The EVPN Data Center

Multihoming models with EVPN

Introduction

In today's data center, EVPN with VXLAN encapsulation (RFC 8365) has become the adopted approach for building a standards based solution to deliver unicast and multicast VPN services across a shared IP leaf-spine infrastructure. For resiliency reasons, within the EVPN environment there is a requirement to provide multi-homing support for servers, switches and routers connecting to the EVPN domain. Where the EVPN multihoming solution will be required to provide support for:

- **Node and Link level resiliency:** The multihoming solution is required to provide layer 2 nodal and link level resiliency, whereby if the local link or VTEP node connecting the server or switch to the EVPN domain fails, traffic should seamlessly failover to the remaining active link(s) or node(s).
- **Active-Active model:** To make use of all available resources and bandwidth, the multi-homing solution needs to provide an active-active forwarding model across all links and nodes under steady state conditions.
- **Loop free layer 2:** While the multi-homing solution should provide a loop-free approach for dual-homing layer 2 services, there will often be a requirement to interact with traditional layer 2 switches, therefore the solution needs to interop with traditional spanning tree domains when required.
- **Unicast and Multicast Services:** The multi-homing model is required to provide the equivalent level of resiliency when connecting both unicast and multicast services to the EVPN domain.
- **Single-homed devices:** While providing support for dual-homed switches/servers to the EVPN domains, there is also a requirement for the solution to provide the capability to support a mix of dual-homed and single-home nodes, while ideally maintaining optimal forwarding for both.
- **Layer 3 services:** The solution should not be limited to layer 2 connectivity and should provide support for connecting layer 3 services, with the ability to provide IGP peering with the layer 3 node when required.

To achieve this level of resiliency and functionality, Arista's EVPN implementation provides two potential multi-homing solutions; Multi-chassis LAG (MLAG) and EVPN all-active (A-A) multi-homing. This whitepaper discusses the details of both solutions, how they address the requirements listed, and provides guidance on the appropriate model based on the design requirements.

EVPN with Multi-chassis LAG (MLAG)

Arista's MLAG approach allows two physical nodes to act as a single logical switch, where downstream servers and switches connect to the logical switch via a port-channel, with the physical links of the port-channel split across the two nodes of the MLAG domain for resiliency. The MLAG nodes are interconnected via a peer-link, which is used under steady-state conditions for state synchronization and keep-alives. The synchronization of state across the peer-link, allows the two nodes to appear as a single logical switch, thereby providing a loop-free topology for dual-homing servers and switches. With the downstream server or switch transparent to the MLAG technology, the port-channel can be a static configuration or a standard based LACP port-channel.

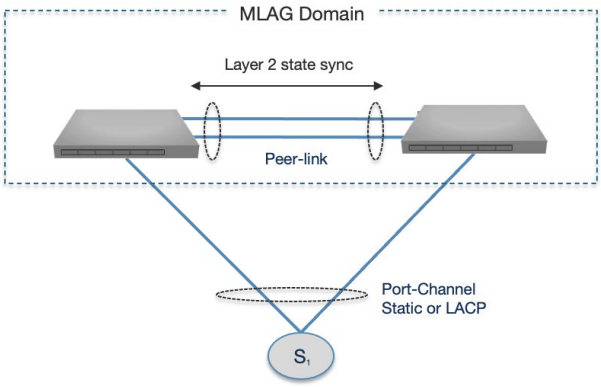

Figure 1: Multi-Chassis LAG (MLAG) topology

Under steady-state conditions layer 2 traffic follows the optimal path for any dual-homed server, this means when traffic egresses the server and is load-balanced onto one of the links of the port-channel, the receiving MLAG node will be responsible for forwarding the traffic via a local link if the destination host is directly connected or the local uplink if the host is learnt remotely. Similarly for layer 3 forwarding, the MLAG nodes provide a virtual gateway functionality (virtual MAC and IP) which is shared across both nodes, this acts as the default gateway for the directly attached host. With this configuration traffic received by either node can be routed at the first-hop without the need to traverse the MLAG peer-link.

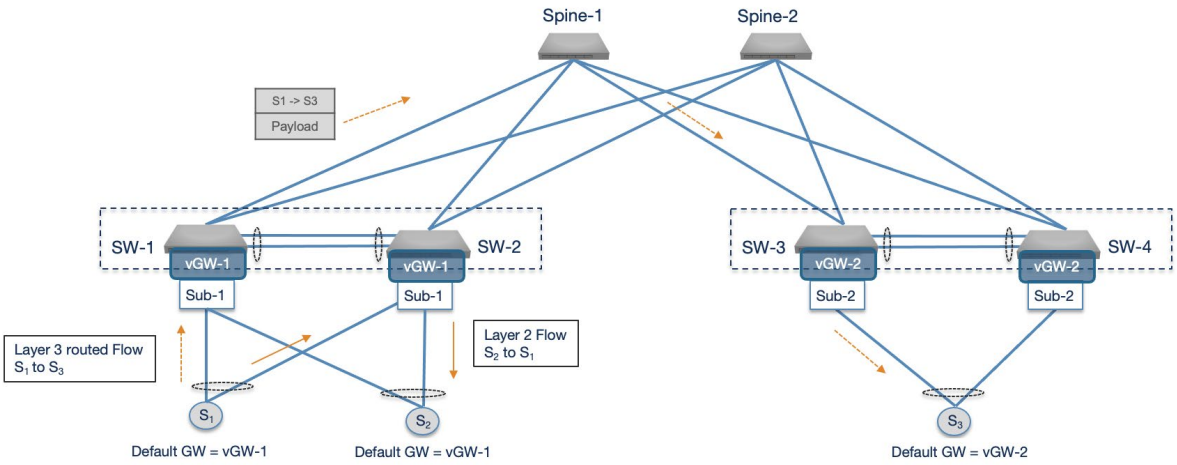

Figure 2: MLAG optimal layer 2 and 3 forwarding under steady-state conditions

Thus in the MLAG topology layer 2 and 3 forwarding under steady state conditions, when all servers are dual-homed, always follows the optimal path and does not traverse the MLAG peer-link. All links and nodes of the topology are therefore active and forwarding traffic regardless of what node of the MLAG domain receives traffic from a locally attached server.

EVPN control-plane with MLAG

Operating within an EVPN topology, an MLAG domain can again be used to dual-home servers and switches to provide active-active layer 2 and 3 forwarding. This is achieved by the two nodes within the MLAG domain operating as a single logical VTEP, sharing the same IP address for their VTEP interface. In the forwarding plane, both nodes are capable of VXLAN encapsulating locally received traffic destined to a remote host, with the shared VTEP IP address used as the source address of any VXLAN encapsulated packet. A VXLAN frame destined to the shared IP address can also be decapsulated and forwarded to a locally attached host by either node. This logical VTEP model therefore provides an active-active optimal forwarding model for both VXLAN encapsulation and decapsulation under steady state conditions.

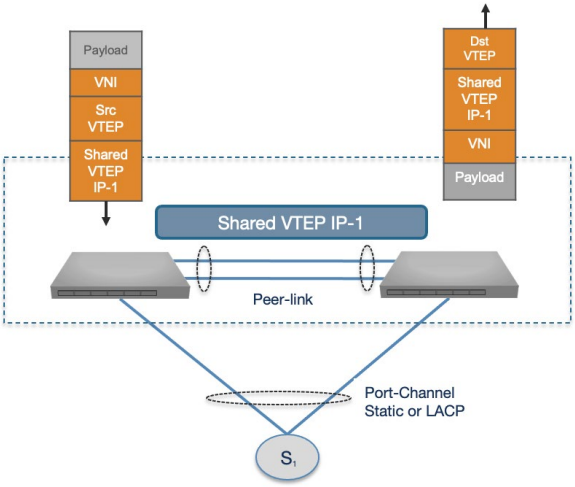

Figure 3: MLAG with EVPN, providing active-active forwarding with a shared logical VTEP IP

For the IPv4 underlay and EVPN overlay control-plane, the individual nodes of the MLAG domain have dedicated underlay and overlay peerings with each of the spine nodes within a leaf-spine topology. In the example, the underlay routes are advertised using BGP, however, any IGP routing protocol could also be deployed. In the topology, each MLAG node has a BGP IPv4 underlay peering to each spine node, and a second BGP EVPN peering to each spine node.

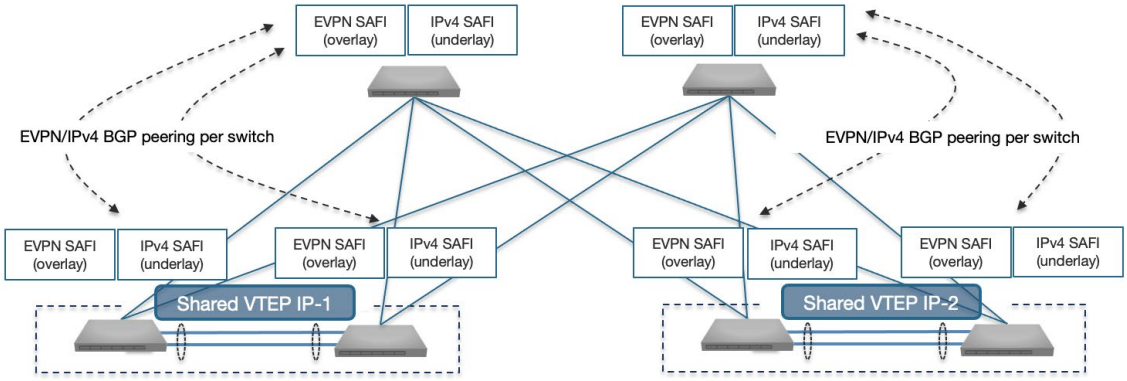

Figure 4: MLAG IPv4 and EVPN BGP peering session per physical switch

The routes advertised across the two BGP peering sessions are as follows:

- **BGP IPv4 (AFI 1 /SAFI 1):** This is the underlay BGP IPv4 peering session, and is used to advertise connectivity to the loopback IPs of the individual nodes and the shared VTEP IP address of the MLAG domain, which will be used for VXLAN encapsulation and the next-hop address of any EVPN advertisement. As stated, the model is not limited to BGP for advertising underlay routes, any IGP routing protocol can be deployed, BGP is a common design within the data center leaf-spine topology for scaling reasons.

- **BGP EVPN (AFI 25 /SAFI 70):** This is the overlay BGP EVPN peering session, and is used to advertise MAC, MAC-IPs and IP-prefixes of the locally connected overlay network. The EVPN routes originated by a node in the MLAG domain, are advertised with the MLAG shared VTEP IP address as the next-hop, which is a loopback IP address advertised in the BGP underlay.

To achieve the active-active forwarding behavior, a node locally learning a MAC or MAC-IP binding shares the state with the peer node of the MLAG domain via the peer-link and in the EVPN control plane advertises a EVPN type-2 route with a next-hop equal to the shared IP VTEP address. With connectivity to the shared VTEP IP address advertised by both nodes in the IP underlay, remote VTEPs will learn two equal cost underlay paths to the EVPN route one via each node of the MLAG domain. Thus traffic destined to the EVPN route, will be load-balancing via ECMP in the underlay, to both nodes of the MLAG domain.

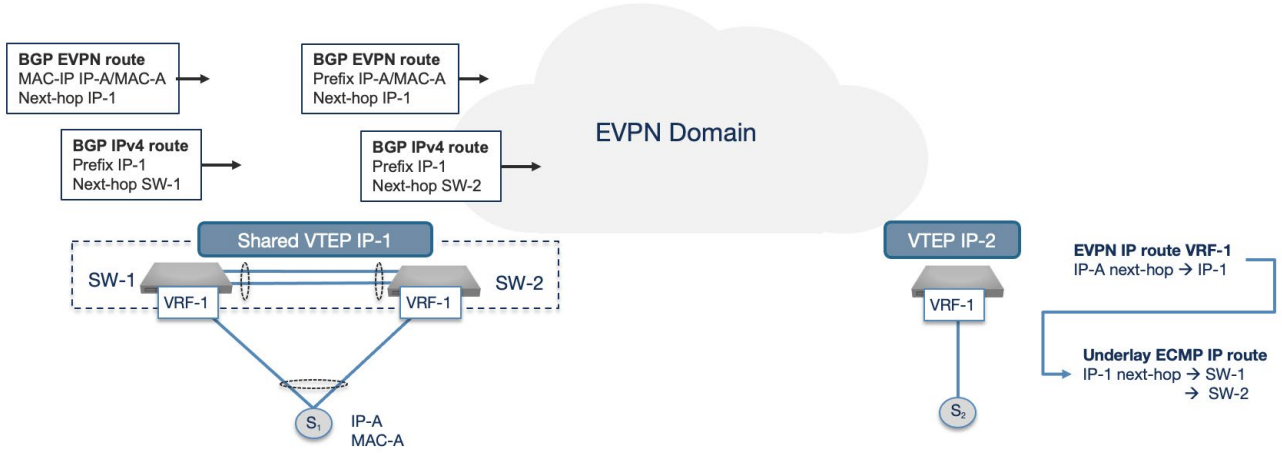

Figure 5: EVPN with MLAG, active-active layer 2 and 3 forwarding model with shared VTEP IP

Broadcast Unknown unicast and Multicast (BUM) Traffic

In the MLAG topology, BUM traffic received from a locally attached host, is forwarded across the peer-link, and flooded to remote VTEPs based on the associated flood-list for the VNI, which would be populated from advertised type-3 (IMET) routes. An MLAG peer node receiving BUM traffic across the peer link, is responsible for forwarding the traffic to any single attached host, and performing split-horizons to prevent the BUM traffic being forward to any local port-channel which has an active link on both nodes.

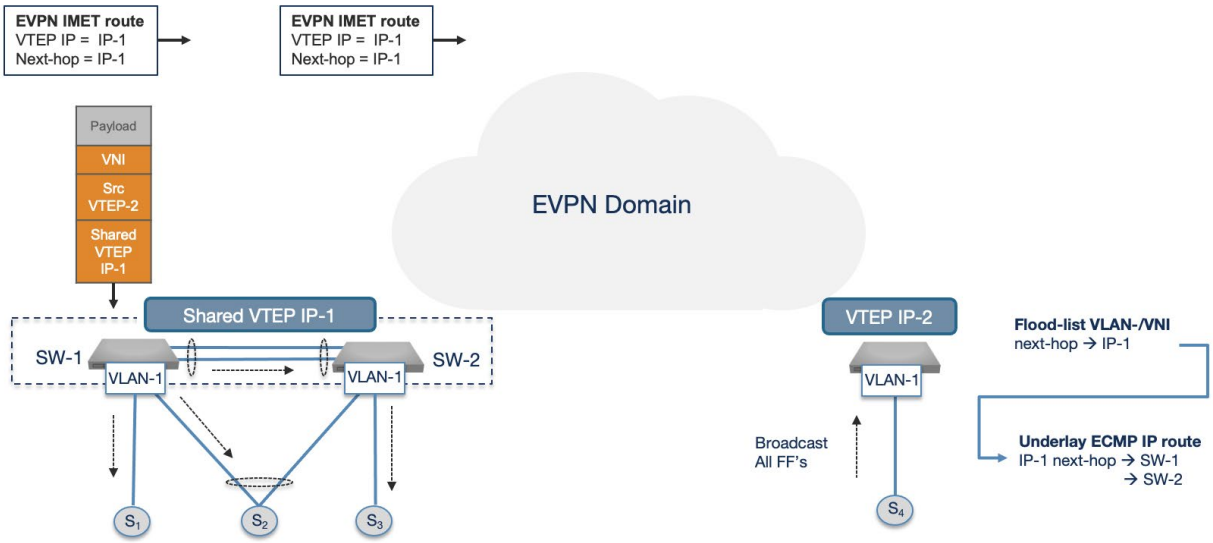

Figure 6: EVPN with MLAG, IMET route advertisement and BUM traffic forwarding

For a remote VTEP receiving BUM traffic from a locally attached host, it will follow normal EVPN procedures, forwarding the traffic to all VTEPs in the associated flood-list for the VNI, the flood-list being populated based on advertised type-3 IMET routes. For an MLAG topology, the type-3 (IMET) routes are advertised by both nodes with the same next-hop (shared VTEP IP). The BUM traffic is therefore forwarded to the shared VTEP IP of the MLAG domain, which due to ECMP in the underlay could be received by either node, the receiving node would decapsulate the VXLAN frame and forward the BUM traffic to the local links of the VNI/VLAN and across the peer-link for any single-homed devices on the peer node.

State synchronization

With the MLAG approach the topology is restricted to two nodes, with the nodes required to be interconnected via a peer link in order to synchronize state. While this can be restrictive when a higher level of resiliency is required, or available cabling/interfaces are limited on the nodes within rack, the synchronization of state and healthcheck across the peer-link does mean the MLAG model doesn't introduce any new EVPN routes and therefore EVPN state churn across the EVPN domain when comparing to the A-A model.

Traffic Load-balancing

In the MLAG approach, EVPN routes are advertised with the shared VTEP IP address of the MLAG domain as the next-hop. Connectivity to the shared VTEP IP address is advertised by both nodes in the IP underlay, resulting in the remote VTEPs having a 2-way ECMP path to the EVPN route one via each node of the MLAG domain. This means traffic destined to a host advertised in an EVPN route, will be load-balanced in the network underlay rather than load-balanced in the network overlay. Performing the load-balancing in the network underlay can improve network re-convergence in the event of a link or node failure within the MLAG domain, while reducing the amount of EVPN state churn across the VTEPs of the EVPN domain, when comparing to an A-A model (see the "Failover" section for more detail).

Spanning Tree

Acting as a single logical switch to the downstream dual-homed nodes, the MLAG approach provides inherent spanning tree support if required. One of the nodes in the MLAG domain is elected the STP master and advertises spanning tree BPDU with the elected bridge-id of the MLAG domain, termed the MLAG system Identifier (MSI). Downstream nodes thus only see a single logical node in the spanning tree topology and don't block any ports. If the elected spanning-tree "master" of the MLAG domain fails, the backup node takes up ownership and starts advertising spanning tree BPDUs with the same elected MSI value, the failure is therefore transparent and seamless to the downstream nodes. Note in this model, spanning-tree is operating downstream to the end-hosts and switches, spanning-tree BPDUs are not forwarded across the VXLAN tunnels to the remote VTEPs in the EVPN domain.

Layer 2 nodes single-homed

The topology supports the ability to attach single-homed hosts and switches to the MLAG domain, however traffic destined to the single-homed host can't always be guaranteed to follow the optimal path. The MAC address of a single-homed host will be learnt by the directly attached node and shared with the peer node of the MLAG domain via the peer-link. In the EVPN control plane, the type-2 route for the host's MAC will be advertised with the shared VTEP IP address as the next-hop.

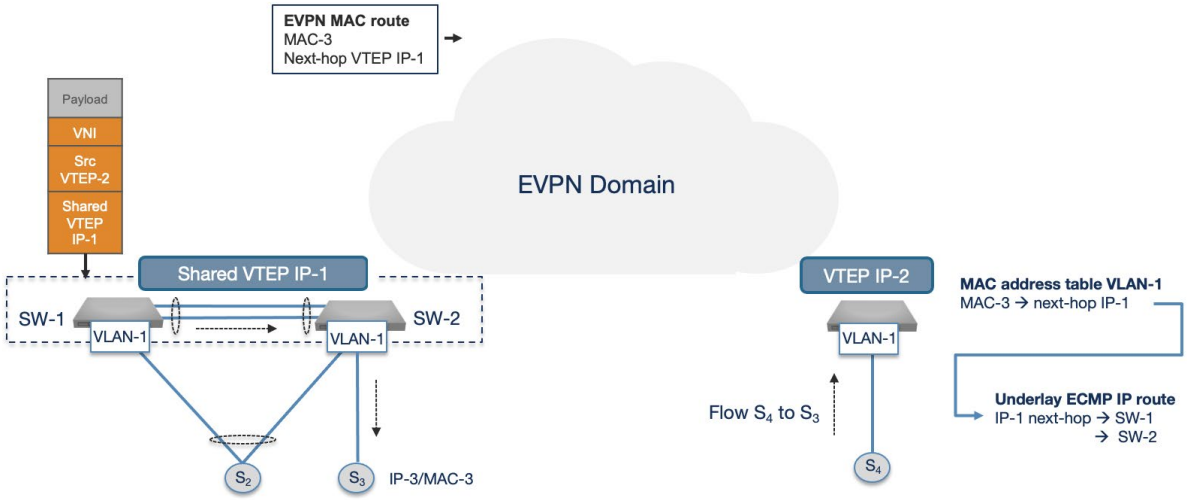

Figure 7: EVPN with MLAG, forwarding model for single-homed nodes

Advertising the type-2 route with a next-hop of the shared VTEP IP address, means VXLAN traffic destined to the MAC could be received by either node of the MLAG domain, if it's received by the node that is not directly connected to the host, the packet will be VXLAN decapsulated and forwarded to the host via the peer-link; if it's received by the node directly connected to the host it will follow the optimal path.

Anycast Gateway

To provide EVPN Integrated Routing and Bridging (IRB) for directly attached hosts, MLAG supports an anycast GW. The anycast GW, is a virtual IP and MAC address, that is configured on each of the VLANs shared between the nodes of the MLAG domain. With the virtual GW shared across the two nodes, and both nodes capable of responding to ARPs destined to the virtual IP and routing traffic destined to the virtual MAC, an active-active routing model is achieved. Thus regardless of how traffic is load-balanced on the port-channel from the host, either node of the MLAG domain will be capable of routing the traffic directly without the need to switch traffic across the peer-link.

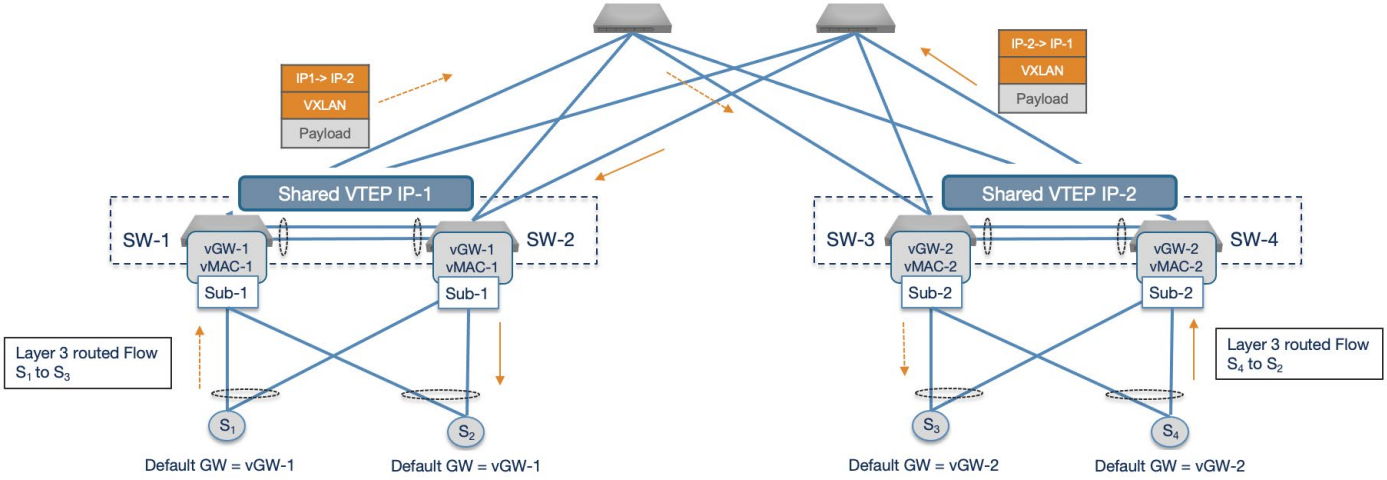

Figure 8: EVPN with MLAG and anycast GW for layer 3 forwarding

MLAG with Fastpath return

In the anycast GW model, traffic routed by an MLAG node to an end-host is forwarded with the source MAC of the node's unique system MAC, the virtual MAC is only used for ARP responses to the virtual GW IP. Certain network appliances and storage arrays

support a feature called "Fastpath" or "Symmetric return", where the GW MAC address is learnt in the forwarding path by inspecting the source MAC of the received packet rather than an ARP response from the GW. With routed packets to the appliance in the anycast GW model, using the system MAC of the node performing the routing action, this can have an adverse effect on the forwarding behavior of the MLAG topology. The traffic forwarded by the "fast-path" appliance to the GW, would use the system MAC of one of the nodes in the MLAG domain as the destination MAC for the packet rather than the virtual MAC. Due to load-balancing on the port-channel, either node of the MLAG domain may receive the traffic, if the destination MAC is not owned by the receiving node, it will be bridged over the peer link for routing rather than routed directly. To provide support for the "Fastpath" model, while maintaining optimal first-hop routing, MLAG with EVPN provides the capability, via user configuration, to allow a node to route traffic destined to the system MAC of the peer node in the MLAG domain, this functionality is called "mlag peer mac routing".

Layer 3 nodes dual-homed

In the MLAG model, for resiliency layer 3 nodes can be dual-homed to the MLAG domain, via dedicated layer 3 point-to-point links to each node of the MLAG domain. An IGP or BGP peering session is run across the point-to-point link to exchange routes, the prefixes learnt are then advertised as type-5 (ip-prefix) routes into the EVPN domain by both MLAG nodes.

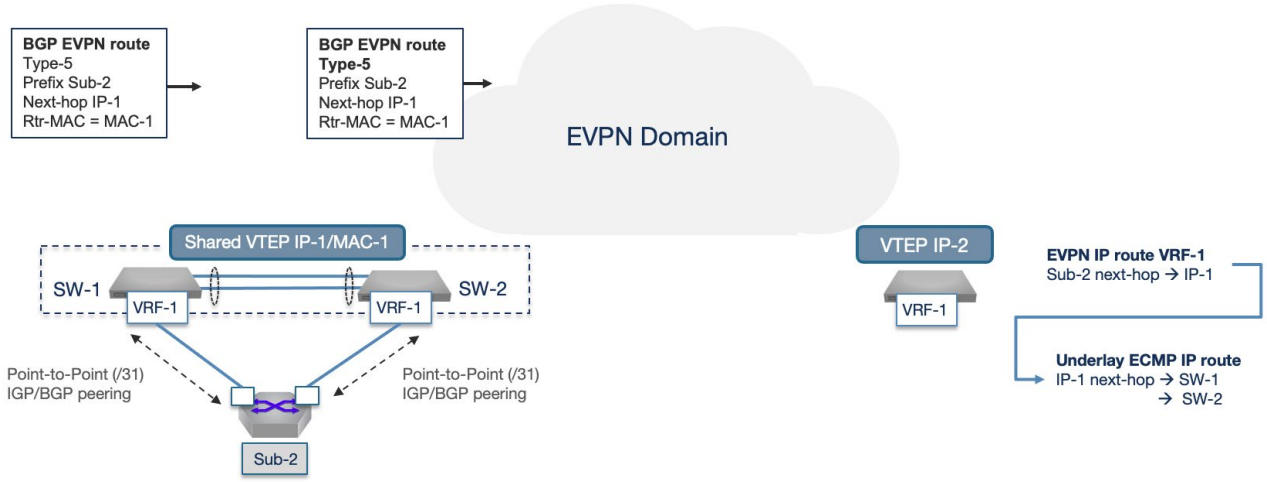

Figure 9: EVPN with MLAG, Layer 3 nodes dual-homed with the MLAG shared router-mac

The EVPN type-5 routes are advertised with both a next-hop and a router-mac which would be the inner destination MAC of the VXLAN frame when forwarded to the advertised prefix. By default, the MLAG nodes will advertise EVPN type-5 routes with the same next-hop which would be shared VTEP IP, but with their own router MAC. This default behavior can result in sub-optimal forwarding when attaching layer 3 nodes to the MLAG domain, as traffic destined to an advertised prefix will be forwarded to the shared VTEP IP, on removing the VXLAN header if the inner destination MAC is not own by the receiving node, the packet will be switched across the peer-link for routing. To provide optimal routing in this model and avoid traffic traversing the peer-link under steady-state conditions the "MLAG shared router MAC" functionality can be enabled. The functionality provides the ability to change the default behavior, allowing nodes in an MLAG domain to advertise the type-5 routes with the same next-hop (shared VTEP IP) and a configurable shared router-mac. Thus ensuring optimal layer 3 forwarding with type-5 prefixes, avoiding the need for traffic to traverse the peer link.

Layer 3 nodes single homed

In the previous Layer 3 model both nodes of the MLAG domain are connected via dedicated L3 point-to-point links to the same downstream router and therefore advertise the same prefixes into the EVPN domain. If there is a requirement to single-home a layer 3 node; connect the router to only one of the nodes in the MLAG domain, there is the potential for suboptimal forwarding or traffic blackholing. The type-5 routes for the prefixes learnt from the single homed router will be advertised with the shared Virtual VTEP IP

as a next-hop, thus resulting in sub-optimal forwarding and potentially traffic blackholing as only one node in the MLAG domain has a route to the advertised type-5 prefixes.

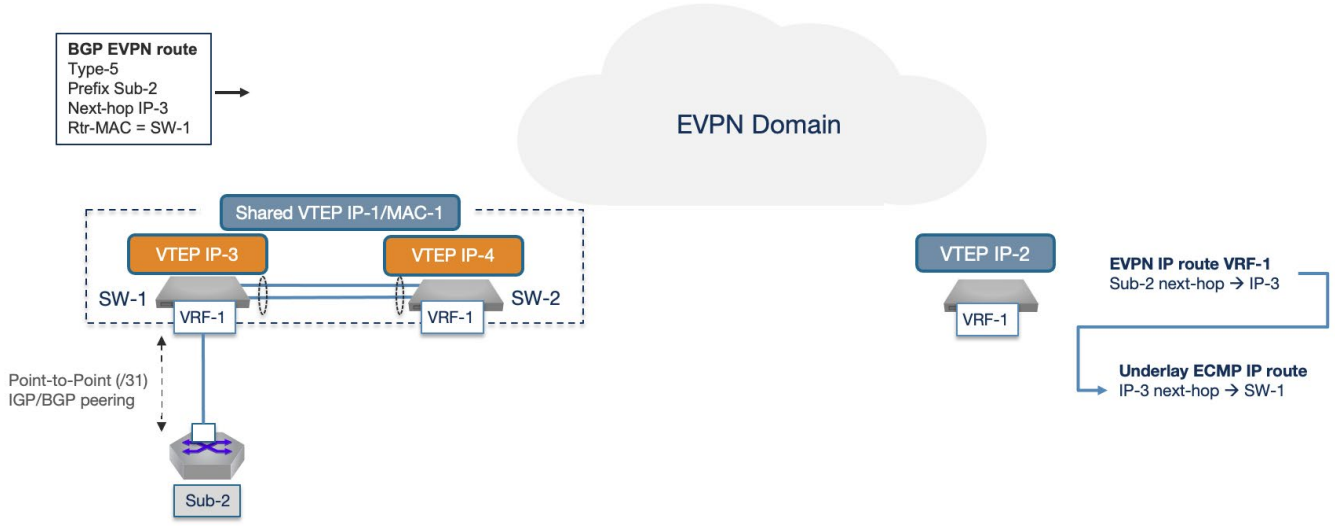

Figure 10: EVPN with MLAG, optimal forwarding for single-homed layer 3 nodes

To provide optimal layer 3 forwarding with this topology, while avoiding traffic blackholing, the MLAG topology provides the ability to configure multiple VTEP IP addresses; 1) Shared VTEP IP address for dual-homed nodes 2) Secondary physical VTEP IP address unique to the node. With this model the prefixes learnt from the single-homed router can be advertised with a next-hop that is unique to the attached node, ensuring traffic destined to the prefix is always forwarded to the correct node in the MLAG domain, for any prefixes or MACs that are dual-homed they follow the standard behavior and are advertised with a next-hop of the shared VTEP IP address. An alternative solution to this problem, would be to configure a peering-session within the VRF between the two nodes across the peer-link, allowing both nodes to learn the prefix and be able to route the traffic across the peer-link if required. This approach would avoid any traffic being black-holed, although it would result in sub-optimal forwarding in certain scenarios.

EVPN Multicast with MLAG

In an EVPN multicast deployment, both multicast sources and receivers can be dual-homed to the EVPN domain via MLAG. From a multicast source perspective, both nodes are able to VXLAN encapsulate a multicast flow received from a locally attached source, the node performing the encapsulation of a specific flow, would be based on the source's load-balancing of the flow across the port-channel. The receiving MLAG node is able to VXLAN route or bridge the multicast flow as required, thus providing an active-active forwarding model where the multicast flow from the source will always follow the optimal path without the need to traverse the MLAG peer-link.

In a PIM underlay solution, each node of the MLAG domain will advertise unique underlay (S,G) group for transporting the VXLAN encapsulated packet, with the underlay to overlay group mapping advertised by the node in a type-10 (S-PMSI) route. To receive the VXLAN encapsulated multicast stream, remote VTEP with interested receivers would join the advertised underlay group.

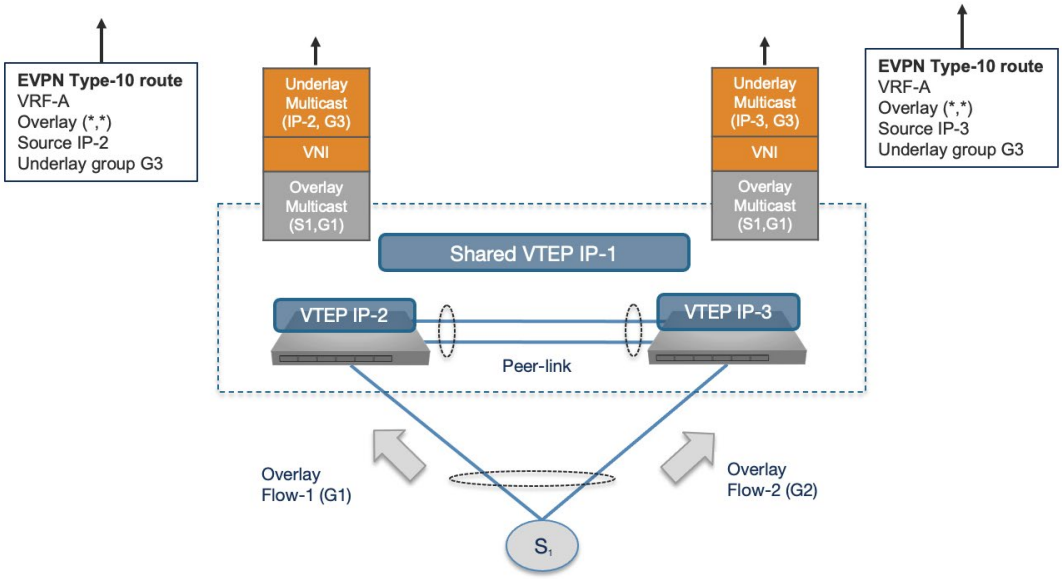

Figure 11: EVPN with MLAG, optimal forwarding for single-homed layer 3 nodes

For dual-attached multicast receivers, both nodes of the MLAG domain will advertise an associated type-6 SMET route for the interested receiver, where the receiver's IGMP join is forwarded across the peer link, to synchronize IGMP state between the nodes. To prevent double delivery of the multicast stream, and optimize bandwidth utilization across the fabric, in a PIM underlay solution only one of the MLAG peers will join the associated underlay group for the stream's VRF. Thus under steady state conditions, only one node of the MLAG domain will receive the multicast streams(s) for a specific VRF, if the DR for a specific subnet resides on the peer MLAG node, then the stream will traverse the peer link for routing into the subnet by the elected DR node.

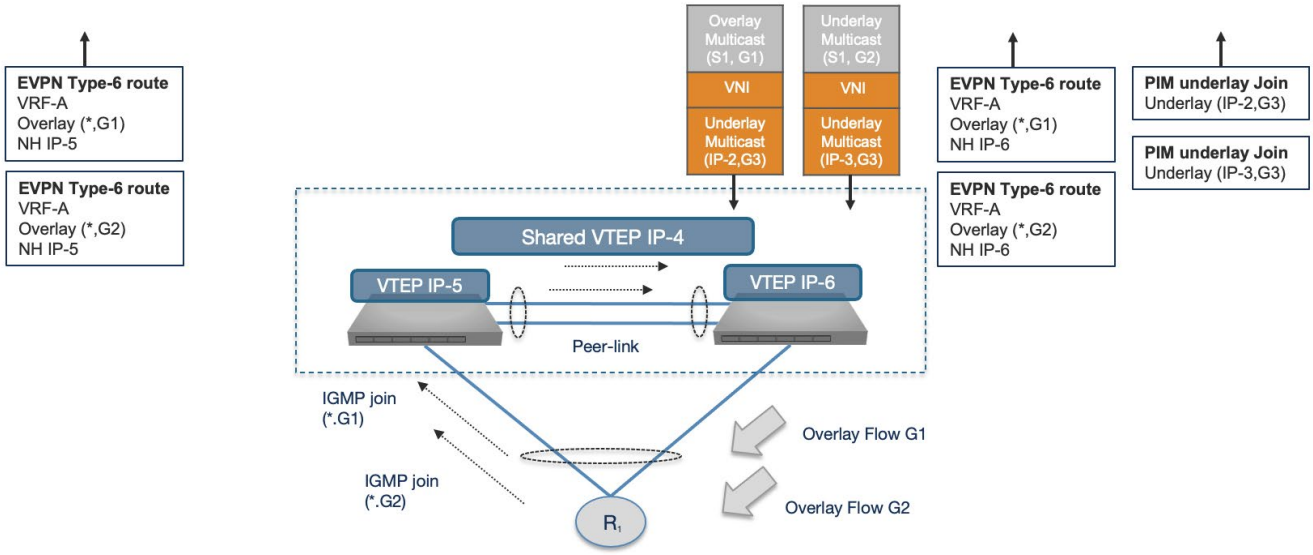

Figure 12: MLAG with EVPN multicast, forwarding model for dual-homed receivers

To achieve this multicast forwarding model, the MLAG nodes are configured with both shared VTEP IP address and the unique VTEP IP address. The shared VTEP IP is used to achieved standard unicast forwarding behavior when advertising type-2, 3 and 5 routes, the unique VTEP IP address is used for the advertisement of the overlay to underlay group mapping within the Type-10 (S-PMSI AD) route, where the unique VTEP IP would be the nodes source IP for the advertised underlay group.

Table 1: Operational behavior of MLAG with EVPN multicast	
Configuration	Requires the configuration of an additional unique VTEP IP address on each node of the MLAG domain
EVPN routes/state	The nodes of the MLAG domain synchronize IGMP state via the peer-link, no additional EVPN routes apart from the type-6 (SMET) routes are required.
Dual-homed Multicast Source	The two nodes of the MLAG domain are able to perform VXLAN encapsulation of the multicast stream for steady-state active-active forwarding.
Dual-homed multicast receiver	Both nodes of the MLAG domain advertise an SMET route, but only the primary node of the domain for the VRF will join the associated underlay group for the interested receiver.
BW Optimisation/Failover	Only the primary node of the MLAG domain joins and receives the underlay group, so under-steady state conditions no additional fabric bandwidth is consumed. However, in the event of a primary node failure, there will be a need to build the PIM state on the new primary node, this has the potential of affecting failover performance.

MLAG Failover

In the MLAG model, EVPN routes are advertised with the shared VTEP IP as the next-hop. With connectivity to the shared VTEP IP advertised by both nodes in the IP underlay, this means load-balancing of the EVPN routes will be achieved in the IP underlay via a 2-way ECMP path. Performing the load-balancing in the network underlay rather than the overlay, greatly reduces the EVPN state churn and simplifies the failover behavior when a link or node failure occurs within the MLAG domain.

In the event of a link failure on a locally attached dual-homed host, the MLAG node experiencing the link failure, updates its MAC table to learn the MAC of the host across the peer-link of the MLAG domain. As both nodes still have connectivity to the host, directly or via the peer-link, there is no need to withdraw the associated type-2 route(s) for the host(s). Traffic destined to the host from a remote VTEP will therefore still be load-balanced to both nodes via ECMP in the underlay.

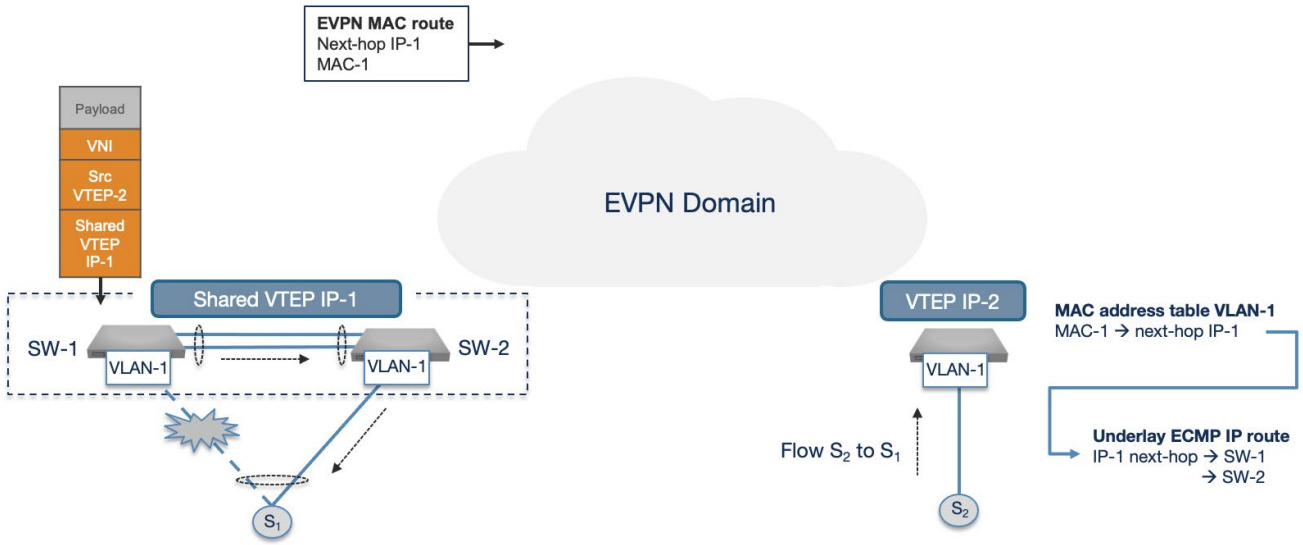

Figure 13: EVPN with MLAG, forwarding behavior after a link failure

In the example above if the VXLAN encapsulated packet is received on SW-2 it will be decapsulated and forwarded to the host via the directly connected active local link, alternatively if the VXLAN encapsulated packet is received on SW-1 it will be decapsulated and forwarded across the peer-link to SW-2 for bridging to the host. Therefore in this failure scenario there is no BGP convergence event in the EVPN overlay or the IP underlay.

In the event of a node failure, the downlink to the host will become inactive and traffic egressing the host will be load-balanced across the remaining links to the active node in the MLAG domain. The node failure will also result in the uplinks to the spine failing and therefore bringing down the associated BGP underlay sessions, this will result in the advertised shared VTEP IP for the failed

node being withdrawn in the network underlay by the Spine nodes. Consequently the remote VTEPs will only learn the next-hop of the advertised EVPN routes for the MLAG domain via the remaining active node of the domain. Thus in this failure scenario, failover can be quickly detected based on the convergence of the IP underlay and the withdrawal of the IPv4 route. There is no need to wait for the EVPN routes for the failed node to be withdrawn by the spine nodes for the remote VTEPs to detect the failure.

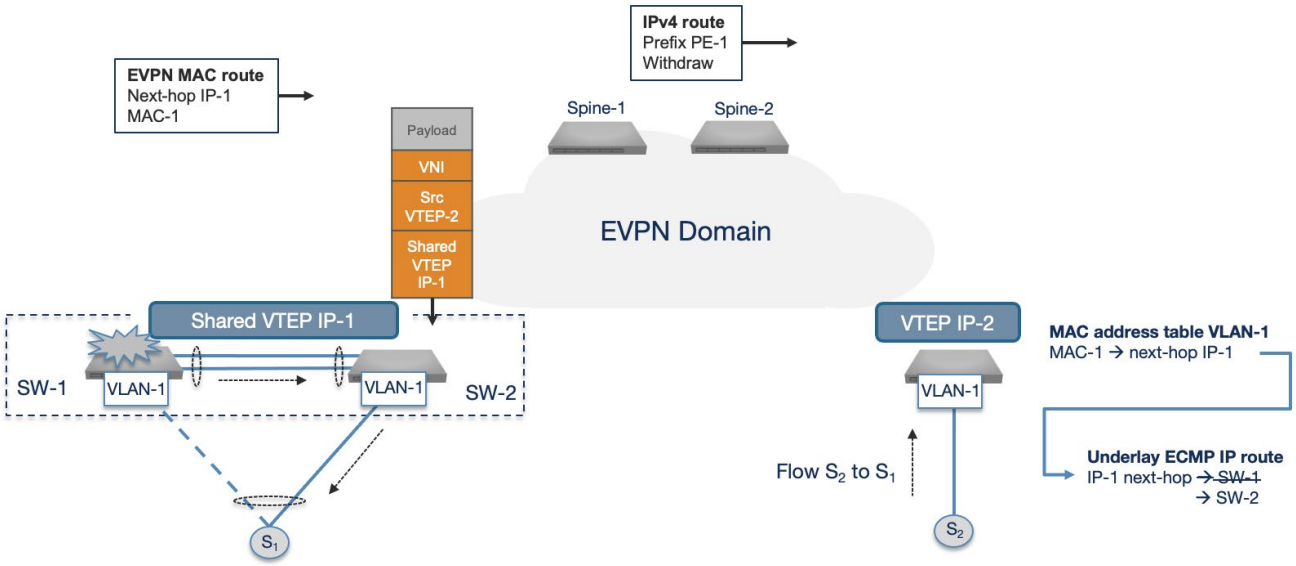

Figure 14: EVPN with MLAG, forwarding behavior after an MLAG node failure

Thus in the MLAG model, by providing a peer-link to interconnect the nodes of the MLAG domain, there is no need for any EVPN route withdrawal and therefore EVPN state churn on the remote VTEPs of the EVPN domain in the event of a local link failure. Secondly by providing load-balancing of the EVPN routes in the IP underlay via the shared VTEP IP, a node failure can be quickly detected through re-convergence of the IP underlay, there is no need to wait for the EVPN overlay routes to be withdrawn for an the remote VTEPs to detect the failure.

EVPN with All-Active multihoming

In the EVPN All-Active (A-A) multihoming model, defined in RFC 7432, the nodes are not interconnected via a peer link, like the MLAG approach, rather peer discovery is achieved with the introduction of two new EVPN route types; Type-4 (Ethernet Segment Route) and Type-1 (Ethernet Auto-Discovery Route). Without the requirement for a peer link, as highlighted in the figure below, an EVPN A-A topology is not restricted to just a pair of nodes.

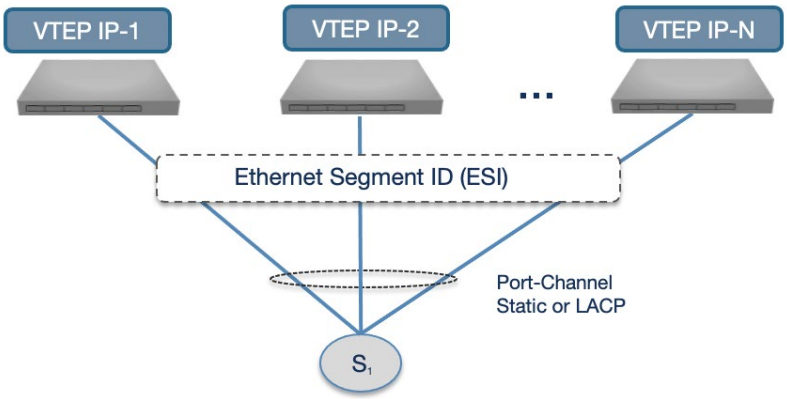

Figure 15: EVPN all-active multihoming topology

In the A-A topology, nodes connecting to the same downstream port-channel are defined as sharing the same Ethernet Segment (ES), which has a unique 10 byte identifier termed the Ethernet Segment Identifier (ESI) for the port-channel. The ESI is configured on all nodes that are members of the same port-channel. Like the MLAG model, an A-A topology is transparent to the downstream multi-homed device which can be configured with either a static or LACP based port-channel with the individual links of the port-channel split across the nodes that are members of the ESI.

The EVPN A-A approach, provides active-active layer 2 and 3 forwarding across the EVPN domain for any multi-homed device, however, unlike the MLAG approach, the nodes connected to the shared ESI, act as independent VTEPs, each configured with a unique VTEP IP address. In the forwarding plane, all nodes are capable of VXLAN encapsulation of locally received traffic on the ESI destined to remote hosts, and decapsulation of VXLAN traffic destined to local hosts on the ESI.

EVPN control-plane with All-Active

Each node in the A-A topology, like the MLAG approach, has a dedicated underlay and overlay peering with each of the spine nodes. In the example below both the underlay and overlay routes are advertised using BGP, although any IGP routing protocol could be deployed for the underlay. BGP is used in the example as its a common design approach for scaling a data center leaf-spine topology. With this topology, each A-A node has a BGP IPv4 peering with each of the spine nodes, and a separate BGP EVPN peering with each of the spine nodes, as each nodes operates as an independent VTEP, the route information advertised in both sessions is different to the MLAG model.

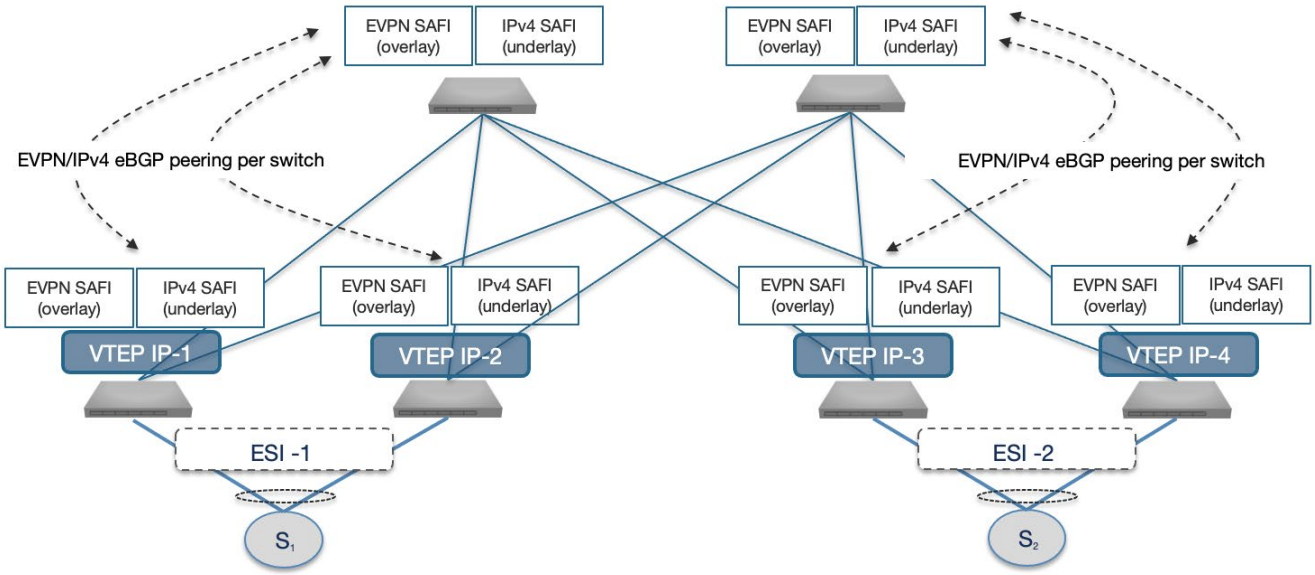

Figure 16: BGP EVPN/IPv4 peering sessions with an EVPN A-A topology

- **BGP IPv4 (AFI 1 /SAFI 1):** This is the underlay BGP IPv4 peering session, and is used to advertise connectivity to the node's unique VTEP loopback IP, which would be used for VXLAN encapsulation and the next-hop IP address of any EVPN route advertised by the VTEP. As stated, the model is not limited to BGP for advertising underlay routes, any IGP routing protocol can be deployed, BGP is a common design within the data center leaf-spine topology for scaling reasons.
- **BGP EVPN (AFI 25 /SAFI 70):** This is the overlay BGP EVPN peering session, and is used to advertise MAC, MAC-IPs and IP-prefixes learnt on the ethernet segment. The type-2 (MAC/MAC-IP) routes originated by a node connected to an Ethernet Segment are advertised with the node's unique VTEP IP as the next-hop and the associated Ethernet Segment Identifier (ESI).

As each node connected to a shared ethernet segment acts as its own independent VTEP, EVPN type-2 (MAC), type-3 (IMET) and type-5 (ip-prefix) are advertised with the node's unique VTEP IP as the next-hop, rather than a shared VTEP IP which would

be the approach in an MLAG topology. In the case of the type-2 routes the associated non-zero ESI is also included in the route advertisement.

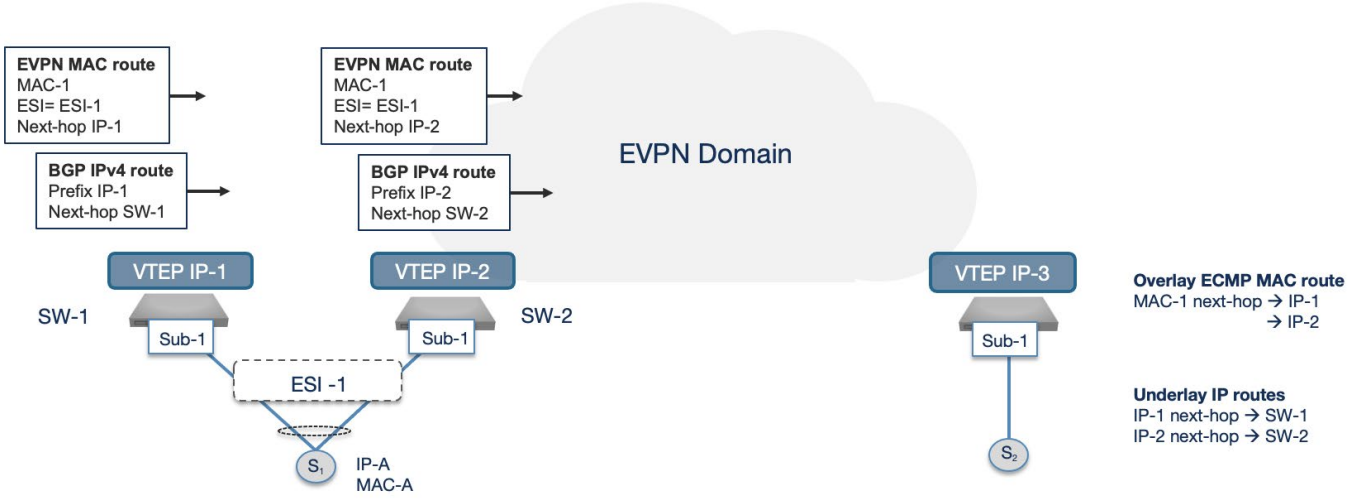

Figure 17: EVPN A-A, Type-2 route advertisements with ESI, MAC and next-hop

Each node on the shared ES will also originate a type-3 (IMET), with the VTEPs unique next-hop, resulting in remote VTEPs populating their flood-list with all nodes connected to the ES.

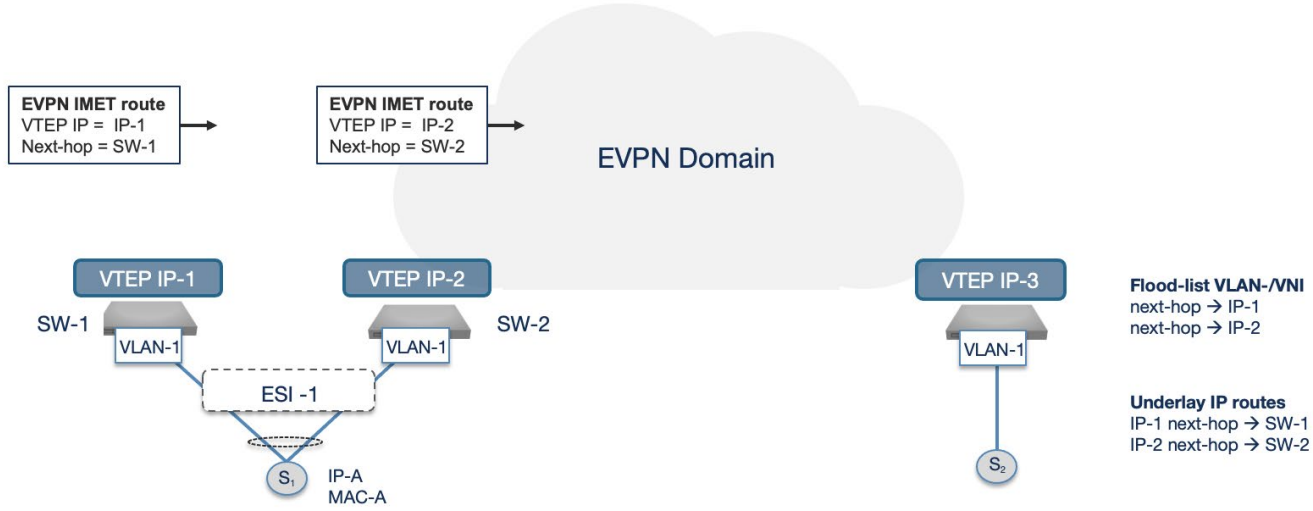

Figure 18: EVPN A-A, Type-3 IMET routes from each VTEP connected to the shared ESI

To avoid duplicate packets when forwarding BUM traffic and provide an active-active forwarding model with fast-failover across the different nodes connected to an ethernet-segment, the A-A topology utilizes additional EVPN type-1 and type-4 routes.

EVPN Type-4 Ethernet Segment Route

The nodes participating in an A-A topology advertise their connectivity to a particular ethernet segment via Type-4 Ethernet Segment (ES) routes. The type-4 route is advertised with a unique route-target (ES-Import Route Target) which is derived from the ESI value of the associated ES. Any node connected to the same ES, will create a rule to automatically import the route based on the RT, allowing dynamic discovery of peer nodes connected to the same ES.

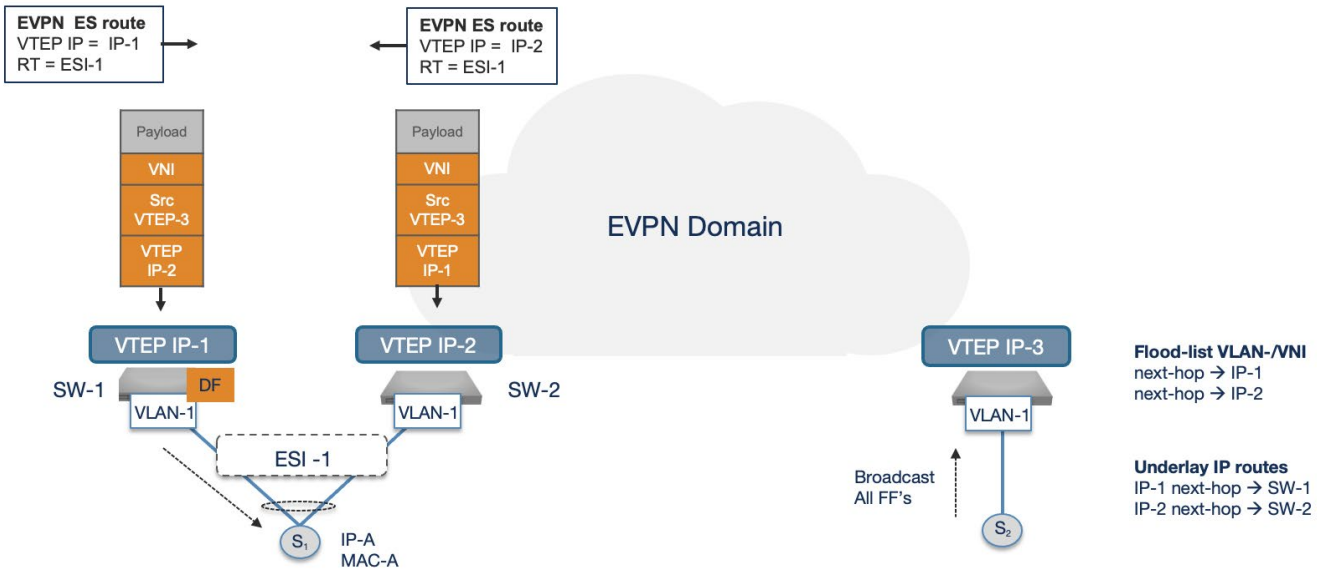

Figure 19: EVPN A-A, BUM traffic forwarding onto the ES by the elected DF node

The type-4 route is also used to elect a Designated Forwarder (DF), for forwarding BUM traffic onto the Ethernet segment. All nodes can forward BUM traffic out of an Ethernet segment but to avoid packet duplication only the elected DF forwards BUM traffic onto the Ethernet segment. A separate DF election is carried out for each EVI configured on the Ethernet segment, with multiple EVIs configured the DF functionality can be load-balanced across the nodes of the ES. In a vlan-based model, there would be a DF election per VLAN (VLAN per EVI) on the Ethernet segment, and for a vlan-aware-bundle model there would be a DF election per EVI (N * VLANs per EVI).

EVPN Type-1 Ethernet Auto-Discovery (AD) Routes

The nodes advertise their connectivity to a ES to remote VTEPs, using type-1 Ethernet Auto-Discovery (AD) routes. The type-1 route has two sub-types; Ethernet AD per Ethernet segment and Ethernet AD per EVI.

The Ethernet A-D per ES route is advertised by a node to announce reachability to a particular ethernet- segment. The main role of the AD per ES route is to facilitate the fast mass withdrawal of MACs, after the loss of connectivity to an ES. In the event a node loses connectivity to an Ethernet Segment (link failure on the port-channel), it will withdraw the type-1 AD per ES route, consequently any remote VTEP will remove the node as a next-hop for any MAC address advertised for the ES being withdrawn. This allows for fast convergence, rather than the remote VTEP waiting for each type-2 route to be withdrawn individually, the withdrawal of a single type-1 AD per ES route, allows the remote VTEPs to quickly remove the node as a next-hop for all MACs learnt on the ES.

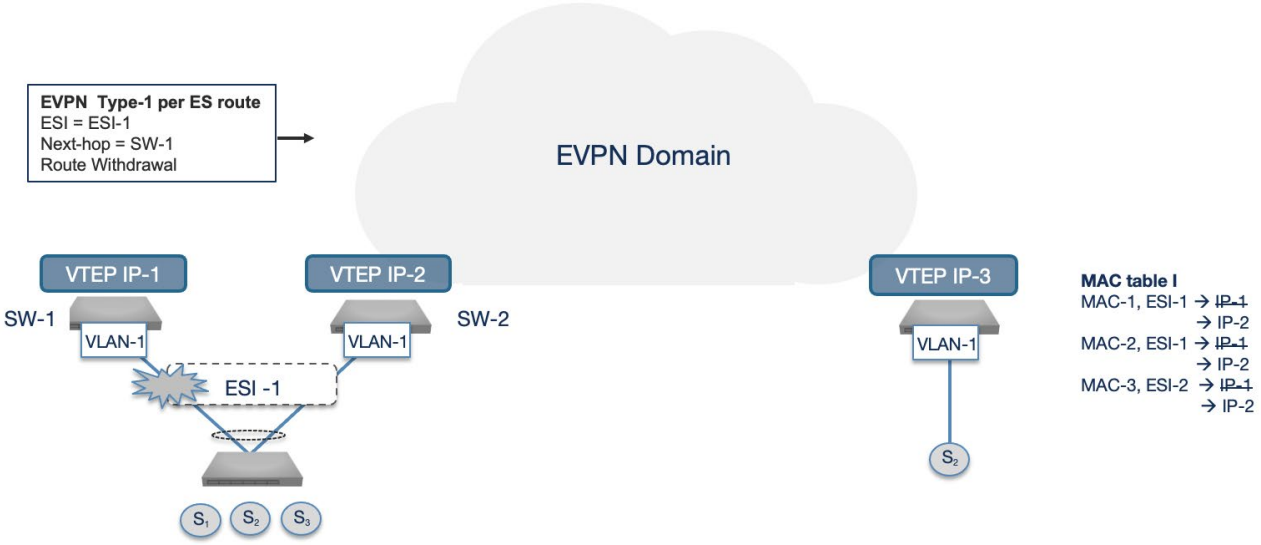

Figure 20: EVPN A-A Mass MAC withdrawal with Type-1 AD per ES route

The Ethernet A-D per EVI route is advertised by a node to announce connectivity to each individual EVI configured on the ES. The main role of the AD per EVI route is mac address aliasing. In an A-A multihoming topology a dual-homed device will load-balance traffic across the active links of the port-channel. The load-balancing algorithm can often result in only one of the upstream nodes receiving the traffic and locally learning the MAC. With only one node learning the MAC address and advertising the resultant type-2 MAC route, remote VTEPs will only learn and forward traffic to a single next-hop rather than load-balance traffic across all nodes connected to the EVI on the Ethernet segment.

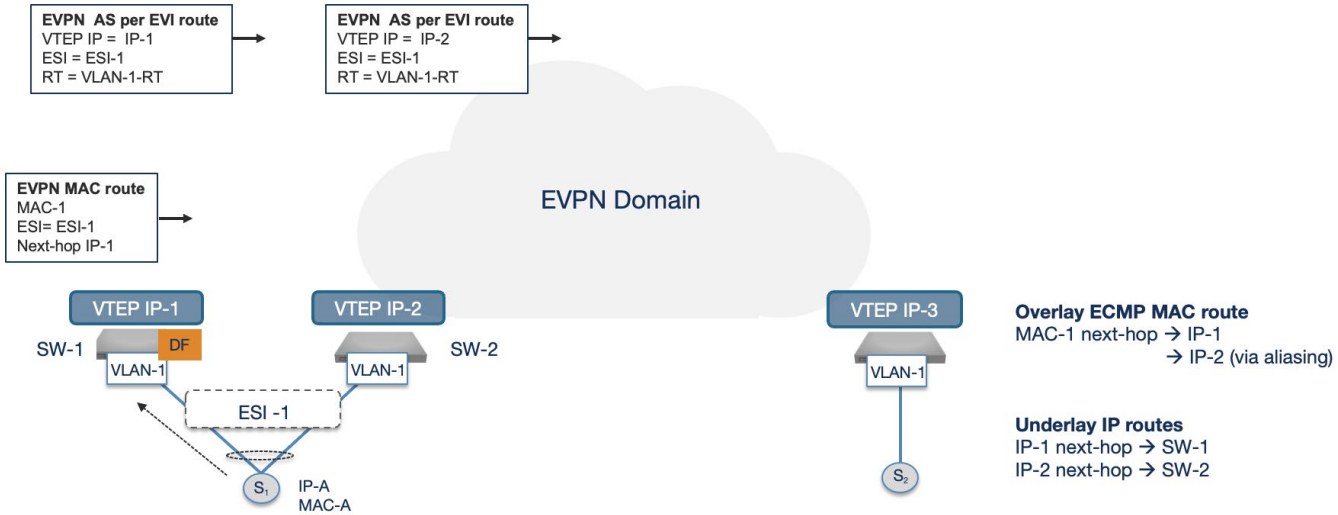

Figure 21: EVPN A-A aliasing with with Type-1 AD per EVI route

This load-balancing inefficiency is addressed with MAC address aliasing. With MAC address aliasing, the reachability to a host connected to an EVI on an ES, is determined using a combination of the type-2 MAC route which contains the associated ES and EVI and the nodes advertising connectivity to the ES and EVI through the type-1 A-D per EVI route. Thus a remote VTEP doesn't need to receive a type-2 route from each node attached to the ES, in order to load-balance traffic across all nodes connected to the ES. With this MAC aliasing approach, a node can only be added as a next-hop for the MAC, if both type-1 routes (auto-discovery per EVI and auto-discovery per ES) are advertised for the associated ES.

MAC-IP proxy-bit

The MAC aliasing model only applies to MACs in the type-2 route, in the case of a symmetric IRB model where type-2 MAC-IP routes are also advertised, creating host routes (/32) in the VRF routing table, support for the "proxy MAC-IP" bit is required to build an ECMP path for the host-route. With "proxy MAC-IP" enabled across the nodes on a ES, when a node locally learns the MAC-IP binding of a host on the Ethernet segment, it advertises a type-2 MAC-IP route with the associated ESI. Any node connected to the same ES, on receiving the type-2 MAC-IP route, if they haven't already advertised a type-2 route for the binding, will re-advertise the route with the next-hop changed to their local VTEP IP address, the route is advertised with the proxy-bit set to indicate its been proxied rather than locally learnt. Consequently remote VTEPs will receive a type-2 MAC-IP host-route from each VTEP on the ES, resulting in routed traffic to the host being layer 3 load-balanced across all VTEPs of the ES. The nodes within an ES will also process and install the type-2 routes received from peer nodes connected to the same Ethernet segment. This action of synchronizing routes, ensures locally learnt MAC-IP bindings by one node are learnt by all other nodes connected to the same ES.

Broadcast, Unknown unicast and Multicast (BUM) Traffic

In an A-A topology, BUM traffic received from a locally attached host on an ethernet-segment, is flooded to remote VTEPs based on the flood-list of the VNI, which is populated by type-3 (IMET) route advertisements. As nodes of the ES act as an independent VTEPs, they will each advertise a type-3 route, consequently they will be members of the flood-list and receive the BUM traffic. On receiving the BUM traffic, the nodes connected to the Ethernet segment perform split-horizon forwarding by checking the source IP of the VXLAN encapsulated frame, only forwarding the BUM traffic on interfaces that are not connected to ESs shared with source VTEP IP of the frame. With this split-horizons approach, it is the responsibility of the originating node to replicate the BUM traffic to all directly attached Ethernet segments regardless of the DF election, a functionality referred to as local bias forwarding. In the case of BUM traffic received from a remote VTEP which is not a member of the ES, all nodes on the shared Ethernet segment will receive the BUM traffic due to their unique type-3 IMET route advertisements, however only the elected DF for the EVI will forward the traffic onto the local Ethernet segment.

Traffic Load-balancing

In the A-A approach, whether it be MAC-aliasing or proxy MAC-IP the load-balancing of traffic across the nodes of the ES will be achieved in the network overlay. The advertised MACs and MAC-IP will have multiple next-hops (each VTEP on the ES) in the overlay network. This is a different approach to the MLAG model, where MACs and host-routes have a single next-hop in the overlay, which is the shared logical VTEP, with the load-balancing achieved in the underlay by each node advertising connectivity to the shared logical VTEP. This change in the load-balancing behavior will have an effect on how a failure on the ES is detected and processed by nodes across the EVPN domain (see the "Failover" section for more details).

State synchronization

In the A-A model, the nodes sharing connectivity to an Ethernet segment are not required to be interconnected via a peer link, this can provide a major advantage over the MLAG topology, as ports and cables are not wasted interconnecting nodes, further it removes the restriction on the number of VTEPs that can be used to connect to a single ES, thereby providing an improved level of resiliency over an MLAG topology. These benefits do mean additional EVPN state in comparison to an MLAG topology, as the A-A topology utilizes EVPN type-1 and type-4 routes rather than a peer-link and shared VTEP IP, to synchronize state, and provide active-active forwarding and fast-failover in the event of a failure. The amount of additional EVPN state can be considerable and will depend on the number of Ethernet segments configured on the nodes and the number of EVIs (VLANs) active on each ES. Taking a data center Compute leaf as an example, where 30 servers are deployed within a rack, with the servers dual-homed via a port-channel to a pair of VTEP nodes configured in an A-A topology. With the servers hosting multiple VMs, each port-channel is configured with 10 VLANs, which map to a unique EVI on the VTEP nodes. With this type of topology, the additional EVPN state created for dual-homing the servers within a single rack would be:

- **Type-4 (Ethernet segment) routes:** With a type-4 route being advertised for each of the 30 Ethernet segments configured on each VTEPs,this would mean 60 type-4 route advertisements.
- **Type-1 (Auto-Discovery per ES) routes:** With a type-1 AD per ES route being advertised for each of the 30 Ethernet segments configured on each VTEPs; this means 60 type-1 (AD per ES) advertisements.
- **Type-1 (Auto-Discovery per EVI) routes:** With a type-1 AD per EVI route being advertised by each VTEP, for each VLAN configured on each of the 30 Ethernet segments; this means (2 x 30 x 10) 600 type-1 type-1 (AD per EVI) advertisements.
- **Total EVPN A-A Route advertisement for all servers in the rack:** 60 type-1 (AD per ES) + 60 type-1 (AD per ES) + 600 type-1 type-1 (AD per EVI) = 720 EVPN routes advertisements

Thus while the A-A approach can provide operational benefits (no peer-link, improved levels of resiliency) in comparison to an MLAG topology, the additional EVPN state that it generates within the topology should also be taken into consideration when comparing both models.

Spanning Tree

The EVPN standard doesn't define a mechanism for an EVPN A-A topology to interact with a downstream spanning tree domain. If the downstream multi-homed devices are end-nodes meaning there is no need to interact with a spanning-tree topology but rather protect against end-host sending STP BPDUs, STP BPDU-Guard can be enabled on the interfaces of the VTEPs connecting to the ES. If the downstream devices are instead layer 2 switches and there is therefore a need to interact with the Spanning-Tree topology, Arista provides a spanning tree "super-root" functionality. With the "super-root" functionality, all VTEP nodes connected to the same ethernet segment are configured with a single shared "super-root" bridge-id, forcing the nodes to be the root bridge for the downstream Spanning-Tree domain. This approach can be extended across all VTEPs within the EVPN domain, where they are all configured with the same "super-root" bridge-id, resulting in the EVPN domain being seen as a single Spanning-Tree bridge, to any downstream layer 2 switch.

Layer 2 nodes single-homed

In the A-A model, node's connected to a shared ES act as independent VTEPs and consequently advertise EVPN routes with a unique next-hop rather than a shared next-hop, which would be the case in the MLAG model. This approach provides the benefit that traffic forwarding to a single-homed host on a node (whether it be due to topology requirements or a consequence of a link failure on a peer node of the ES) will always follow the optimal path to the VTEP directly connected to the single-homed device, rather than traverse the peer-link which would be the case in an MLAG topology.

Anycast Gateway

To provide EVPN Integrated Routing and Bridging (IRB) for directly attached hosts, A-A supports the same anycast GW solution as MLAG. The anycast GW, is a virtual MAC and IP address, that is configured on each of the VLANs shared between the nodes of the A-A topology. With the virtual GW configured, all the VTEPs connected to the ES are able to respond to ARPs destined to the virtual IP and route traffic destined to the virtual MAC, thus providing an active-active routing model across all nodes. This means regardless of how traffic is load-balanced on the port-channel from the host, the node on the ES receiving the traffic is able to route the traffic directly to the destination subnet, thus ensuring optimal layer 3 routing regardless of how traffic is load-balanced by the host.

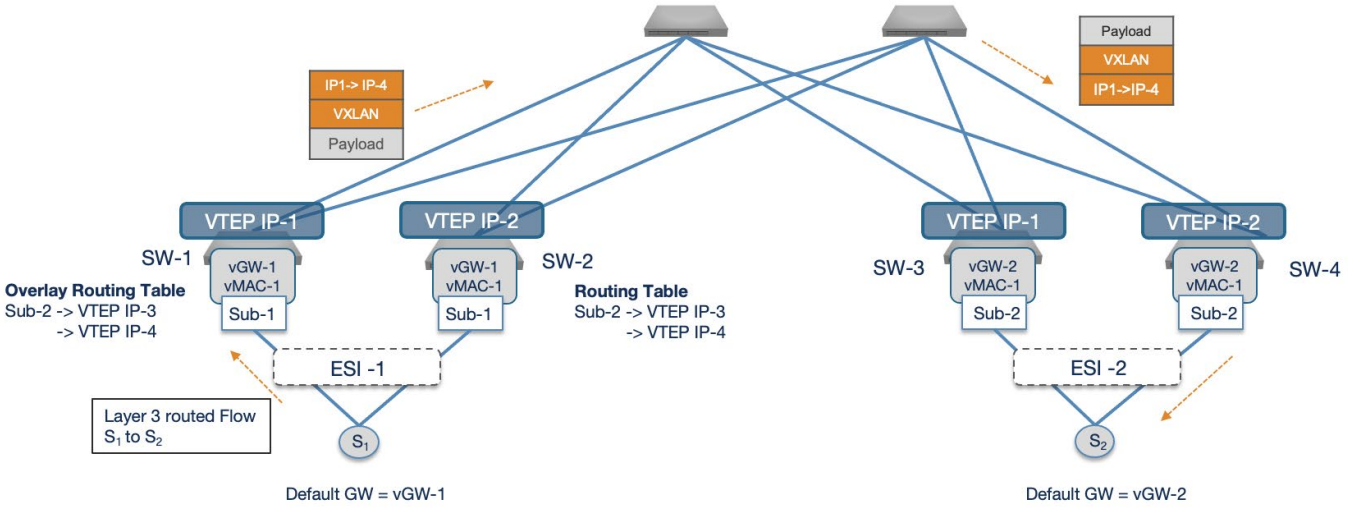

Figure 22: EVPN with A-A and anycast GW for layer 3 forwarding

A-A with Fastpath return

In the anycast GW model, traffic routed by a node within the A-A topology is forwarded with the source MAC of the node's system MAC, rather than the virtual MAC which is only used for ARP responses to the virtual GW IP. This forwarding behavior will have an adverse effect when connecting network appliances supporting "Fastpath" or "Symmetric return", where the GW MAC address is learnt in the forwarding path by inspecting the source MAC of the received routed packet rather than an ARP response from the GW. In an A-A topology, this would mean traffic forwarded by the "fast-path" appliance will use the system MAC of one of the nodes of the ES as the destination MAC, due to load-balancing on the port-channel, any node on ES may receive the traffic, if the destination MAC is not owned by the receiving node, it will be VXLAN bridged over the EVPN domain to the relevant node rather than route the traffic directly.

To provide support for the "Fastpath" model, while maintaining optimal first-hop routing, the nodes within the A-A topology provide support for advertising their system router MAC for an associated GW IP address, as defined in RFC 7432 (Section 10.1). In this approach each node that is a member of the ES, will advertise a type-2 route for the virtual GW IP, included in the route is a default GW community used to advertise the node's unique router MAC. Any node configured with the same virtual GW IP, programs the advertised MAC as a local router MAC, thereby allowing all nodes in the ES to route traffic directly even when the destination MAC is owned by another node of the ES.

Layer 3 nodes dual-homed

A layer 3 node can be dual-homed to the A-A topology, while removing the need to configure a shared router mac, which would be required in an MLAG topology. In the typical deployment model the layer 3 node would be connected via dedicated layer 3 point-to-point links to each node of the ESI for resiliency, with an IGP/BGP session running across the point-to-point links to exchange routes, with the routes advertised as EVPN Type-5 prefixes into the EVPN domain. The configuration is highlighted in the figure below.

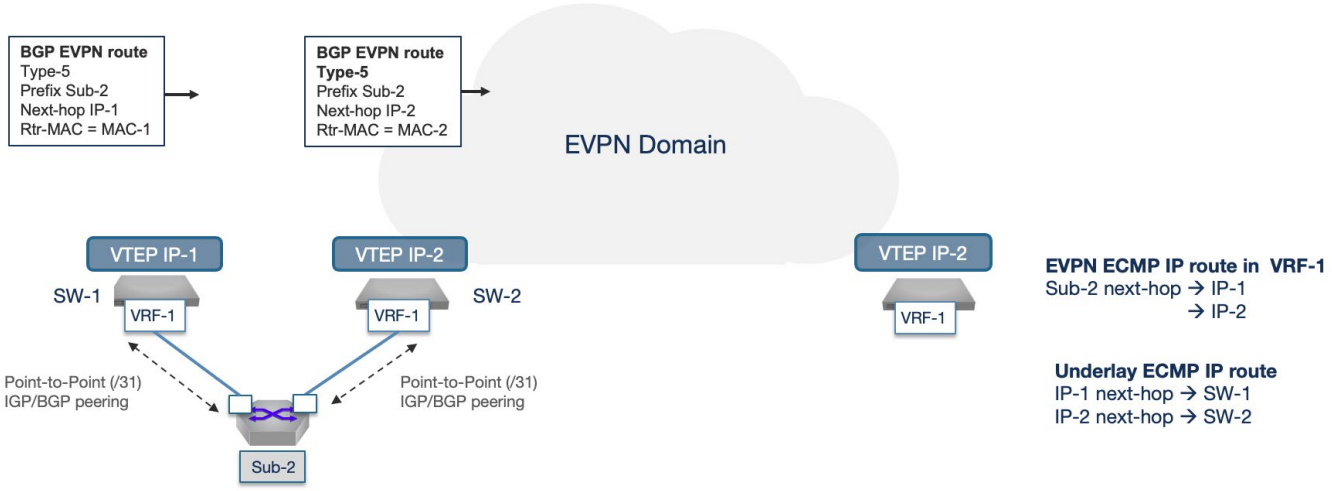

Figure 23: EVPN A-A topology with dual-homed L3 nodes

As the nodes of the A-A topology act as independent VTEPs, the Type-5 routes would be advertised by each node with a unique next-hop and a router-mac, representing the node's VTEP interface. Consequently remote VTEPs will have two unique routes to the prefix, resulting in a 2-way overlay ECMP path for the prefix with a next-hop of each VTEP. By acting as independent VTEPs and generating unique next-hops for the type-5 routes, the A-A EVPN topology inherently supports the ability to connect Layer 3 nodes to the topology while maintaining optimal layer 3 forwarding for the advertised prefixes. This is in comparison to the MLAG model, which requires support for a shared router MAC to ensure traffic isn't routed across the peer-link.

Layer 3 nodes single-homed

To support a single-homed layer 3 node in an MLAG topology, there is a requirement to configure a multi-VTEP IP address for the physical node in addition to a shared virtual VTEP IP address, this is to ensure the traffic is routed directly to the correct node rather than traverse the peer-link of the MLAG domain. In the A-A topology this additional multi-VTEP configuration is not required, as the nodes in the A-A topology act as independent VTEPs. Thus routes exchanged with the single-homed layer-3 node are advertised as type-5 routes into the EVPN domain with a next-hop and a router-mac, representing the node's unique VTEP IP address. With the next-hop of the type-5 route representing the node's unique VTEP IP, rather than a logical VTEP IP (which would be the case for an MLAG topology), traffic will follow the optional path and be routed directly to the VTEP where the layer 3 node is connected.

EVPN Multicast with A-A

In an EVPN multicast deployment, both multicast sources and receivers can be dual-homed to the EVPN domain using A-A multi-homing. To synchronize IGMP state across the nodes of a shared ES, the A-A approach introduces an additional set of EVPN routes; 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 VTEPs on the same ES. Similarly the Type-8 route (leave-sync) is used to synchronize locally received leave messages between VTEPs on the Ethernet segment.

From a multicast source perspective in the A-A topology, both nodes are able to VXLAN encapsulate a multicast stream received from a locally attached source. The actual node performing the encapsulation of a specific stream, would be based on the source's load-balancing of the stream across the port-channel. The receiving node is able to VXLAN route or bridge the multicast flow as required, thus providing an active-active forwarding model where the multicast stream from the source will always follow the optimal path, regardless of which node on the ES receives the stream.

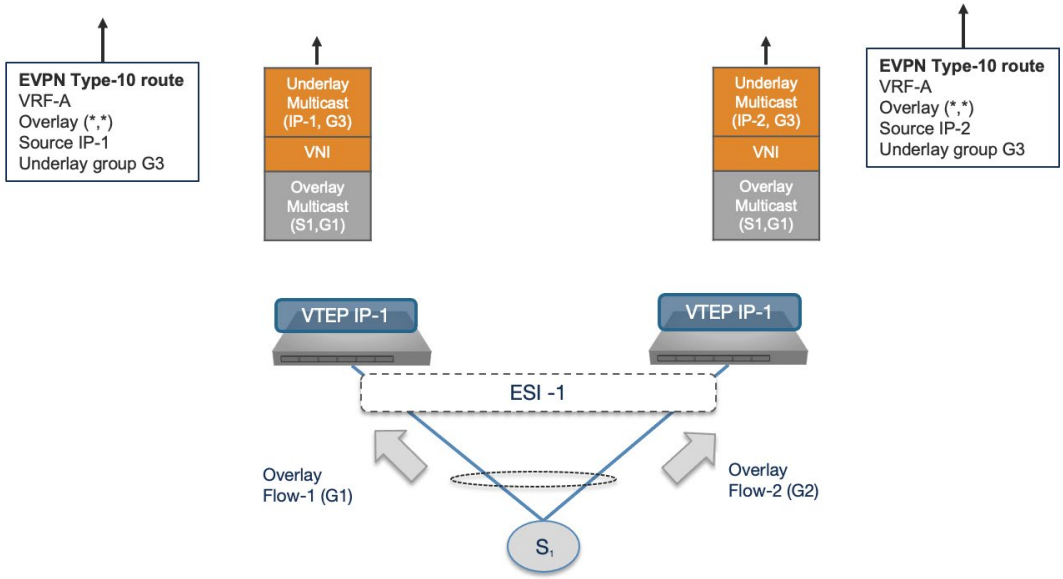

Figure 24: EVPN A-A with dual-homed multicast source

In a PIM underlay solution, each node of the ESI will advertise a unique underlay (S,G) group for transporting the VXLAN encapsulated packet, with the underlay to overlay group mapping advertised by the node in a type-10 (S-PMSI) route. To receive the VXLAN encapsulated multicast stream, remote VTEP with interested receivers would join the advertised underlay group.

For multi-homed multicast receivers, the elected Designated Forwarder (DF) for the bridge-domain of the interested receiver, is responsible for advertising the associated type-6 SMET route. As the IGMP join from the receiver, due to load-balancing on the port-channel, could be received by any VTEP connected to the ES, the VTEP receiving the IGMP join will advertise a EVPN type-7 (sync-join) route. The type-7 route is imported by all other VTEPs on the ES, and used to populate their local IGMP snooping table, however only the elected DF will advertise an SMET route in response to receiving the type-7 route. Although only the DF advertises the SMET route, in a PIM underlay solution all nodes on the ES will join the associated underlay group for the stream's VRF. Thus under steady state conditions, all VTEPs connected to the ES, will receive the multicast stream for the locally attached receiver. To avoid duplicate delivery, the DF node will decapsulate the VXLAN packet and forward the multicast stream onto the ES, the non-DF nodes will drop the received VXLAN frame.

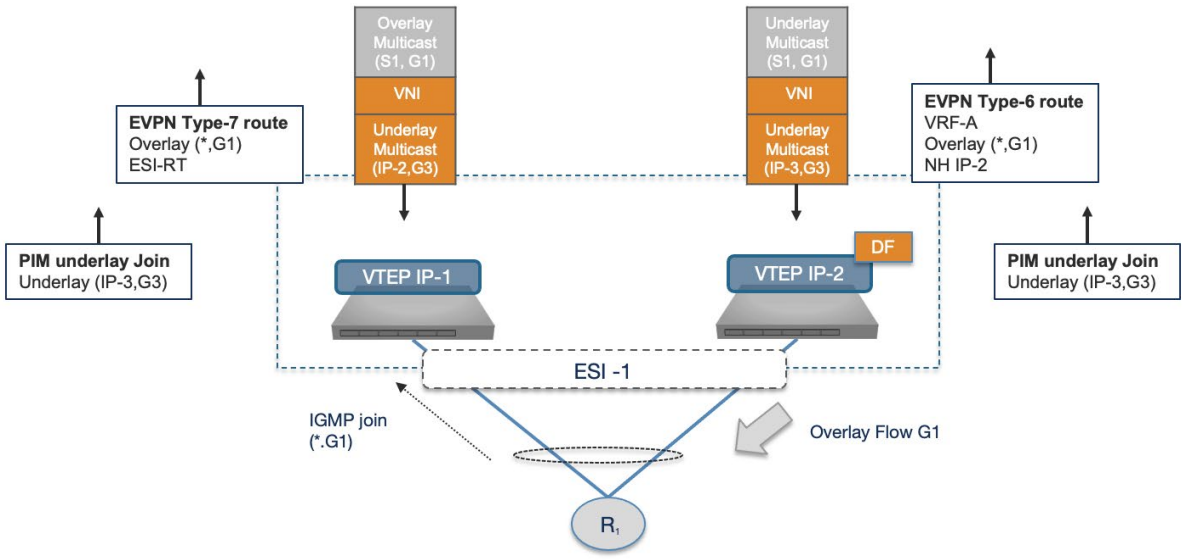

Figure 25: EVPN A-A with dual-homed multicast receiver

Unlike the MLAG multicast forwarding model, which requires the configuration of an additional unique VTEP IP address, there is no extra configuration required on an A-A topology to support either a dual-homed multicast receiver or source. The table below provides a summary of the A-A operation for both multi-homed multicast sources and receiver.

Table 2: Operational behavior of All-Active with EVPN multicast	
Configuration	No additional VTEP configuration required
EVPN routes/state	To synchronize IGMP state across the nodes of the ES, utilizes type-7 and type-8 routes in addition to the type-6 (SMET) route
Dual-homed Multicast Source	All VTEPs nodes on the ES are able to perform VXLAN encapsulation of a multicast stream from a locally attached source, thus providing steady-state active-active forwarding.
Dual-homed multicast receiver	Elected DF responsible for advertising the SMET route, all nodes of the ES join and receive the associated underlay group, but only the DF is responsible for forwarding the stream onto the ES.
BW Optimisation/Failover	All nodes on the ES join and receive the underlay group, so during -steady state conditions additional fabric bandwidth is consumed. However building PIM state on the non-DF nodes will improve failover performance in the event of a DF failure.

All-Active Failover behavior

In the EVPN A-A topology, under steady state-conditions load-balancing across all nodes connected to an ES is achieved through MAC aliasing and the use of the "proxy IP-MAC" functionality. In the event of a link failure to a local multi-homed host, the node experiencing the link failure, will withdraw its advertised type-1 route for the ES and any associated type-2 routes. On receiving the type-1 (AD per ES) withdrawal, remote VTEPs will remove the node as a next-hop for all MACs learnt on the ES. This mass MAC withdrawal provides a fast failover mechanism in the event of an ES link failure, as there is no need for the remote VTEPs to wait for each individual type-2 on the ES to be withdrawn, although it will require the remote VTEPs to process the type-1 withdrawal and shrink their ECMP path for each MAC associated with the ES. With the VTEP removed as a next-hop for any MAC on the ES, traffic will be forwarded via ECMP to the remaining active VTEPs on the Ethernet Segment. To avoid any blackholing of traffic during this re-convergence, the node exhibiting the local link failure can also be configured to pre-calculate a backup path for the withdrawn routes via a VXLAN tunnel to peer node connected to the same ES, this is termed VTEP PIC edge.

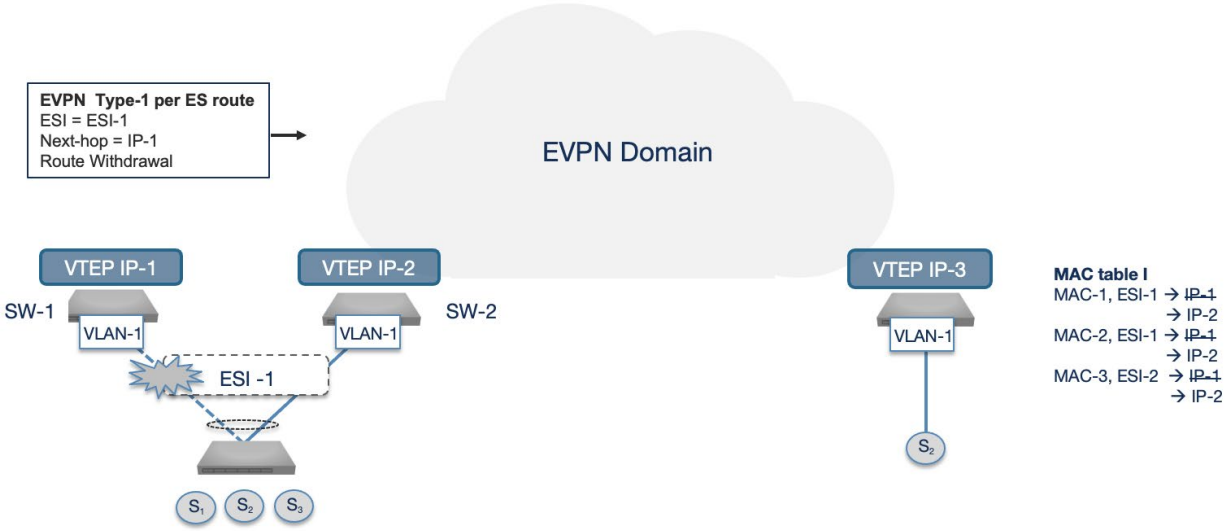

Figure 26: EVPN A-A, failover behavior in the event of an ESI link failure

In the event of a node failure, the downlink to the host will be disabled and traffic egressing the host will be load-balanced across the remaining links of the port-channel to the active VTEP(s) of the ES. The node failure will also result in the uplinks to the spine failing and therefore bringing down the associated BGP IPv4 underlay sessions. Consequently the spine nodes will withdraw the underlay route for the failing node's VTEP IP, unable to resolve the next-hop of any advertised EVPN route, the EVPN routes will also be withdrawn by the spine nodes. The remote VTEPs within the EVPN domain, will therefore need to process the withdrawal of both underlay IP routes and overlay EVPN routes. On receiving the withdrawn underlay IP route and overlay EVPN routes, the failing VTEP will be removed as next-hop for any MACs learnt on the ES, this means each remote VTEP will shrink it's overlay ECMP group to contain only the remaining active VTEPs of the ES.

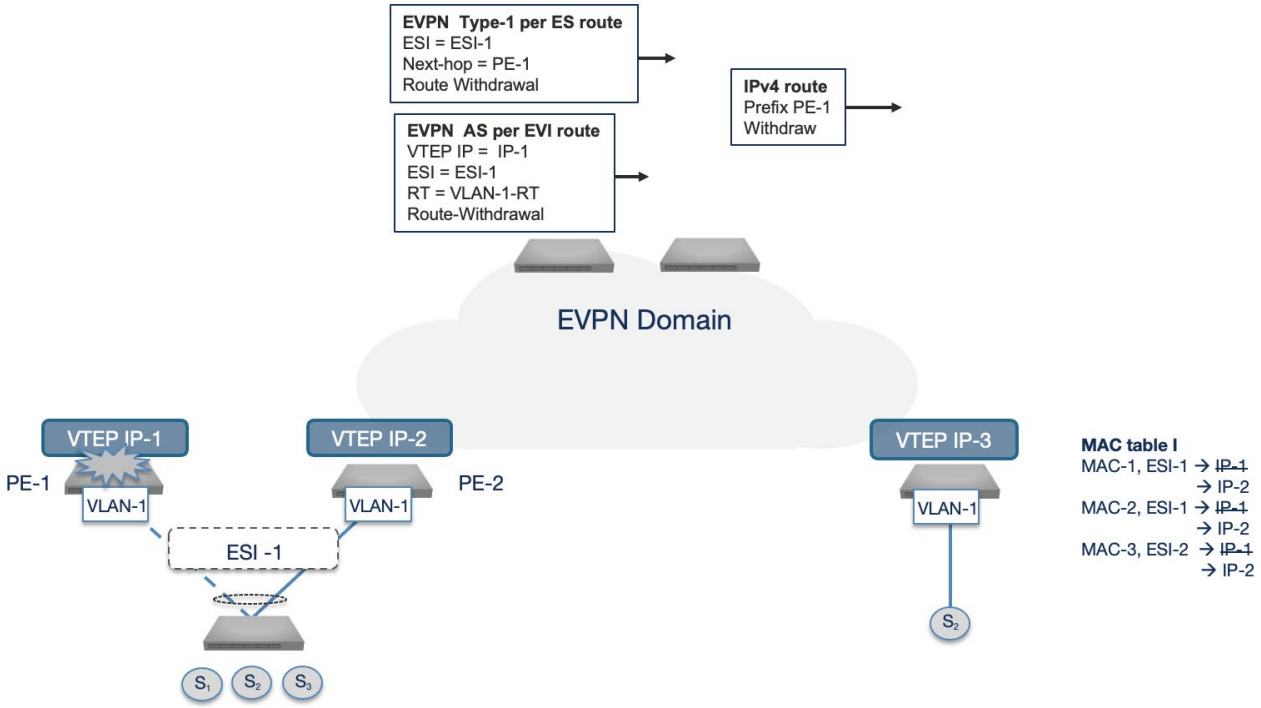

Figure 27: EVPN A-A multihoming behavior in the event of a PE failure.

With the A-A model relying on both the convergence of the underlay and EVPN overlay during a failover event, when compared to a similar MLAG topology, the A-A approach will result in more state churn across the EVPN domain and potential slower convergence. For example, with a link failure in the MLAG model, due to the peer-link and the shared VTEP IP address, there is no need to withdraw any EVPN routes and consequently no EVPN state churn on the remote VTEPs. For an A-A model, there is a requirement to withdraw the type-1 route for the ES and all associated type-2 routes learnt on the ES, and more importantly each remote VTEP in the EVPN domain needs to process the withdrawn routes, shrinking their overlay ECMP forwarding table for any MAC or MAC-IP learnt on the ES. This additional EVPN state churn also holds true for the node failure scenario, in an MLAG model the IP underlay route to the shared VTEP IP is withdrawn by the spine, the remote VTEPs are only required to shrink their underlay ECMP route to the shared VTEP IP, there is no need to alter the forwarding table of EVPN overlay routes as the next-hop is unchanged. Alternatively, with the A-A model, where the load-balancing is being achieved in the overlay, the withdrawal of the IP underlay and EVPN overlay routes will result in each remote VTEP re-program their ECMP overlay group for the MAC and MAC-IP routes learnt on the ES.

Comparison of MLAG and EVPN All-Active models

As outlined Arista's EOS software supports two models for providing active-active multi-homing of downstream devices within an EVPN domain; MLAG and EVPN All-Active. While both approaches provide support for an active-active multi-homing topology and are not mutually exclusive as they provide interoperability, the perceived benefit of one approach over another will be very much dependent on individual requirements and their priority within the overall design, e.g Level of resiliency, cabling within the rack, failover performance, scaling, multicast requirements, etc. The table summaries the operation of both approaches, and their perceived benefits based on specific deployment requirements.

Table 3: Operational comparison between MLAG and EVPN All-Active

Feature/Requirement	MLAG	EVPN All-active
Nodes in a Domain	Support for 2 nodes in an MLAG domain	Support for up to 16 nodes in an ES
Inter-switch link	Yes, used for synchronizing layer 2 state and link failures	No requirement, synchronization of state achieved via BGP EVPN routes (type-1 and type-4)
VTEP IP Address	Single shared VTEP IP between both nodes of the MLAG domain, and is the next–hop used by both nodes for any advertised EVPN route.	Unique VTEP IP for each node attached to the ES, which would be the next-hop for any EVPN route advertised by the VTEP.
Load-balancing	Load-balancing in the underlay. EVPN routes are advertised with a single share next-hop, which is the shared VTEP IP. The means load-balancing can be achieved in the underlay via ECMP to the shared next-hop.	Load-balancing in the overlay, EVPN routes are learnt with unique next-hops to each VTEP attached to ES. ECMP in the overlay to load-balance the traffic to each VTEP connected to the ES.
EVPN Multicast	Support for dual-homing multicast receivers and sources. Utilises type-6 SMET routes for advertising local IGMP joins, IGMP state sync via the peer-link. Both nodes within the domain are able to VXLAN bridge/route multicast traffic for any directly attached source. For multicast receivers only one VTEP joins the associated underlay group, saving fabric bandwidth but can result in multicast traffic being forwarded across the peer-link.	Support for multi-homing multicast receivers and sources. Utilises type-6 SMET routes for advertising local IGMP joins, IGMP state sync using type-7 and type-8 routes. All VTEPs connected to the ES are able to VXLAN bridge/route multicast traffic for any directly attached source. For multicast receivers only the DF forwards the traffic to the interested receiver, however, all VTEPs join the associated underlay group, consuming more fabric bandwidth but offering faster failover in the event of a DF failure.
EVPN state	Doesn't require any additional EVPN routes to function, with state synchronize achieved via the peer link	Introduces new EVPN routes; type-1 (AD per ES) and type-4 (ES) for each active ES on the VTEP and type-1 (AD per EVI) routes for each active EVI on the ES. The additional EVPN routes can be considerable when deploying multiple ESs on a VTEP with multiple VLANs on each ES.
Failover	ECMP re-convergence in the underlay. Consequently fast failover, as re-converge is achieved by the withdrawal of the shared VTEP IP in the underlay, minimal EVPN state churn in on the remote VTEPs.	ECMP re-convergence in the overlay. Consequently slower failover at scale, as re-converge is now achieved by each remote VTEP by processing the withdrawn EVPN overlay routes.
Spanning Tree	Inherent SPT support with MLAG. The MLAG domain acts as a single logical switch, from a spanning tree perspective, with the primary node of the domain responsible for processing and sending BPDUs downstream	No inherent support within EVPN. Each VTEP connected to an ES acts as an independently Spanning-tree bridge. Requires vendor specific support (Arista "super-root" functionality), to allow VTEPs on the ES to behave as a single Spanning-Tree bridge to any downstream L2 switch,
Fastpath return support	Supported within an MLAG topology, when "mlag peer mac routing" is enabled on both VTEPs of the MLAG domain.	Inherent supported within an A-A topology, with the configuration and advertisement of Default GW community MAC.
Single homed L2 nodes	Supported with potential for suboptimal forwarding via the peer-link, as type-2 routes are advertised with a next-hop of the shared VTEP IP.	Inherent support, type-2 routes advertised with a unique next-hop.

Table 3 (contd.): Operational comparison between MLAG and EVPN All-Active		
Feature/Requirement	MLAG	EVPN All-active
Single homed L3 nodes	Requires a multiple VTEP IP configuration. Otherwise potential for suboptimal forwarding via the peer-link	Inherent support, type-5 routes advertised with a unique next-hop
Interoperability	Arista nodes within an MLAG domain, interop with remote VTEPs configured as either MLAG domains or in an EVPN A-A topology.	IETF Standards based, Arista or third-party nodes within a shared ESI, interop with remote VTEPs configured as either MLAG domains or in an EVPN A-A model.

Summary

Arista's EOS software supports two flexible models for providing active-active multi-homing of downstream devices within an EVPN domain; MLAG and EVPN all-active. The models are not mutually exclusive as they offer interoperability within the same EVPN domain, the preference for one model over another will depend on the specific requirements of the design e.g. Brownfield vs Greenfield, complexity, STP interaction, cabling standards, level of redundancy, and single-homing demands. As outlined in the whitepaper both models take different approaches to addressing each of these requirements, thereby offering different levels of benefit for each. Therefore, the importance of an individual requirement(s) within the final design should be taken into account when choosing between an EVPN All-Active model and an MLAG approach.

Reference Materials

RFC 8365: A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)

RFC 9251: IGMP and MLD proxy for EVPN

EVPN Optimized Inter-Subnet Multicast (OISM)

IP-MAC-proxy: draft-rbickhart-evpn-ip-mac-proxy-adv-01

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