The EVPN Data Center

EVPN Gateway for hierarchical multi-domain EVPN and DCI

Introduction

In today's data center, EVPN with VXLAN encapsulation (RFC 8364) has become the de-facto standard for delivering VPN services across a shared leaf-spine IP infrastructure, where servers and applications can be dynamically deployed and re-deployed on-demand. The maturity of EVPN and the benefits it provides, has also seen its footprint grow within the data center and across data centers for service continuity. This dramatic growth raises a number of scaling, fault containment, and Data Center Interconnect (DCI) challenges, when attempting to provide seamless layer 2 and 3 EVPN services. Networking vendors have historically attempted to address the scaling and DCI challenges with proprietary solutions that work alongside EVPN, while providing a level of benefit, they have resulted in additional complexity, hardware dependency and a vendor lock-in. This whitepaper discusses Arista's implementation of a standards based EVPN GW solution, to deliver hierarchical multi-domain EVPN designs for improved scale and fault-containment both within the data center and across data centers as part of a DCI solution to provide seamless EVPN connectivity across the WAN.

Standard based solution

Modern leaf-spine fabrics can scale to support 100s of leaf nodes within a single fabric, however, in the context of an EVPN environment, this single flat scaled-out fabric approach can present a challenge, due to the amount of EVPN state involved (MAC-IP routes, IP-prefixes, flood-lists, next-hops etc), while affecting convergence times and the blast radius of any failure event. To achieve the required level of scale without linearly growing EVPN state on each node within the fabric, Arista's EVPN GW solution introduces hierarchy into the end-to-end EVPN design while maintaining the capability to deliver any service anywhere across the topology.

With Arista's EVPN GW solution, an EVPN topology is divided into multiple self-contained EVPN domains, while maintaining seamless layer 2 and 3 VPN connectivity between the domains. This multi-domain hierarchical approach not only improves scale by reducing the amount of state held within each individual domain, but also ensures better fault-containment and improved convergence by reducing the amount of state churn between domains. Further to ensure service continuity in the event of a data center failure, this multi-domain approach can be extended across sites, to form a DCI, allowing the extension of EVPN services between DCs, with the ability to support both VXLAN and MPLS encapsulations when transversing the WAN.

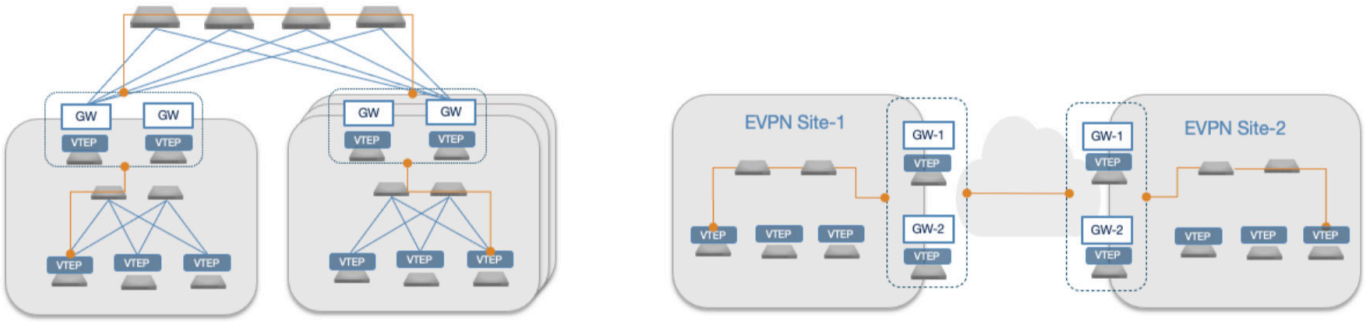

Figure 1: EVPN GW for multi-domain EVPN designs within the DC and inter-DC as a multi-domain DCI

To ensure vender interoperability, which can be a critical requirement when interconnecting EVPN domains both within the data center and across data center's, Arista's EVPN GW is built using a series of of ratified IETF standards, summarized in the table below:

iETF standard/Draft	Overview
BGP MPLS-Based Ethernet VPN **RFC 7432**	EVPN control plane for L2 VPNs with MPLS encapsulation.
A Network Virtualization Overlay Solution using EVPN **RFC 8365**	EVPN control plane for L2 VPNs with an NVO environment with VXLAN encapsulation
Interconnect Solution for EVPN Overlay networks ***RFC 9014***	EVPN Gateway for Layer 2 interop between EVPN-VXLAN and EVPN-/MPLS domains, along with VPLS and PBB.
EVPN Interworking with IPVPN ***evpn-ipvpn-interworking***	Gateway procedures for providing Layer 3 interconnect between domains (VXLAN and MPLS)

Based on the standards outlined in the table, Arista's GW implementation looks to address two EVPN multi-domain scaling and DCI deployment scenarios.

EVPN GW with EVPN-VXLAN interconnect

In this topology the EVPN GW is deployed to provide hierarchy when interconnecting Layer 2 and 3 VPN services across different EVPN-VXLAN domains within the same data center or across data centers interconnected via an IP backbone. This deployment model can be defined as an "EVPN-VXLAN-to-EVPN-VXLAN" gateway as the control-plane between the domains is EVPN and the data-plane encapsulation between the domains is VXLAN.

EVPN GW with EVPN-MPLS interconnect

When interconnecting EVPN-VXLAN domains, certain DCI deployment scenarios will require the EVPN GW to interconnect via an MPLS backbone, where the MPLS transport in the backbone is being utilized for historical reasons or to provide traffic engineering and fast reroute functionality. For this scenario, there are additional requirements for the EVPN GW node, as it will now be required to translate between VXLAN and MPLS forwarding planes along with the EVPN-VXLAN and EVPN-MPLS control-planes. This EVPN GW deployment model can be defined as an "EVPN-VXLAN-to-EVPN-MPLS" gateway as the control-plane between the two domains is EVPN with a data-plane encapsulation of VXLAN and MPLS in each domain.

EVPN GW with EVPN-VXLAN interconnect

In this scenario, for scale, fault-containment or DCI reasons, multiple EVPN-VXLAN domains are deployed within or across data centers via an IP interconnect, with a requirement to provide layer 2 and 3 VPN services across the domains. To address these requirements, GW nodes would be deployed within each EVPN domain. To provide hierarchy when interconnecting the domains, the GW nodes are defined with a local EVPN peering domain; for learning EVPN routes from VTEP nodes in the local domain and a remote EVPN peering for learning EVPN routes from GW nodes connected to remote EVPN domains.

The concept of local and remote EVPN domains on the GW nodes is used to scope and aggregate the EVPN routes advertised between the domains. Only type-2 (MAC-IP) and type-5 (IP-prefix) routes are advertised between domains, type-1 (A-D routes), type-3 (IMET routes) and type-4 (ES routes) have domain level scope and are not re-advertised across domains by the GW. Thus reducing the amount of EVPN state that is advertised between domains, while maintaining layer 2 and 3 VPN connectivity across domains

When the type-2 and type-5 routes are re-advertised by a GW into the remote domain, the next-hop of the route is changed to the GW's IP, any equivalent EVPN route received from the remote domain is re-advertised into the local domain again with the next-hop changed to the local GWs IP address.

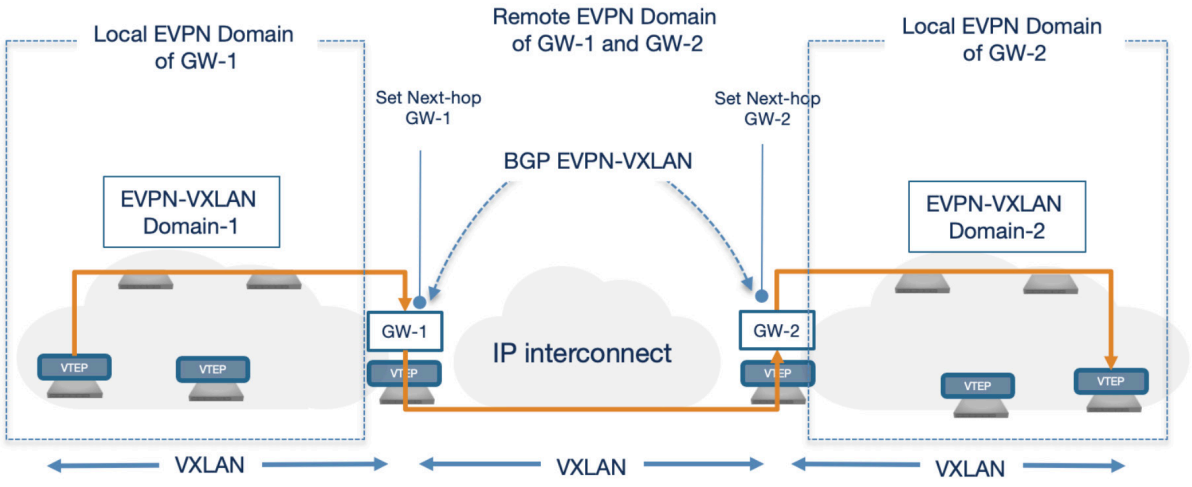

Figure 2: EVPN-VXLAN GW with IP interconnect

With this approach, nodes within an EVPN domain only learn the next-hop of the neighboring nodes in the same domain and their local GW(s), so there is no need for end-to-end IP underlay connectivity across domains. The GW node(s) learn the next-hop of all nodes in their own local domain and the next-hop of the remote GWs, with connectivity to nodes in a remote domain hidden. Thus introducing hierarchy into both the IP underlay and the EVPN overlay network for improved scale.

Layer 3 forwarding model

To provide seamless layer 3 VPN connectivity between the EVPN domains, the GW nodes participate in both the EVPN control-plane and the VXLAN forwarding plane. To achieve hierarchy for improved scale when interconnecting domains at layer 3, the next-hop of any type-5 (ip-prefix) route re-advertised by the GW across domains, is changed to the GW itself. By participating in the EVPN control-plane the GW can control via route-map policy the IP prefixes advertised between domains, changing the next-hop of the type-5 route also allows the IP underlay connectivity within a domain to be hidden from any remote domain.

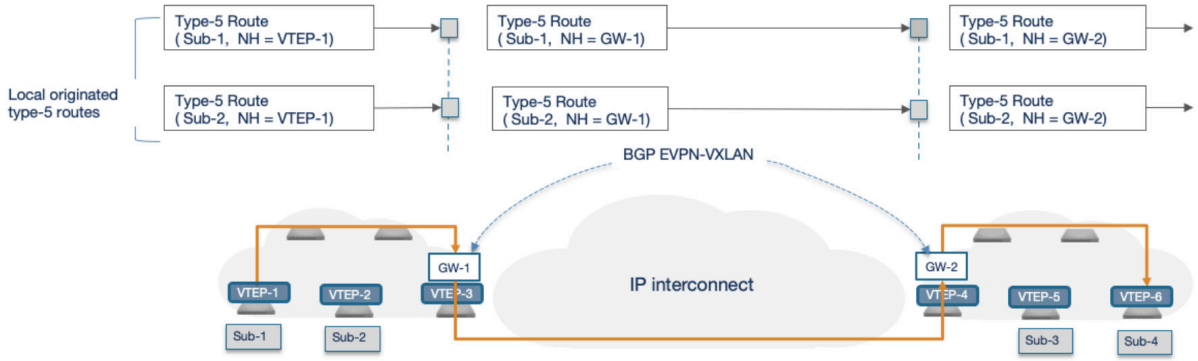

Figure 3: EVPN-VXLAN GW, EVPN layer 3 control plane with IP interconnect

The resultant end-to-end control-plane and forwarding plane for routing traffic between domains when an EVPN GW is deployed is illustrated in the figure below. In the figure a tenant has hosts in each EVPN domain, with the hosts residing in different IP subnets (host-1/sub-1 and host-2/sub-2).

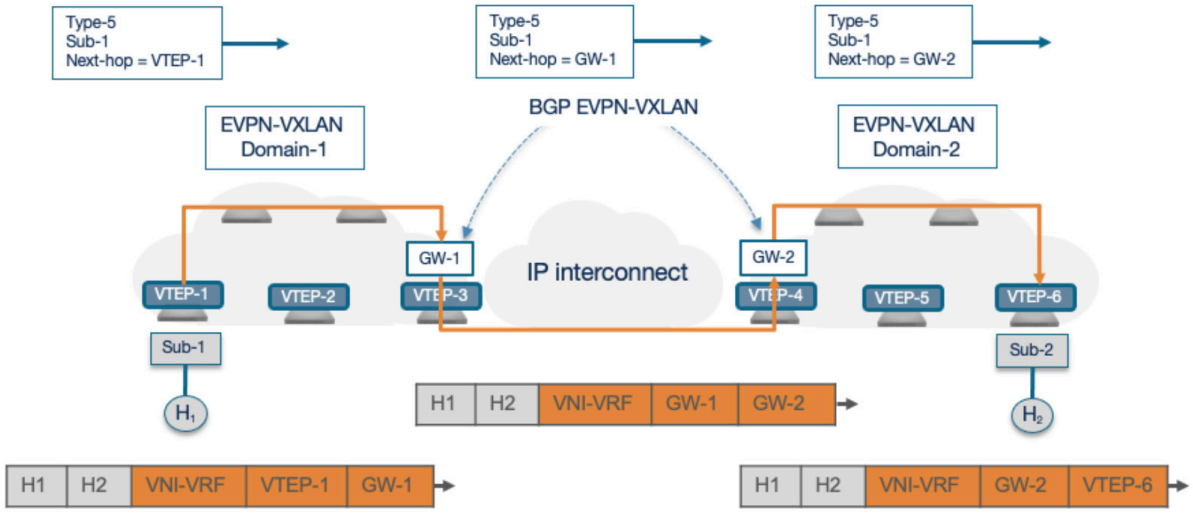

Figure 4: EVPN-VXLAN GW layer 3 forwarding plane with IP interconnect

In the topology a GW node is deployed at the edge of each EVPN domain, with EVPN peering sessions, with the local leaf nodes and the corresponding remote GW node. The layer 3 VPN connectivity for the tenant between the two domains, is achieved as follows:

1. VTEP-1 advertises a type-5 (ip-prefix) route for the subnet of host-1 (sub-1), the route is advertised with VTEP-1 as the next-hop and the route-target (RT) for the corresponding tenant's VRF.
2. As a member of the tenant VRF, the local GW node of the EVPN domain (GW-1) imports the route, and adds the entry to the VRF routing table with a next-hop of VTEP-1.
3. The GW also re-advertises the route to its remote GW peer (GW-2), with the next-hop of the type-5 route now changed to GW-1.
4. The remote GW-2, again as a member of the tenant's VRF, imports the type-5 route and adds an entry in the VRF for the subnet (sub-1) with a next-hop of GW-1.
5. GW-2 then re-advertises the type-5 route into it's local EVPN domain with a next-hop of GW-2,
6. VTEP-6 as a member of the VRF, imports the route, based on the RT and adds a route entry to the tenant's VRF for the subnet (sub-1) with a next-hop of GW-2. For the type-5 route (sub-2) advertised by VTEP-6, the process is repeated in reverse, as the route is received and re-advertised by first GW-2 and then GW-1 in the remote EVPN domain

From a forwarding plane perspective for the traffic flow between host-1 and host-2, VTEP-1 acting as the default GW for host-1, on receiving the frame destined to host-2, performs a route lookup for the destination subnet (sub-2), learning the destination subnet with a next-hop of GW-1. VTEP-1 VXLAN encapsulates and forwards the packet to GW-1. Receiving the packet, GW-1 removes the VXLAN header and performs a route lookup in the tenant's VRF based on the VNI of the VXLAN frame. GW-1 has a route for the destination prefix in the tenant's VRF with a next-hop of GW-2. The frame is VXLAN encapsulated by GW-1 and forwarded to GW-2 again via the tenant's VRF VNI. GW-2 receiving the encapsulated frame from GW-1 removes the VXLAN header and performs a route lookup in the associated tenant's VRF learning the destination subnet with the next-hop of VTEP-6. GW-2 adds a VXLAN header and forwards the frame to VTEP-6, which removes the outer VXLAN header and routes the packet to the local attached host (host-2).

Layer 2 forwarding model

The GW node also provides support for stretching layer 2 VPNs between EVPN domains. Like the layer 3 model the GW again participates in both the EVPN control-plane and the VXLAN forwarding plane. However, to scale the flood-list of the L2 domain and control unnecessary L2 state churn between domains, the scope of the L2 EVPN routes are controlled by the GW node. As illustrated below, type-1, type-3 and type-4 EVPN routes only have domain level scope and are not re-advertised by the GW nodes across domains.

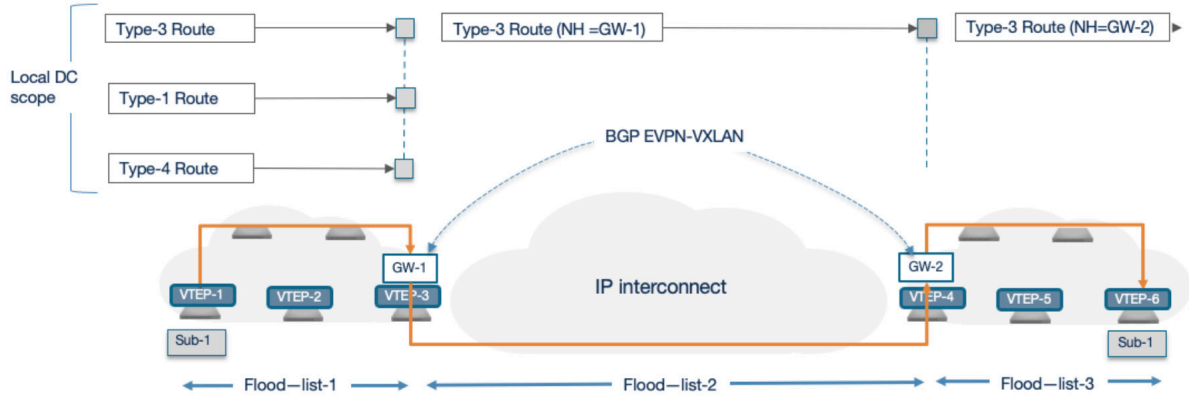

Figure 5: EVPN-VXLAN GW, EVPN layer 2 control plane with IP interconnect

The GW itself originates its own type-3 route which is advertised to the remote GW and leaf nodes in the local domain. Consequently the GW node will construct two flood-lists for any stretched layer 2 bridge-domain, a local flood-list containing the local leaf nodes in the bridge-domain and a second remote flood-list containing any GWs which are part of the bridge-domain. Thus BUM traffic originated from a leaf node would only be flooded to the local leaf and GW nodes that are part of the same bridge-domain, its then the responsibility of the GW to forward the BUM traffic via its remote flood-list to any GW node advertising membership (via type-3 routes) of the bridge-domain. The remote GWs would then flood the BUM traffic into their local domain, using their own local flood-list. With this approach the flood-list on any leaf node for a stretched layer 2 domain, will be limited in scale to local leaf nodes and the local GW, there will be no remote leafs in the flood-list. The approach thus ensures the underlay hierarchy is retained, as leaf nodes don't need to learn the next-hop of leaf nodes in the remote domains, in order to flood BUM traffic across domains. The type-1 and type-4 routes, which are responsible for advertising multi-homing ESI topologies, also only have EVPN domain level scope, this ensures any state churn on a local ESI doesn't result in unnecessary state churn in the remote EVPN domains, which don't need to know about the link failure of an ESI, if the corresponding hosts are still residing in the same EVPN domain with the same GW next-hop.

The resultant end-to-end control-plane and forwarding plane for bridging traffic between domains when the EVPN GW is deployed is illustrated in the figure below.

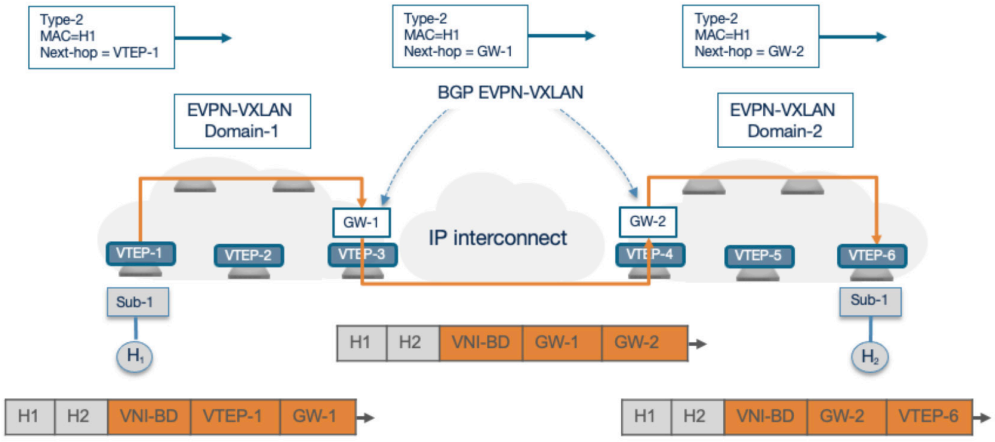

Figure 5: EVPN-VXLAN GW, EVPN layer 2 control plane with IP interconnect

In the topology a tenant has host-1 in EVPN domain-1 and host-2 in EVPN domain-2, with both hosts residing in the same subnet, connectivity between the two hosts is achieved as follows

1. VTEP-1 locally learns the MAC address of host-1, and advertises a type-2 MAC-IP route for the host along with the associated route-target for the bridge domain.
2. As a member of the bridge-domain, the local GW node of the EVPN domain (GW-1) imports the type-2 route into its local MAC and ARP table.
3. The GW re-advertises the type-2 route to the remote GW (GW-2), with the next-hop of the route changed to GW-1.
4. The remote GW (GW-2) as a member of the bridge domain imports the type-2 route (based on the route-target) and adds it to its local MAC and ARP table.
5. GW-2 then re-advertises the type-2 route into its local EVPN domain with a next-hop of GW-2.
6. VTEP-6 as a member of the bridge-domain, imports the route (based on the route-target) and adds a MAC entry for the host with a next-hop of GW-2.
7. For the type-2 route for host-2 advertised by VTEP-6, the process is repeated in reverse, as the route is received and re-advertised by first GW-2 and then GW-1 in the remote EVPN domain

From a forwarding plane perspective for the layer 2 traffic flow between host-1 and host-2, VTEP-1, on receiving the frame destined to host-2, performs a mac table lookup for the destination MAC, learning the MAC address via a VXLAN tunnel with a next-hop of GW-1. VTEP-1, VXLAN encapsulates and bridges the packet to GW-1. Receiving the packet, GW-1 removes the VXLAN header and performs a MAC table lookup, based on the VNI of the VXLAN frame for the inner destination MAC. Learning the MAC via a VXLAN tunnel with a next-hop of GW-2. The frame is VXLAN encapsulated by GW-1 and bridged to GW-2, on receiving the encapsulated frame, GW-2 removes the VXLAN header, performs a MAC lookup, learning the destination MAC via a VXLAN tunnel with a next-hop of VTEP-6, GW-2 adds a VXLAN header and forwards the frame to VTEP-6, which removes the outer VXLAN header and bridges the packet to the locally attached host (host-2).

EVPN GW with EVPN-MPLS interconnect

In certain deployment scenarios there will be a requirement to interconnect EVPN domains across an MPLS backbone rather than an IP backbone, where MPLS is being used in the WAN for historical reasons or to provide additional traffic engineering (TE) or fast reroute (FRR) capabilities. Arista's EVPN GW node, can also be used to address this requirement, while maintaining both layer 2 and 3 services between the EVPN domains. In this model the EVPN GW node, will operate an EVPN-VXLAN control plane and forwarding plane for connectivity to the local domain and an EVPN-MPLS control and forwarding plane for interconnecting the EVPN domains across the MPLS WAN.

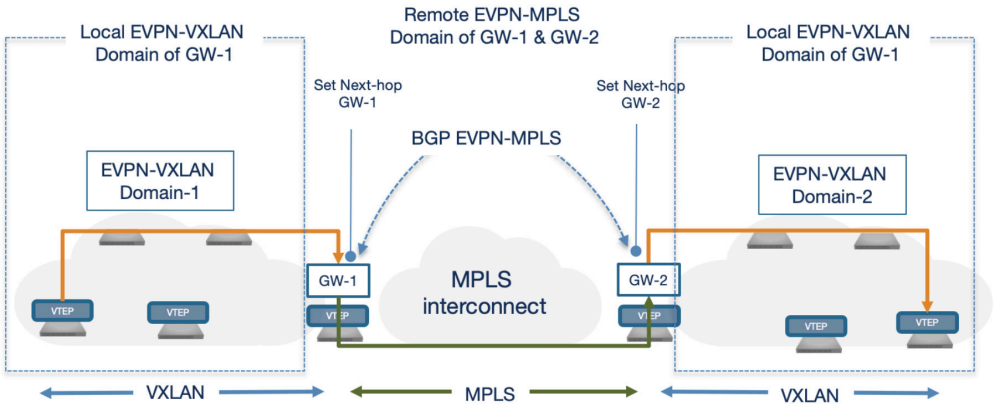

Figure 6: EVPN-VXLAN GW with EVPN-MPLS interconnect

The EVPN GW in this scenario is deployed in a similar manner to the VXLAN-to-VXLAN model, where GW nodes are deployed within each EVPN domain, with BGP EVPN-VXLAN peerings within the local domains, the difference now being that the peerings with the remote GW nodes are BGP EVPN-MPLS rather than BGP EVPN-VXLAN. The advertisement of EVPN routes between domains follows an identical model, where type 1, type-4 and type-3 routes only have domain level scope, and the GW re-originates its own type-3 route to create separate flood-lists (local and remote) for any bridge-domain that is stretched between the domains.

Type-2 (MAC/MAC-IP) and type-5 (ip-prefix) routes advertised by a node in the local EVPN-VXLAN domain are re-advertised by the GW to the peer GW nodes, as EVPN-MPLS labeled routes with the next-hop of the route set to the GW's loopback IP. With the GW advertising its loopback IP and associated label via LDP, RSVP-TE or SR, allowing the remote GWs to resolve the next-hop of the EVPN-MPLS route over an MPLS LSP. Any EVPN-MPLS route received from a remote GW is imported and re-advertised into the local domain as an EVPN-VXLAN route with the appropriate VNI label programmed and next-hop changed to the local GWs IP address.

Layer 3 Forwarding model

The figure below provides an overview of the EVPN control-plane and data-plane, for providing layer 3 services between the domains, with the GWs interconnected via EVPN-MPLS. In the figure a tenant has a host residing in each EVPN-VXLAN domain, with the hosts in separate subnets (host-1/sub-1 and host-2/sub-2).

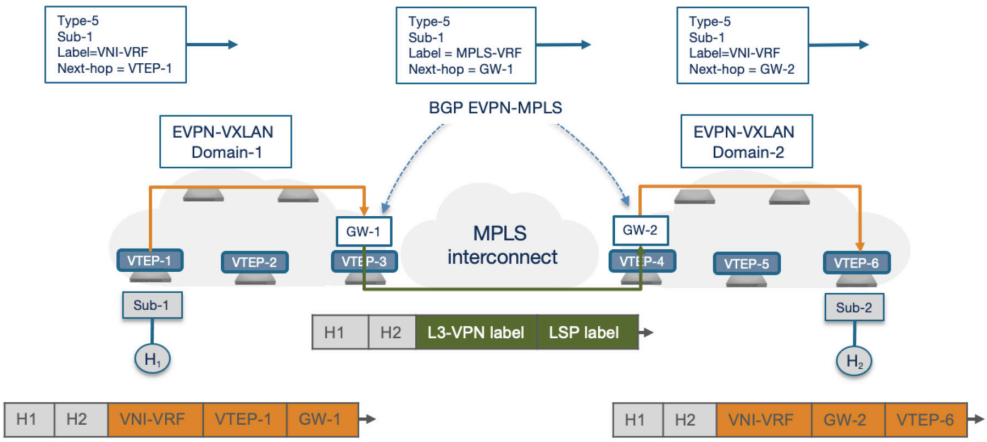

Figure 7: EVPN-VXLAN layer 3 GW with EVPN-MPLS interconnect

A GW node is deployed within each of the EVPN domains, with EVPN-VXLAN peering sessions within the local domain and EVPN-MPLS peerings with the remote GW nodes. Layer 3 connectivity within a VRF between the two domains, is achieved as follows:

1. VTEP-1 advertises a type-5 (ip-prefix) route for the subnet of host-1 (sub-1), the route is advertised as an EVPN-VXLAN route with VTEP-1 as the next-hop and a route-target (RT) and VNI corresponding to the tenant's VRF.

2. As a member of the VRF, the local GW node (GW-1) imports the route, based on the route-target, and adds the entry to the tenant's VRF routing table with a next-hop of VTEP-1 over a VXLAN tunnel.

3. The GW re-advertises the route to the remote GW (GW-2), as a type-5 EVPN-MPLS route, with the next-hop changed to GW-1 and a route-target and VPN label corresponding to the VRF.

4. The remote GW (GW-2), again as a member of the tenant's VRF, imports the type-5 route and adds an entry in the VRF for the subnet (sub-1) with a next-hop of GW-1 which is resolved over an MPLS LSP.

5. GW-2 then advertises a type-5 EVPN-VXLAN route for the prefix into it's local domain with a next-hop of GW-2, and a route-target and VNI corresponding to the VRF

6. VTEP-6 as a member of the VRF, imports the route, based on the RT and adds a route entry to the tenant's VRF for the subnet (sub-1) with a next-hop of GW-2 over a VXLAN tunnel.

7. For the type-5 route (sub-2) advertised by VTEP-6, the process is repeated in reverse, as the route is received and re-advertised by first GW-2 and then GW-1 into the remote EVPN-VXLAN domain.

From a forwarding plane perspective for the traffic flow between host-1 and host-2, VTEP-1 acting as the default gateway for host-1, on receiving the frame destined to host-2, performs a route lookup for the destination subnet (sub-2), learning the destination subnet with a next-hop of GW-1. VTEP-1 VXLAN encapsulates and forwards the packet to GW-1, receiving the packet GW-1 removes the VXLAN header and performs a route lookup in the tenant's VRF based on the VNI of the received VXLAN frame. GW-1 has a route for the destination prefix in the tenant's VRF with a next-hop of GW-2 which is resolved over an MPLS LSP. The GW adds the relevant VPN and LSP labels and forwards the MPLS packet to GW-2. On receiving the packet, GW-2 removes the MPLS header and performs a route lookup in the tenant's VRF based on the MPLS VPN label, learning the destination subnet with the next-hop of VTEP-6. GW-2 adds a VXLAN header with the VNI of the relavant VRF and forwards the frame to VTEP-6, which removes the outer VXLAN header and routes the packet to the local attached host (host-2).

Layer 2 forwarding model

The resultant end-to-end control-plane and data-plane for bridging traffic between domains is illustrated in the figure below. In the topology a tenant has host-1 in EVPN-VXLAN domain-1 and host-2 in EVPN-VXLAN domain-2, with both hosts residing in the same subnet, the GW nodes of each domain are again interconnected via EVPN-MPLS.

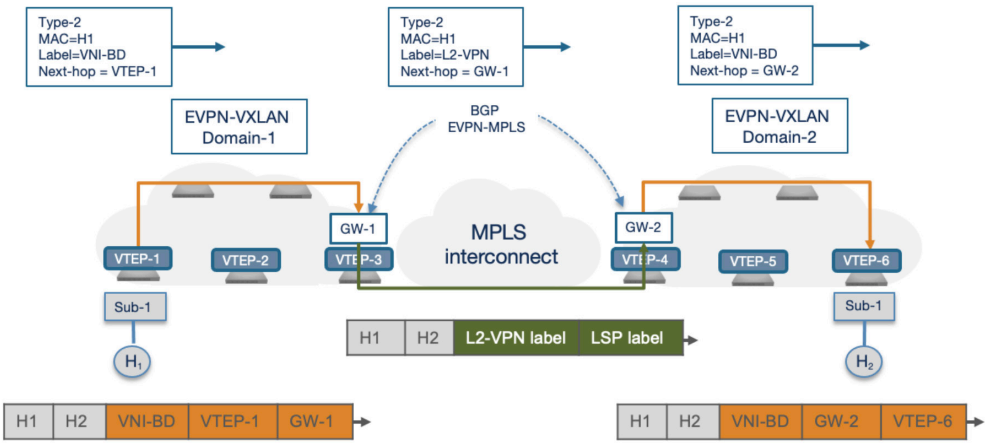

Figure 8: EVPN-VXLAN GW layer 2 forwarding plane with IP interconnect

1. VTEP-1 locally learns the MAC address of host-1, and advertises a type-2 EVPN-VXLAN route for the host along with the associated route-target and VNI for the bridge-domain.
2. As a member of the bridge-domain, the local GW node (GW-1) imports the type-2 route, based on the route-target, adding the MAC to its local bridging table with a next-hop of VTEP-1 which is learnt over a VXLAN tunnel.
3. With the bridge-domain stretched across the two EVPN domains, GW-1 will re-advertises the type-2 route to the remote GW (GW-2) as an EVPN-MPLS route. The next-hop of the type-2 route is changed to GW-1 with a route-target and VPN label corresponding to the bridge-domain.
4. The remote GW (GW-2) as a member of the bridge-domain imports the type-2 route, and adds it to its local MAC table, with the next-hop of the route (GW-1) resolved over an MPLS LSP.
5. GW-2 then re-advertises the type-2 route as an EVPN-VXLAN route into its local EVPN domain with a next-hop of GW-2 and the route-target and VNI associated with the bridge-domain.
6. VTEP-6 as a member of the bridge-domain, imports the EVPN-VXLAN route and adds a MAC entry for the host with a next-hop of GW-2 which is learnt over a VXLAN tunnel.
7. For the type-2 route for host-2 advertised by VTEP-6, the process is repeated in reverse, as the route is received and re-advertised by first GW-2 as EVPN-MPLS route and then re-advertised by GW-1 as EVPN-VXLAN type-2 route into its local EVPN domain

From a forwarding plane perspective for the 2 traffic flow between host-1 and host-2, VTEP-1, on receiving the frame destined to host-2, performs a mac table lookup for the destination MAC learning the MAC address via a VXLAN tunnel with a next-hop of GW-1. VTEP-1 VXLAN encapsulates and bridges the packet to GW-1. Receiving the packet GW-1 removes the VXLAN header and performs a MAC table lookup, based on the VNI of the VXLAN frame, for the inner destination MAC. Learning the MAC with a next-hop of GW-2, which has been learnt across an MPLS LSP, the packet is MPLS encapsulated by GW-1 and forwarded to GW-2. On receiving the MPLS encapsulated frame, GW-2 removes the MPLS header and performs a MAC lookup, based on the L2-VPN label of the MPLS frame, learning the destination MAC via a VXLAN tunnel with a next-hop of VTEP-6. GW-2 adds a VXLAN header and forwards the frame to VTEP-6, which removes the outer VXLAN header and bridges the packet to the locally attached host (host-2).

EVPN GW benefits

Regardless of the deployment scenario (VXLAN-to-VXLAN or VXLAN-to-MPLS), a multi-domain EVPN design offers a number of major operational benefits at scale over a single flat domain topology by introducing hierarchy.

Reduced network underlay state: With the GW node changing the next-hop of any EVPN route advertised between domains, the nodes in each domain only learn the next-hop of neighboring nodes in the same domain and their local GW node .There is no need to provide end-to-end IP connectivity in the underlay to the nodes in the remote domains. This is achieved by the GW participating in the VXLAN/MPLS forwarding plane, performing decapsulation and encapsulation functionality as packets are forwarded between domains. Consequently, simplifying the underlay routing design between domains and reducing the size of the underlay routing table on each node regardless of the number of EVPN domains interconnected.

Hierarchy for the EVPN overlay network: As the GW node participates in the EVPN control-plane of both domains, it can control what routes (type-2 and type-5) are advertised between domains via BGP best path calculation and route policies, providing the ability to optimize, aggregate and summarize routes before they are advertising between domains. This dramatically reduces the number of MAC, MAC-IP and IP prefixes advertised between domains and therefore programmed in the RIB and FIB tables of the nodes in each domain. For layer 3 interconnect, a node would only need to hold specific routes for subnets within its local domain, with summarized routes for subnets residing in any remote domain. For layer 2 interconnect, the GW and leaf nodes would only need to hold MAC and MAC-IP routes for the bridging-domains that are being stretched.

Improved layer 2 flood-list scaling: By introducing separate flood-domains and split-horizons for the forwarding of BUM traffic between domains, the GW node dramatically reduces the size of the layer 2 flood-list on the nodes within each domain, regardless of the number of EVPN domains being interconnected. Type-3 (IMET) routes only have local domain significance, meaning the flood-list of an individual node is constrained to nodes in their own domain and their local GW node, with the GW responsible for flooding BUM traffic to remote GWs, which would be responsible for forwarding the BUM traffic into their own local domain.

Reduced EVPN State churn: By restricting the scope of type-1 and 4 routes, which are responsible for advertising multi-homing ESI topologies, ensures any state churn on a local ESI does not result in any unnecessary state churn in the remote EVPN domains. Remote EVPN domains do not need to know about a link failure on an ESI, if the corresponding host/MAC is still residing in the same EVPN domain with the same GW next-hop. This reduction in state churn across domains, is further optimized by the fact that any MAC move within a domain doesn't result in forwarding state change in remote domains, as the next-hop for the host will still be the same GW node, a forwarding state change across domains will only occur when hosts/MACs move between domains and the resultant next-hop GW IP address changes.

EVPN GW deployment models

The benefits Arista's EVPN GW offers, has seen GW nodes deployed within the data center in a multi-domain data center design, for scale and fault-containment reasons and across geographically dispersed data centers in a multi-domain DCI design, to provide seamless inter data center VPN connectivity, with the flexibility to provide both IP and MPLS DCI design solutions.

Multi-Domain EVPN Data Center design

In the multi-domain data center design, to achieve improved scale within the DC while providing fault-containment, the EVPN topology is constructed from multiple leaf-spine fabrics, where a fabric could represent a row or floor within the data center. Each fabric is configured as a self-contained EVPN domain, with GW node(s) deployed to provide connectivity between domains. To provide high-speed interconnect between a small number of domains where there are no plans for future growth, the EVPN GWs can be interconnected in full mesh topology, with either an iBGP or eBGP design for EVPN GW peerings between domains

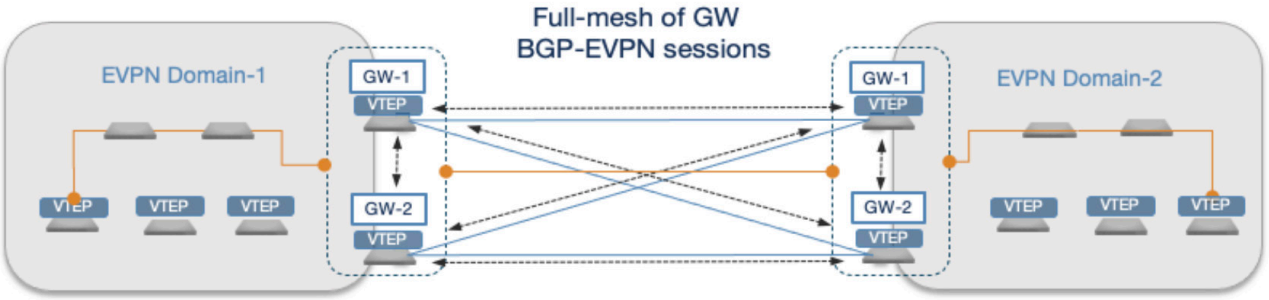

Figure 9: Small scale EVPN multi-domain design, full-mesh of EVPN peerings between GW nodes

If the number of BGP sessions between GW nodes is a concern or there is need to provide scope for future growth of the EVPN footprint, the more scalable approach would be to introduce a Super-spine layer for interconnecting the EVPN GWs at the edge of each domain. The BGP-EVPN peering between GWs being achieved via the Super-spine nodes which can act as Route-Servers (RS) in a eBGP design or Route-Reflectors (RR) in an iBGP design.

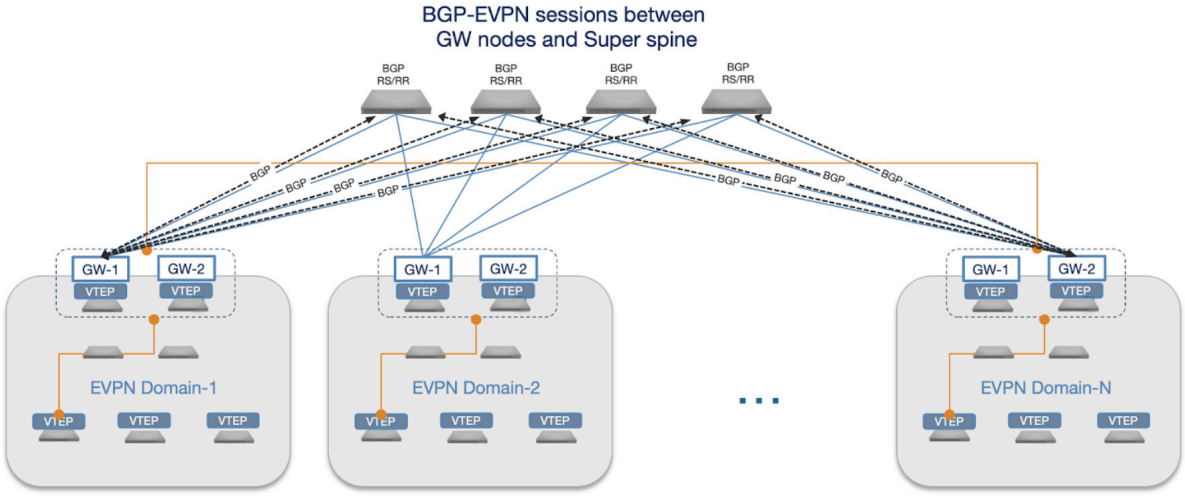

Figure 10: Large scale EVPN multi-domain design, super spine for interconnect & EVPN peering between GWs

Multi-Domain DCI solution

In a DCI design, there is a requirement to interconnect multiple EVPN sites/data centers across an MPLS or IP backbone to provide seamless layer 2 and 3 VPN services between the sites. To provide a scalable BGP EVPN design, for the GW nodes of each site to exchange routes, while minimizing the disruption to the existing backbone, a pair of RRs would typically be deployed within the backbone with the GW nodes iBGP peering with the RRs to exchange EVPN routes. In this type of topology "local-as" can be used on the GW peering session with RR, to hide private AS schemes within each site. This BGP topology for the backbone can be deployed regardless of whether the GW is configure for EVPN-VXLAN-to-MPLS which would be the case with an MPLS backbone or EVPN-VXLAN-to-VXLAN which would be the case with an IP only backbone.

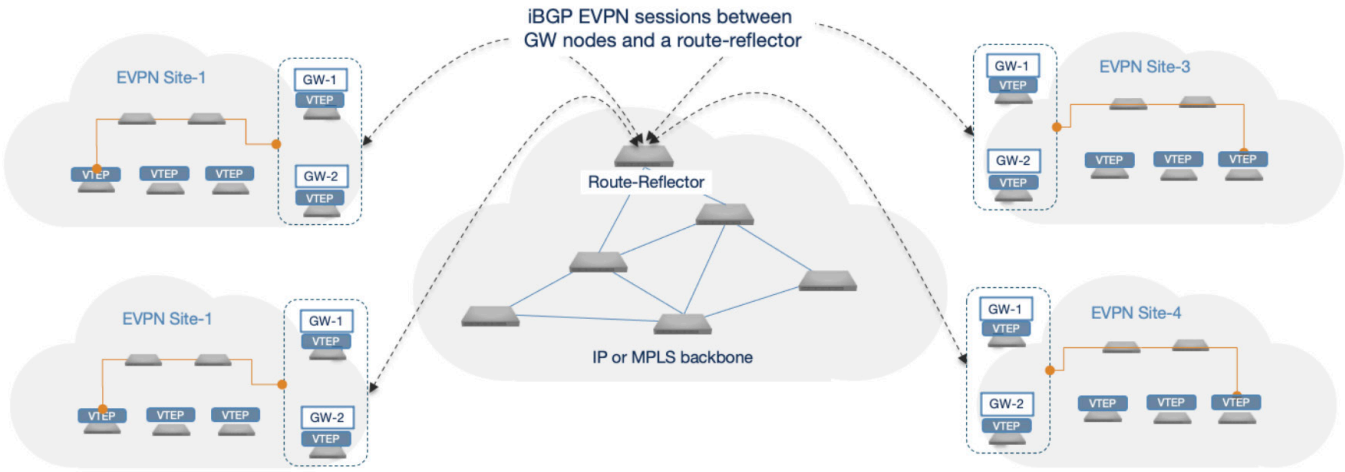

Figure 11: DCI multi-domain design, with an IP or MPLS backbone, with RR and iBGP EVPN peering between GWs

Conclusion

Modern leaf-spine data center topologies can now scale to support 100s of leaf nodes, in the context of a multi-tenant DC design the amount of EVPN state this flat single scaled-out fabric approach involves, can present a challenge to the finite hardware resources of the leaf nodes within the fabric. This scaling challenge is further complicated by the increased trend for stretching L2 and L3 VPNs between data centers to ensure service continuity in the event of a data center failure.

To address these challenges, Arista has introduced a standards based EVPN GW model, for building hierarchical EVPN designs, providing the ability to divide the end-to-end EVPN topology into multiple self-contained EVPN domains for improved scale and fault-containment, while retaining seamless layer 2 and L3 VPN interconnect between domains. This hierarchical EVPN approach, is not limited to just within the data center, with Arista's EVPN GW supporting both EVPN-VXLAN and EVPN-MPLS services, this multi-domain approach can be extended across sites, to form a DCI, allowing the extension of EVPN services between DCs, with the ability to support both VXLAN and MPLS encapsulations when transversing the WAN.

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