about Unicast Routes Needed for RPF/UMH Selection

- MEGs export VPN-IP routes for the multicast sources in EVPN
  - These routes do not necessarily have to be host routes; that depends on the deployment scenario
  - These routes are translated from EVPN unicast routes
- MEGs import VPN-IP routes from L3VPN

EVPN Receivers, MVPN Sources

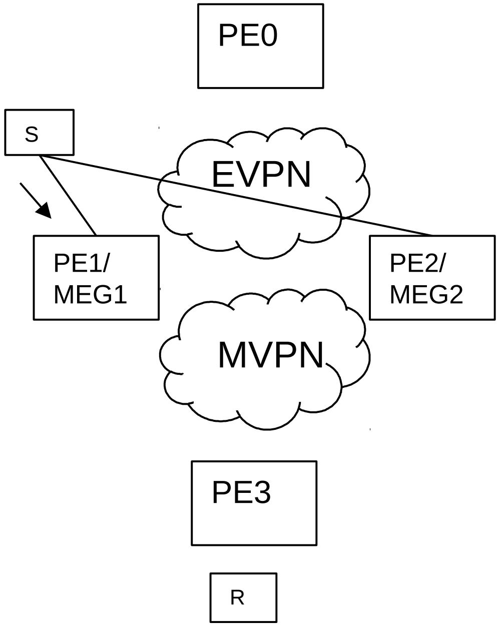
- If an EVPN receiver is attached to a MEG:
  - MEG uses ordinary MVPN procedures to pull the traffic from the MVPN ingress PE
  - Then delivers the traffic to local receivers
- If an EVPN receiver is not attached to a MEG:
  - The PE to which it is attached uses SMET routes to pull the traffic from a MEG
  - The MEG uses MVPN procedures to pull the traffic from the MVPN ingress PE

MVPN Receivers, EVPN Sources

- MVPN egress PE uses ordinary MVPN procedures to pull traffic from the MVPN PE that advertises the best route to the source
    - The ingress PE will be a MEG
    - If the source is not attached to a MEG, the MEG will use OISM procedures to pull the traffic from the real ingress PE
        - In case of ASM, the MEG pulls traffic proactively and send PIM register message to the RP
- There’s an interesting issue when the source is on an all-active multi-homed segment …

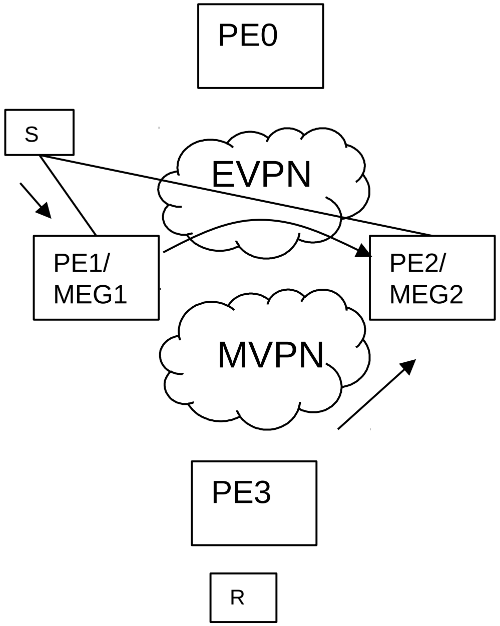
EVPN Sources on All-Active Multi-homed Ethernet Segments I

- Scenario:
  - Source S attaches to EVPN PE1 & PE2, on an all-active multi-homed segment. Both PEs are MEGs.
  - S sends (S,G1) traffic to PE1
  - Receiver R for (S,G1) attaches to MVPN PE PE3
- MVPN requires PE3 to select the ingress PE for (S,G1) traffic.
  - But there is no way for PE3 to figure out which of PE1 or PE2 is the ingress PE



EVPN Sources on All-Active Multi-homed Ethernet Segments II

- What happens if PE3 chooses PE2?
  - PE2 uses EVPN to pull (S,G1) traffic from PE1,
  - PE2 then uses MVPN to push the traffic to PE3
  - So everything still works automatically
    - Because the MVPN and EVPN control planes remain separate
- But won’t PE1 send MVPN *Source Active* routes that force PE3 to choose PE1 as the ingress?
  - No, MVPN nodes do not use the *Source Active* routes to choose the ingress PE.



Still To Do

- Ensure alignment with PIM-Proxy draft
- Ensure alignment with EVPN/IPVPN unicast interoperability draft
- Propose WG adoption