The EVPN Data Center
Multi-tenant Multicast Services

Introduction

In today's data center, EVPN with VXLAN encapsulation (RFC 8365) has become the standard approach for delivering unicast VPN services across a leaf-spine IP fabric, with the ability to achieve an "any server anywhere" model, where servers and services can be dynamically deployed and re-deployed across any leaf node within the fabric, while maintaining optimal forwarding regardless of their location.

Emerging applications in the Broadcast, Financial and Telco industries are now driving demand for point-to-multipoint VPN services across the same EVPN infrastructure, as customers look to deploy video-conferencing, video production, and market data applications. To cater for such demand, Enterprise data centers, as well as emerging Mobile edge computing (MEC) environments now require a standards based EVPN solution, for the efficient and reliable delivery of multicast VPN services. This whitepaper discusses Arista's implementation of a standards based EVPN approach for delivering multi-tenant multicast services, which addresses each of the five main requirements for the successful deployment of business critical, high volume, delay sensitive multicast applications within the data center :

- **Multi-tenancy:** The common model for EVPN is the delivery of multi-tenant unicast services across a shared infrastructure, a multicast forwarding model needs to provide a similar level of functionality, ensuring VRF-aware multicast delivery with support for overlapping multicast groups between tenants sharing the same physical infrastructure.

- **Optimized multicast forwarding:** In an "any server anywhere" deployment model, the forwarding of multicast traffic needs to be achieved in an efficient manner at the first-hop node, eliminating the need for sub-optimal traffic hairpinning, regardless of whether it be a layer 2 or layer 3 multicast flow.

- **Active-Active resiliency:** For resiliency, servers within an EVPN domain are often deployed using multi-homing, any proposed multicast forwarding model therefore has to include support for the dual-homing of multicast sources and receivers within the topology.

- **Fairness and Bandwidth efficiency:** The requirement for multicast is often driven by high-bandwidth applications (video streaming) or latency and delay sensitive applications (market-data) therefore any proposed solution needs to provide a transport mechanism that ensures efficient and fair delivery of multicast traffic.
- **External Domains:** The multicast applications within an EVPN domain will not stand in isolation, the solution must therefore provide the ability to interact with existing multicast sources and receivers which are external to the data center's EVPN domain.

Standards-based Solution

To achieve efficient forwarding of multicast traffic within an EVPN topology, requires the introduction of new EVPN route types to advertise multicast state, a new layer 3 multicast forwarding model to ensure optimal first-hop routing, and a mechanism for effeciently transporting the multicast traffic across the IP infrastructure. The new forwarding model, route types and transport mechanism are defined across multiple IETF standards and drafts which are summarized in the table below:

Draft	Overview
IGMP and MLD proxy for EVPN ***RFC 9251***	Introduction of new Type-6, 7 and 8 EVPN routes for signaling and proxying IGMP/MLD join and leave via BGP EVPN
EVPN Optimized Inter-Subnet Multicast Forwarding (OISM) ***draft-lin-bess-evpn-irb-mcast***	New optimized layer 3 forwarding model for multicast traffic to allow efficient first-hop routing and avoid traffic hairpinning..
Updates on EVPN BUM Procedures ***draft-ietf-bess-evpn-bum-procedure-updates***	PIM Transport Model(s) - Introduce support for Type-9, 10 and 11 routes for Inclusive forwarding and non-IR transport options (PIM, P2MP)

Figure 1: Standard-based Solution

IGMP/MLD Proxy

To provide multicast awareness across the EVPN domain, RFC 9251 introduces a mechanism for the proxying of locally received MLD and IGMP joins on a VTEP, and the signaling of the interested receiver(s) via a new type-6 Selective Multicast Ethernet Tag (SMET) route. The proxy functionality reduces the number of SMET routes advertised across the EVPN domain, by aggregating the local joins for a specific group into a single SMET route advertisement. To support dual-homed multicast hosts (receivers and sources), the functionality also operates within an EVPN All-Active topology, with the introduction of a new type-7 (IGMP/MLD join sync) and type-8 (IGMP/MLD leave sync) routes. The type-7 (join-sync) route is used to synchronize locally received joins between VTEP peers on the same shared ethernet segment (ES). Similarly the type-8 route (leave-sync) is used to synchronize locally received leave messages between VTEP peers on the same shared ethernet segment (ES).

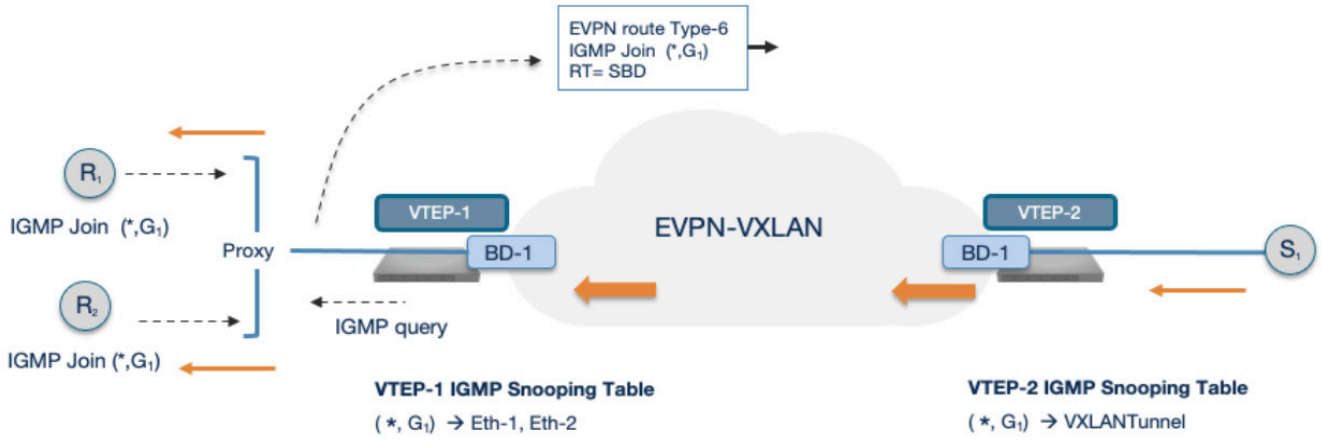

Figure 2: EVPN IGMP proxy functionality with type-6 SMET route

EVPN Optimized Inter-Subnet Multicast (OISM)

The OISM draft utilizes the new routes 6,7 and 8 routes to provide optimized inter-subnet multicast forwarding. The draft introduces the concept of a distributed multicast Designated Router (DR) across all VTEP nodes within the EVPN domain. Where each VTEP acts as a multicast DR for any attached subnets, and is responsible for routing multicast traffic onto the subnet of any locally interested receiver. To facilitate this distributed egress routing model, the multicast traffic is 'VXLAN bridged" from the source to the remote VTEPs via a supplementary bridge-domain (SBD), which all VTEPs in the VRF are members of, with the VTEPs signaling interest in a multicast flow via type-6 advertisements. With this distributed DR approach, the multicast traffic is always routed at the first-hop, for any receivers local to the source, avoiding the need to hairpin traffic across the EVPN domain to a centralized DR for inter-subnet forwarding. For any interested remote receivers, traffic is 'VXLAN bridged" and the remote VTEP acts as the DR for inter-subnet routing to any locally attached receivers.

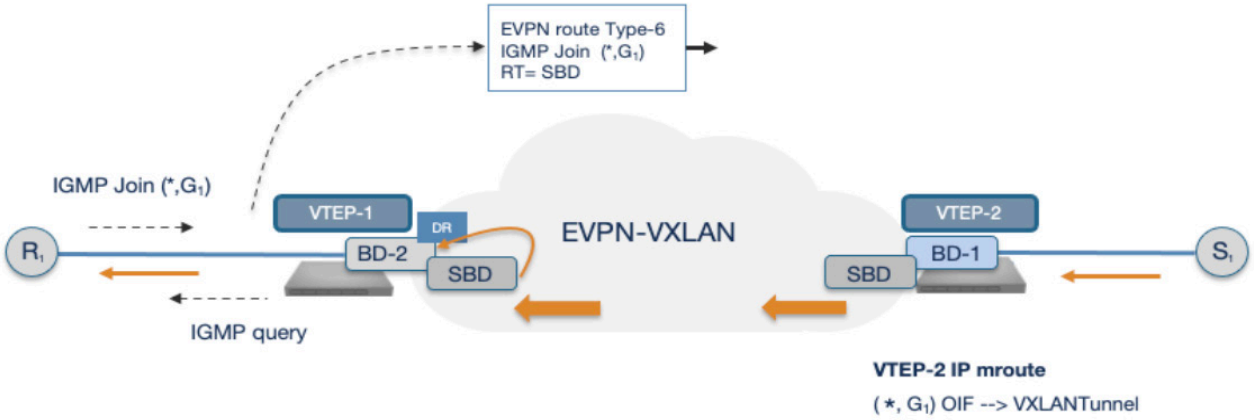

Figure 3: EVPN optimized inter-subnet multicast (OISM)

EVPN BUM procedures

The EVPN multicast model provides a number of options on how the tenant multicast traffic can be VXLAN encapsulated and transported to the remote VTEPs, which have signaled interest in a group via the type-6 route advertisement. The various transport models are summarized below, along with their benefits and drawbacks which will be dependent on the specific application and deployment model.

Ingress Replication

The simplest approach to transporting the overlay multicast traffic is ingress replication (IR). In this model the VTEP attached to the multicast source is responsible for VXLAN encapsulating the multicast flow and forwarding a unique unicast VXLAN encapsulated packet to each of the interested remote VTEPs. While the IR approach removes the need for any additional state or multicast protocol in the IP underlay, it's not an efficient method for high bandwidth multicast flows, as a copy of the flow needs be created by the ingress VTEP for each of the interested remote VTEPs, thus a 1Mbs flow with ten interested remote VTEPs, would result in 10Mbs of bandwidth on the uplink from the ingress VTEP. As the number of interested receivers increases the overhead on the ingress VTEP will linearly increase as each replicated copy will require a unique VXLAN header rewrite. The requirement for a header rewrite for each interested remote VTEP will also result in a time delta between when the first VTEP and last interested VTEP receives the multicast flow. This resultant asynchronous delivery model can have a detremental effect on delay-sensitive multicast applications, which require fair delivery across all receivers regardless of their location. The IR approach is therefore an appropriate model for transporting low to medium bandwidth multicast applications, when the number of interested VTEPs in the flow is low and fairness of delivery is not a key requirement.

Assisted Replication

Assisted Replication. To alleviate the replication processing on the ingress VTEP, a second option termed Assisted Replication (AR) is define in the IETF draft (bess-evpn-optimized-ir), providing the ability for the ingress VTEP to off-load the replication process to an AR-replicator node. In this model the AR-replicator(s) are typically spine nodes in a leaf-spine topology, with AR support enabled on the leaf nodes. The leaf nodes would send a single copy of the multicast traffic to the advertised AR-replicator, and the AR-replicator takes the responsibility of performing the replication to all the interested remote VTEPs. With the capability to support multiple AR-replicator nodes in the topology, the replication process can be load-balanced across different spine nodes in the topology.

The AR approach will alleviate the load on the ingress VTEP, and also reduce the bandwidth consumption on the uplink to the spine (AR-replicator), as only a single copy of the multicast traffic is forwarded by the source VTEP. However, the processing overload of the IR model is not being eliminated; it is only being concentrated on a smaller number of spine node(s) rather than distributed across all leaf nodes, under the assumption that the spine will offer better IR performance than the leaf node. This assumption can't always be guaranteed, typically in small or medium sized deployment the leaf and spine nodes are the same fixed configuration platforms. The AR approach, like IR, will also still result in asynchronous delivery of the multicast traffic, as it still requires a VXLAN header rewrite on the AR-replicator node, making the approach a sub-optimal solution for multicast applications where fairness of delivery is a key requirement.

The AR model like the IR approach, would therefore be an appropriate model for transporting low to medium bandwidth multicast applications, when the number of interested VTEPs in the flow is low and fairness of delivery is not a key requirement. The model itself will only offer an obvious performance benefit over the IR approach, when the Spine nodes provide a marked IR performance improvement over the leaf nodes or when the number of interested receivers are known to be spread across a wide range of VTEPs.

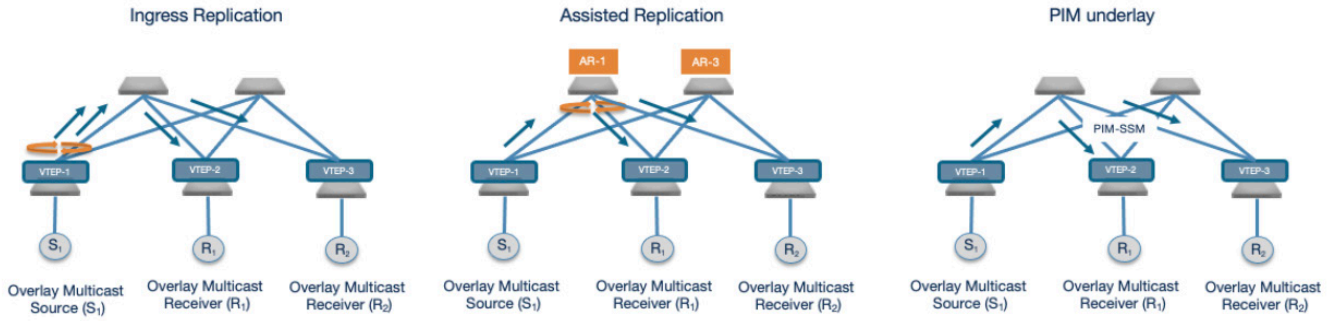

Figure 4: EVPN multicast transport models

PIM underlay

To address the potential drawbacks of both the IR and AR approaches when delivering high-bandwidth and delay-sensitive multicast applications a third approach is a Point-to-MultiPoint (P2MP) transport model using PIM. In this approach, PIM is enabled in the network underlay, and VTEPs with interested overlay receivers join an associated PIM underlay multicast group advertised by the VTEP attached to the source. A new EVPN type-10 (S-PMSI AD) route is used by the source VTEP to advertise the underlay multicast group that will be used to transport the overlay multicast flow. The VTEP attached to the source is then responsible for VXLAN encapsulating the tenant's multicast flow, with the destination address being the advertised underlay multicast group. The packet is then transported across the leaf-spine topology using the shortest-path multicast tree (SPT) to the interested VTEPs.

With this approach, the ingress VTEP only sends a single multicast packet on the uplink, so there is no additional bandwidth on the uplink or replication overhead on the ingress VTEP, addressing the concerns of an IR approach. There is also no additional overhead on the spine node, unlike the AR approach which has the overhead of re-writing the VXLAN IP header for each individual interested VTEP. In the PIM model, the VXLAN IP header is unchanged, and the Spine follows standard multicast forwarding mechanisms, copying the original VXLAN packet to the multiple interested VTEPs. This forwarding of the VXLAN packet to the interested VTEPs is done in parallel, so the multicast flow is delivered synchronously, ensuring all VTEPs receive the stream at the same time. The model therefore provides the best solution for delivering high-bandwidth, latency sensitive multicast traffic within a data center leaf spine topology. The table below provides a summary of the different transport models and their Pros and Cons:

	Ingress Replication (IR)	Assisted Replication (AR)	PIM in the underlay
Leaf and Spine node processing efficiency	Low, ingress VTEP required to create a unique packet for each interested remote VTEP.	Medium, ingress VTEP only required to forward a single copy to the AR-Replicator (Spine). However the AR-Replicator is required to re-write the VXLAN IP header for each interested remote VTEP.	High, ingress VTEP only required to forward a single copy. Spine forwards the original VXLAN packet to the interested VTEPs, no packet rewrite required.
Bandwidth consumed for multicast replication on Leaf	High, required to create copy of the multicast flow for every interested remote VTEP	Low, only a single copy of the multicast packet is forwarded on the uplink from the source VTEP.	Low, only a single copy is forwarded on the uplink from the source VTEP.
Fairness/latency sensitive	Low, due to the requirement to re-write the header of each packet there will be asynchronous delivery on the ingress PE. The fairness of delivery and latency delta will deteriorate further with the number of interested remote VTEPs	Low, due to the requirement to re-write the header of each packet there will be asynchronous delivery on the Spine. The fairness of delivery and latency delta will will deteriorate further with the number of interested remote VTEPs	Good, synchronous delivery across all nodes in the path. Parallel forwarding on the Spine no re-write to ensure all interested VTEPs receive the flow at the same time.

EOS implementation of EVPN multicast

Arista's EOS software provides support for EVPN multicast based on the aforementioned IETF standards, for deployment within the data center and across an EVPN campus environment. The functionality includes support for IGMP proxy and therefore the advertisement of type-6 (SMET) routes, all-active multi-homing of multicast sources and receivers with support for the new type-7 and 8 routes, and the implementation of the OISM draft for optimized routing of multicast flows. To provide support for deployment scenarios where the multicast applications are low-bandwidth and therefore don't require the added complexity of running PIM in the network underlay, the implementation supports the option of running an ingress replication (IR) transport model. When high bandwidth multicast applications are deployed requiring synchronous delivery, which would be the case for video-conference, IP-TV and market data applications, the implementation also supports a PIM transport model. This document focuses on the PIM transport model, however, it should be noted the EVPN control plane described in the document is also relevant for an ingress replication model.

Overlay to underlay multicast group mappings

In the PIM transport model, there is a need to map the overlay multicast groups within a VRF or bridge-domain to a PIM signaled multicast underlay group, which would be used to transport the VXLAN encapsulated multicast streams across the network underlay. The multicast group mapping is dependent on the forwarding mode; layer 3 inter-subnet multicast routing within a VRF, or layer 2 only multicast bridging within a VNI. For the layer 2 only mode, multicast traffic is VXLAN encapsulated and transported using an associated multicast underlay group per bridged-domain, thus providing an N:1 overlay to underlay mapping model per bridge-domain (VNI). For routed multicast within a VRF, to provide greater granularity three different mappings models can be used in combination within the same VRF.

- N:1 mapping: All overlay groups within the VRF are mapped to a single underlay multicast group.
- N:M mapping: Overlay groups within the VRF are dynamically mapped based on activity to a defined range of underlay multicast groups. If the range is oversubscribed, fall back to the defined N:1 mapping for the VRF.
- 1:1 mapping: An overlay multicast group is mapped directly to a specific underlay group.

The flexibility to support different mapping models within a VRF, provides the capability to optimize the fabric bandwidth while minimizing the PIM underlay state. For example in the case of the N:1 mapping model, where all the overlay groups within a VRF are mapped to single underlay group, this will mean a VTEP with an interested receiver for a specific group within the VRF will receive all active multicast streams within VRF, with the VTEP decapsulating the VXLAN streams and only forwarding the stream(s) where there is an interested local receiver, all other streams being dropped by the VTEP. Therefore in this model, all active streams within the VRF are forwarded across the fabric to any VTEP with an interested receiver and selectively dropped on egress. However, while this mapping model will mean additional bandwidth being used across the fabric, it will result in minimal PIM state in the network underlay, as only one underlay group is being used to transport all the overlay groups in the VRF. The opposite is true for the 1:1 mapping model, with this configuration only VTEPs with an interested receiver for that specific group within the VRF will receive the overlay multicast stream, thus optimizing the fabric bandwidth as the multicast VXLAN traffic is only sent to VTEPs with interested receivers. However, while the model results in a reduction in fabric bandwidth, it will mean an increase in PIM state in the network underlay, as a new (S,G) underlay entry needs to be created for each active overlay group in the VRF.

The different overlay to underlay mapping models for VXLAN encapsulating the overlay multicast traffic with a PIM transport are highlighted in the figure below:

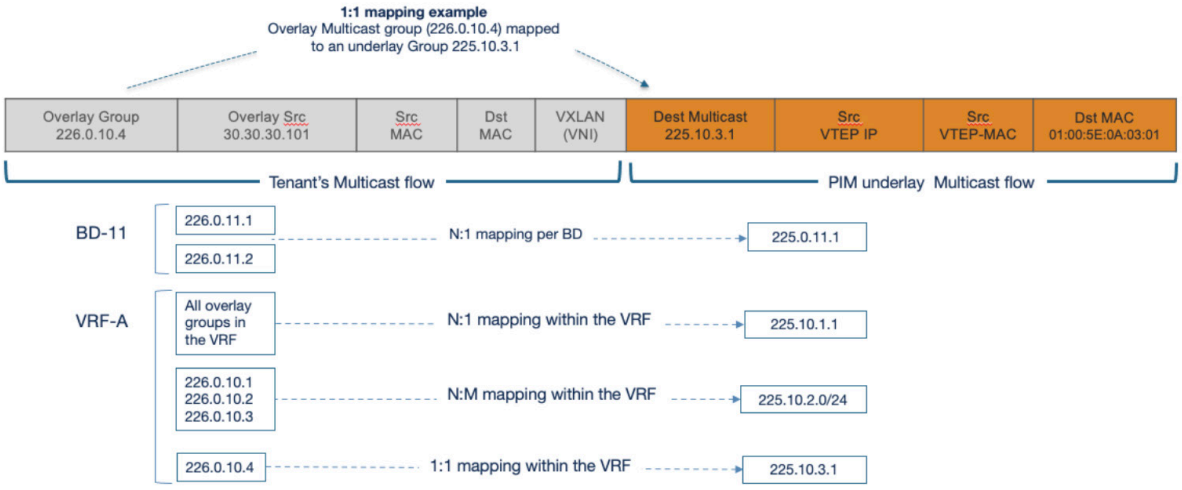

Figure 5: EVPN multicast group overlay to multicast underlay group mapping model

As the different mapping models provide their own unique benefits whether it be less PIM state or optimized fabric bandwidth, the appropriate model to deploy for specific streams will be dependent on the distribution of receivers and sources for the stream within the fabric. If multicast receivers for a specific set of streams within the VRF are distributed across all VTEPs in the fabric, then bandwidth is unlikely to be concern, as all VTEPs are required to receive the streams, therefore the N:1 model would offer the benefit of reducing the underlay PIM state. Alternately, if there are only a small number of streams in the VRF and they are high bandwidth, with only a subset of VTEPs having interest in the streams, then the 1:1 model would offer the benefit of optimizing the bandwidth utilization, with the increased PIM being less of concern due to the small number of streams. In a live environment, there is more likely to be a mix of both high-bandwidth streams and distributed receivers, within a single VRF, therefore the ability to use a combination of all mappings models becomes essential in order to minimize the PIM state while optimizing the fabric bandwidth.

Deployment examples

Arista's introduction of EVPN multicast functionality, along with configurable overlay to underlay group mappings, provides the ability to achieve a bandwidth efficient model for delivering both layer 2 and 3 multicast VPN services within the data center, some of the typical use cases for the functionality and its benefits are summarized in the following sections.

Use case 1: Resilient tenant Layer 2 multicast delivery

In this deployment scenario there is a requirement to bridge multicast traffic for a tenant hosted within an EVPN domain. The figure below illustrates the topology, where tenant-A has a dual-homed source attached to VTEP 1 & 2 with an interested dual-homed receiver in the same VLAN on VTEP 3 & 4. In the topology the dual-homing of the multicast source and receiver is achieved using EVPN All-Active (A-A).

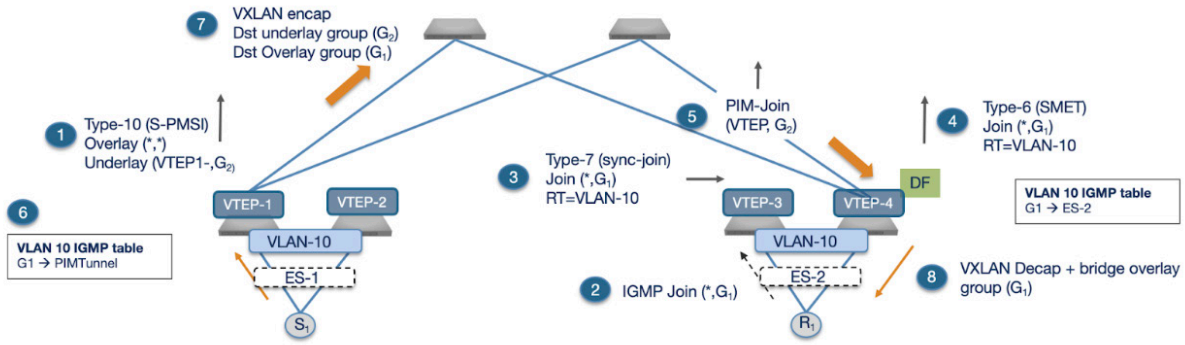

Figure 6: EVPN layer 2 multicast bridging forwarding model

The individual steps to advertise the underlay group mappings, and the receive the associated overlay multicast stream using an EVPN control plane, are outlined below.

1. Each VTEP advertises a type-10 (S-PMSI AD) route, carrying the underlay to overlay multicast mapping for the VLAN. In the case of the example topology, the advertised mapping for VLAN 10 for all VTEPs is (*.*-> VTEP, G2), meaning all multicast traffic in the VLAN will be transported using the multicast underlay group G2 with a source address equal to the VTEP that is VXLAN encapsulating the multicast stream.
2. The local receiver R1 signals its interest in the overlay multicast group (G1) via an IGMP join, which due to load-balancing on the port-channel is received by VTEP-3.
3. To synchronize the IGMP state between the peers of the shared ethernet segment (ES-2), VTEP-3 advertises a EVPN type-7 (sync-join) route.
4. As the elected designated forwarder (DF) for VLAN 10 on the ethernet segment (ES-2), VTEP-4 is responsible for advertising the EVPN type-6 (SMET) route in response to receiving the type-7 join-sync route for the ESI from VTEP-3.
5. With an interested receiver for the group attached to the ESI, VTEP 3 & 4 will now join the associated underlay groups advertised in the type-10 routes from each VTEP for VLAN 10.
6. The type-6 advertisement is received by both VTEP 1 & 2 and imported into their local IGMP snooping table for VLAN 10, this is based on the route-target (RT) of the type-6 advertisement. The IGMP snooping table entry would be for the multicast overlay group G1, with a receiver learnt across a VXLAN PIM tunnel.
7. With an interested receiver for the overlay group (G1), VTEP-1 on receiving the multicast flow from (S1), will VXLAN encapsulated the packet with the advertised underlay multicast group for VLAN 10 (G2) as the destination IP, which is transport via the PIM enabled underlay to both VTEP 3 & 4.
8. VTEP-4 as the elected DF for VLAN 10, on receiving the VXLAN packet, removes the VXLAN header and based on its local IGMP snooping table forwards the packet to the interested receiver (R1) on the ES. VTEP-3 as the non-DF for the VLAN drops the received VXLAN frame.

In this model:

- Multicast sources and receivers in the same bridge-domain can be deployed on any VTEP within the EVPN domain, with the ability to both single or dual-homed any source or receiver.
- The layer 2 multicast traffic in the overlay is only received by VTEPs with interested receivers, and forwarded to local ports on the VTEP based on the local IGMP snooping table.
- In this layer 2 only model, the mapping of overlay multicast groups to separate underlay multicast groups is achieved at a VLAN level, providing support for overlapping groups across different layer 2 VLANs within a single shared EVPN domain.
- The multicast overlay traffic is transported efficiently using PIM in the underlay, where only a single copy is sent by the source VTEP regardless of the number of VTEPs interested in the flow. The spine node in the topology is responsible for copying the multicast packet to each interested leaf node, there is no need for the spine to re-write the outer IP header for each interested VTEP this ensures fair and synchronous delivery of the multicast flow to all interested VTEPs.

Use case 2: Resilient tenant Layer 3 multicast delivery

In this deployment there is a requirement to provide tenant aware multicast routing, with the capability to again support overlapping multicast groups between tenants. A typical topology to meet these requirements using the OISM forwarding model is illustrated in the figure below, where tenant-A has a dual-homed source attached to VTEP 1 & 2 in subnet 10 (sub-10) with an interested dual-homed receiver in subnet 11 (sub-11) attached to VTEP 3 & 4. To provide connectivity between the two subnets, a common supplementary bridge-domain (SBD) is configured across all the VTEPs in the tenant's VRF, this is used to associate received VXLAN multicast traffic with the tenant's VRF. In the topology the dual-homing of the multicast source and receiver is achieved using EVPN all-active (A-A) multi-homing.

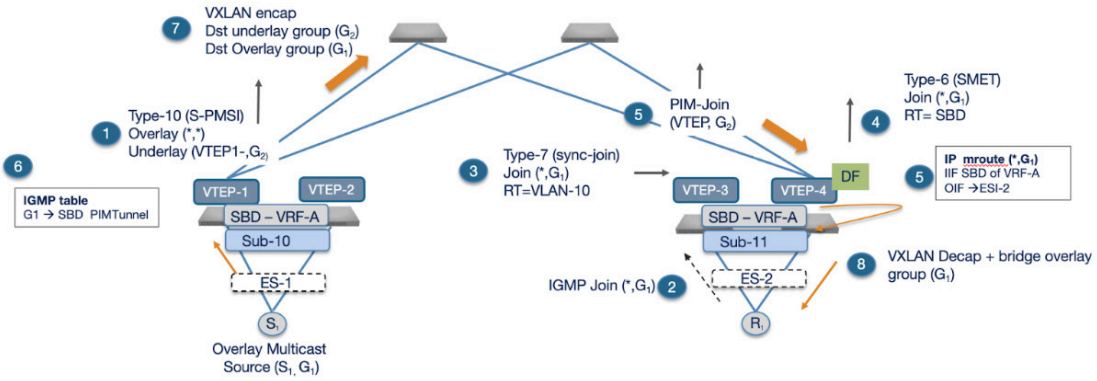

Figure 7: EVPN layer 3 multicast routing forwarding model with OISM

Although the example only illustrates a layer 3 multicast forwarding model and control-plane, the functionality can be used in conjunction with the aforemented layer 2 multicast model, to support both bridging and routing of multicast streams across the EVPN domain. The control and forwarding plane for the layer 3 multicast model is outlined in the following steps.

1. Each VTEP advertises a type-10 route, carrying the underlay to overlay multicast group mapping for the VRF. In the case of the example topology, for simplicity a single underlay multicast group (G2) is configured for all overlay groups in the VRF. Thus the type-10 route advertisement for the VRF would be (*.*-> VTEP, G2), meaning all multicast traffic in the VRF will be transported using the multicast underlay group G2 with a source address equal to the VTEP that is VXLAN encapsulating the multicast flow.

2. The local receiver R1 signals its interest in the overlay multicast group (G1) via an IGMP join, which due to load-balancing on the port-channel is received by VTEP-3.

3. To synchronize the IGMP state between the peers of the shared ethernet segment (ES-2), VTEP-3 advertises a EVPN Type-7 (sync-join).

4. As the elected designated forwarder (DF) for VLAN-11(sub-11) on the ethernet segment, VTEP-4 is responsible for advertises the type-6 (SMET) route in response to receiving the type-7 join-sync route. The advertised SMET route in this layer 3 OISM model is advertised with the RT of the tennant's supplementary bridge-domain (SBD).

5. With an interested receiver for the group attached to the ESI, VTEP 3 & 4 will now join the associated underlay groups advertised in the type-10 routes from each VTEP for the VRF. The VTEPs will also add an OIF entry for sub-11 to the IP mroute table of the VRF along with an IIF entry which will be the SBD for the VRF.

6. The type-6 advertisement is received by both VTEP 1 & 2 and imported into the IGMP snooping table for VLAN 10, with the receiver learnt across a VXLAN PIM tunnel. The route is imported based on the route-target (RT) of the type-6 route which is the RT of the common SBD for the tenant's VRF. In the OISM forwarding model, multicast traffic is VXLAN bridged from the source's VLAN/VNI to the interested remote VTEPs, hence the reason why the type-6 route is imported into the IGMP snooping table and not the mroute table of the VRF.

7. With an interested receiver for the overlay group (G1), VTEP-1 on receiving the multicast flow from (S1), will VXLAN encapsulated the packet with the advertised underlay multicast group for the VRF (G2) as the destination IP which is used to transport the VXLAN packet in the PIM enabled underlay to VTEP 3 & 4.

8. VTEP-4 as the elected DF for VLAN 11 (Sub-11), on receiving the VXLAN packet, it removes the VXLAN header and routes the multicast traffic based on the VRFs mroute table to the interested receiver (R1) on the ES. VTEP-3 as the non-DF for the VLAN drops the received VXLAN frame.

In this model:

- Multicast sources and receivers in the same or different subnets can be deployed on any VTEP, with the ability to both single or dual-homed any multicast source or receiver

- Both layer 2 and 3 multicast traffic in the EVPN overlay is only received by VTEPs with interested receivers in the VRF, with the forwarding on local ports of the VTEP based on the IGMP snooping table and IP mroute table of the VRF.

- The routing of the multicast traffic is achieved on the egress VTEP, using a distributed multicast DR model across all VTEPs. This means only a single copy of the flow needs to be sent to any interested VTEP, where the egress VTEP is responsible for bridging and routing the stream to any local receiver. This forwarding model ensures optimal multicast forwarding, while avoiding any traffic hairpinning.

- The mapping of tenant overlay multicast groups to separate underlay multicast groups at a VRF level, provides support for overlapping multicast groups across VRFs within a single EVPN domain.

- The multicast overlay traffic is transported efficiently using PIM in the underlay, where only a single copy is sent to the Spine regardless of the number of VTEPs interested in the flow, with the Spine forwarding the packet without the need to re-write the IP header for each individual interested VTEP ensuring fair and synchronous delivery of the multicast flow to all interested VTEPs.

Use case 3: Resilient tenant Layer 3 multicast with external receivers/sources

This deployment scenario is an extension of the previous use cases, with the addition of the multicast receivers within the EVPN domain being interested in a multicast source residing outside the EVPN domain. An example topology to meet these requirements is illustrated in the figure below where tenant-A (VRF-A) has a dual-homed receiver attached to VTEPs 3 & 4 in VLAN 11 (Sub-11) with interest in a multicast source in the external PIM network. The PIM network is connected to the EVPN domain via an option-A hand-off between VTEP-1 and the PIM router, where VTEP-1 has a PIM neighbourship within the VRF with the PIM router. For simplicity, the PIM router is also acting as the RP for the PIM network. In this type of topology VTEP-1 is termed the PIM edge Gateway (PEG) for the EVPN domain.

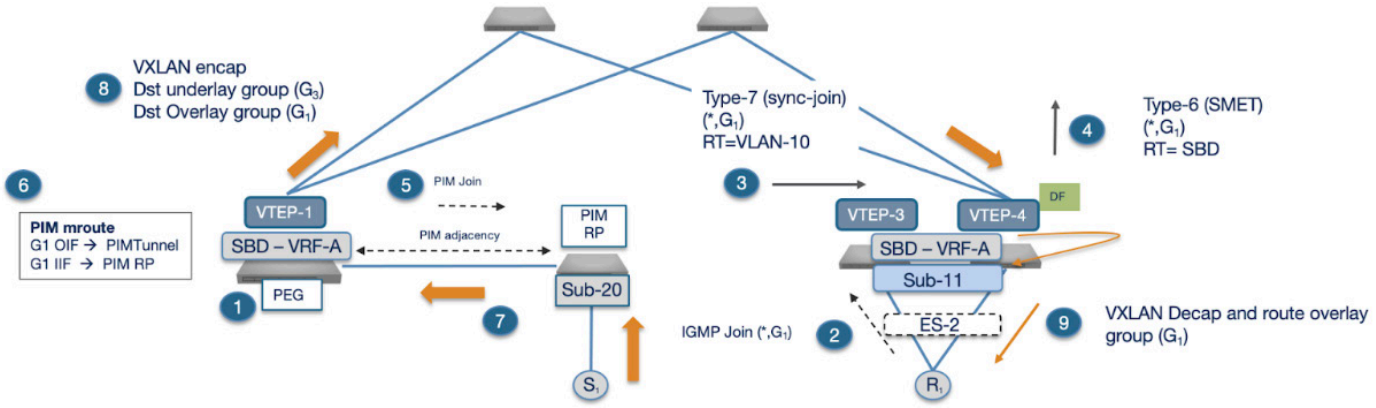

Figure 8: EVPN layer 3 multicast routing forwarding model with OISM and external PIM domain

Note: Each VTEP again advertises a type-10 route, carrying the underlay to overlay multicast group mapping for the VRF. In the case of the example topology, for simplicity a single underlay multicast group (G2) is configured for all overlay groups in the VRF. Thus the type-10 route advertisement for the VRF would be (*.*-> VTEP, G2), meaning all multicast traffic in the VRF will be transported using the multicast underlay group G2 with a source address equal to the VTEP that is VXLAN encapsulating the multicast flow. To simplify the overall diagram, this step is omitted from the figure above.

The individual steps to join the underlay group within the EVPN domain and receive the multicast stream from the external source (S1) residing in the PIM network are as follows:

1. VTEP-1 is configured with a PIM interface in the tenant's VRF connecting to the external PIM router and its associated RP. With this configuration VTEP-1 advertises its PEG capability via an external community with its IMET route advertisement. As the only PEG router in the EVPN domain it is elected the PEG DR router for the VRF. The elected PEG DR is responsible for sending PIM joins for receivers within the EVPN domain, to the upstream RP.

2. The local receiver R1 signals its interest in the overlay multicast group (G1) via an IGMP join, which due to load-balancing on the port-channel is received by VTEP-3.

3. To synchronize the IGMP state between the peers of the shared ethernet segment (ES-2), VTEP-3 advertises a type-7 (sync-join) route.

4. As the elected designated forwarder (DF) for VLAN-11(sub-11) on the ethernet segment, VTEP-4 is responsible for advertising the type-6 (SMET) route in response to receiving the type-7 route. The SMET route is advertised with the RT of the supplementary bridge-domain (SBD) of the VRF.

5. The type-6 advertisement is received by VTEP-1 and imported based on the route-target (RT), which is the RT of the common SBD for the VRF. As the elected PEG DR for the VRF, VTEP-1 in response to receiving the SMET route will also send a PIM join upstream to the RP for the multicast group (*,G1).

6. At this point an mroute entry in the VRF will be created for the overlay group with an IIF pointing to the RP and OIF entry of the SBD.

7. The RP on receiving the PIM join adds VTEP-1 as an OIF entry for the multicast group (*,G1). On the assumption that S1 is already forwarding multicast traffic for the group, the flow will now be forwarded downstream to VTEP-1 via the RP-tree.

8. VTEP-1 on receiving the multicast flow, will VXLAN encapsulated the packet with the advertised underlay multicast group for the tenant's VRF (G3) as the destination IP, which is used to transport the VXLAN packet in the PIM enabled underlay to VTEP 3 & 4.

9. VTEP-4 as the elected DF for VLAN 11 (Sub-11), on receiving the VXLAN packet, removes the VXLAN header and routes the multicast flow to the interested receiver (R1) on the ES. VTEP-3 as the non-DF for the VLAN drops the received VXLAN frame.

In this model:

- Multicast sources and receivers in the same or different subnets can be deployed on any VTEP, with the ability to single or dual-homed both sources and receivers if required.
- The sources and receivers can reside both within the EVPN domain and within an external PIM network
- The routing of multicast traffic is achieved on the egress VTEP, using a distributed multicast DR model across all VTEPs, ensuring optimal multicast routing and bridging avoiding the need for any traffic hairpinning within the topology.
- Routing of external multicast flows into the EVPN domain, follows a standard PIM model using the PEG node, where the PEG will register interest in the flow via the PIM RP when receiving type-6 advertisements and switch over to the SP-Tree for delivery across the PIM domain if required.
- Multicast routing to external receivers, also follows a standard PIM model using the PEG node, which forwards the tenant flows via the RP-Tree initially, with the capability to switch to an SP-Tree if required.
- The mapping of tenant overlay multicast groups to separate underlay multicast groups at a VRF level, provides support for overlapping multicast groups across VRFs within the EVPN domain.
- The multicast overlay traffic is transported efficiently using PIM in the underlay, where only a single copy is sent to the Spine regardless of the number of PEs interested in the flow, with the Spine forwarding the packet without the need to re-write the IP header for each individual interested VTEP ensuring fair and synchronous delivery of the multicast flow to all interested VTEPs

Conclusion

EVPN has become the standard model for delivering multi-tenant unicast services within the data center, with the maturity of EVPN there is now an evolving demand for point-to-multipoint VPN services as customers look to deploy video-conferencing, video production, IPTV and market data across the same infrastructure.

Networking vendors have historically looked to address this multicast demand by introducing proprietary solutions to work alongside EVPN, however, the drive for openness, alongside the scale and performance requirements of the multicast applications, has placed major restrictions on these proprietary approaches. By delivering an open and standards based EVPN multicast solution, Arista addresses these historical concerns regarding vendor lock-in, by providing an open standards based solution that supports the high-bandwidth, delay sensitive multicast applications typically seen in the data center. The benefits of this innovative Arista approaches include:

- Open standards based EVPN solution, no vendor lock-in
- Fair and synchronous delivery for high bandwidth and latency sensitive multicast applications, with a PIM transport model
- Optimized first-hop multicast routing, no traffic hairpinning within the EVPN fabric
- Resilient EVPN All-Active multihoming of both multicast sources and receivers.
- Support for overlapping multicast groups between tenants sharing the same EVPN domain
- Seamless integration with existing multicast PIM domains

Reference Material

https://datatracker.ietf.org/doc/html/rfc9251

https://tools.ietf.org/html/draft-ietf-bess-evpn-irb-mcast-04

https://tools.ietf.org/html/draft-ietf-bess-evpn-bum-procedure-updates-05

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office 1390
Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062

Copyright © 2023 Arista Networks, Inc. All rights reserved. CloudVision, and EOS are registered trademarks and Arista Networks is a trademark of Arista Networks, Inc. All other company names are trademarks of their respective holders. Information in this document is subject to change without notice. Certain features may not yet be available. Arista Networks, Inc. assumes no responsibility for any errors that may appear in this document. 03/23