ARISTA

# EVPN Unleashed: Scaling Multi-Domain Fabrics with EVPN Gateways

28 April 2026 - SwiNOG #41

Remi Locherer <remi@arista.com>

modus one
smart networks

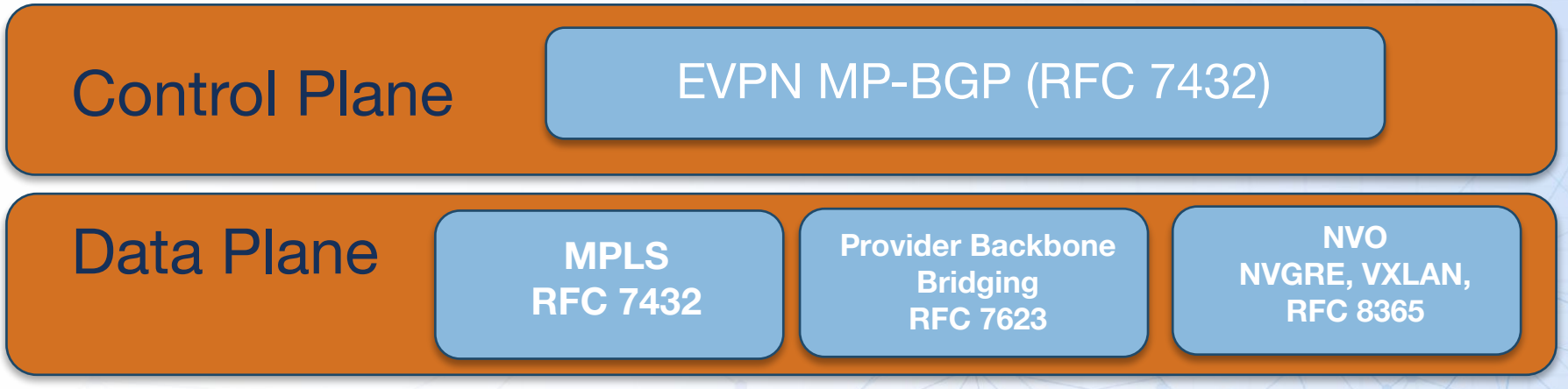
# Quick EVPN Refresher

EVPN, IETF defined standard RFC 7432

- Specifics BGP control plane and new address family to advertise MAC/IP and IP prefixes.
- Providing Layer 2 and 3 VPN services on single interface, with a single MP-BGP control plane.

Multiple forwarding plane options, with an equivalent BGP EVPN control plane

- RFC 7432 – MPLS forwarding plane – Metro and WAN focus
- RFC 8365 – VXLAN, NVGRE, MPLSoGRE – Data Center focus
- RFC 7623 – Provider Backbone Bridging – Metro Ethernet focus



modus one ARISTA

# EVPN: Core Route Types

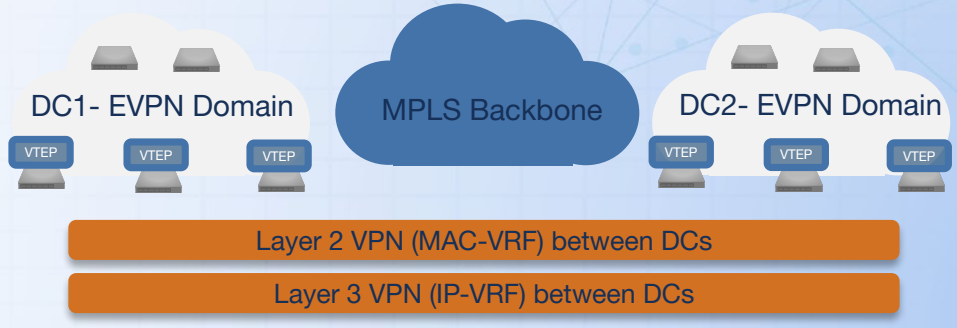
| Route Type | Description |
|---|---|
| 1 | Auto-Discover Segment route<br>For A/A multi-homing forwarding, and to allow remote discovery of dual-homed Segments. |
| 2 | MAC address Route<br>Advertisement of locally learnt/provisioned MAC address and optionally IP addresses. |
| 3 | Inclusive Multicast Ethernet Route<br>For advertisement EVI membership for the creation of ingress replication lists |
| 4 | Ethernet Segment Route<br>For multi-homing deployments to allow Peer discovery on same segment and DF election |
| 5 | IP prefix Route<br>Advertisement of a IP prefix and next-hop, no MAC address for the route is advertised. |

# EVPN GW – Hierarchical EVPN for Scaling and DCI

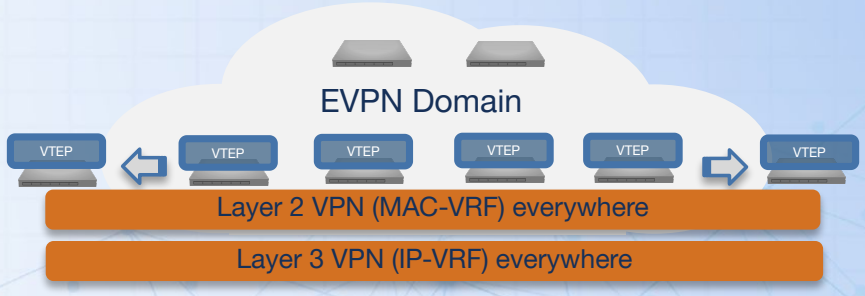
## EVPN DCI challenge – need to extend L2/3 VPNs across MPLS

- Need  to extend Layer 2 & 3 VPN between geographically dispersed DCs
- MPLS deployed in the backbone for traffic engineering or FRR reasons
- EVPN-VXLAN interop with IP-VPN at the DC Edge PE
- For both layer 2/3 requirement EVPN-VXLAN interop with EVPN-MPLS on the




## EVPN scaling challenge - VNIs everywhere across ever larger LS fabrics

- Challenge on the size/amount of flood-lists on individual nodes/VTEPs
- Amount of EVPN state across nodes in a single domain – MAC, ARP, IP & IMET routes
- Level of EVPN state churn across all nodes, during any failure or network change
- Requirement to segment the EVPN domain and introduce hierarchy for scale



modus one ARISTA

# EVPN GW – Hierarchical EVPN for scaling and DCI

IETF BESS working group, number of RFCs/Drafts for EVPN GW behavior

- Support for both Layer 2 (type-2 & 3) and Layer 3 (type-5) DCI solutions
- Interop across different BGP Address families and data-plane encapsulations (VXLAN, PBB, MPLS)

| Draft | Overview |
|---|---|
| A Network Virtualization Overlay Solution using EVPN **RFC 8365** | EVPN control plane for L2 VPNs with an NVO environment with VXLAN, NVGRE and GENEVE encap– DCI using GWs and DCI using ASBRs |
| EVPN and IP-VPN Integrated Solution **draft-ietf-bess-evpn-ipvpn-interworking** | Layer 3 DCI interop between EVPN-VXLAN/MPLS and IP-VPN WAN for layer 3 DCI |
| Multi-site EVPN based VXLAN using Border Gateways **draft-sharma-bess-multi-site-evpn** | GW DCI solution focused only on EVPN-VXLAN, support for a single control planes (EVPN) and single data-plane (VXLAN) |
| Interconnect Solution for EVPN Overlay networks **RFC 9014** | EVPN GW solution for L2 interconnecting of multiple control planes (VPLS/EVPN) and data-planes (MPLS, VXLAN, PBB) |
| EVPN multicast forwarding for EVPN to EVPN GWs **draft-rabnic-bess-evpn-mcast-eeg** | EVPN GW solution for providing seamless multicast interconnect between EVPN domains, across VXLAN and MPLS data-planes |
| Domain Path (D-PATH) for Ethernet VPN (EVPN) Interconnect Networks **draft-ietf-bess-evpn-dpath** | D-path community for EVPN routes to provide loop-free route advertisement between EVPN domains for layer 2. |

L3 GW solution

Industry adopted L2 GW

modus one ARISTA

# EVPN GW – EVPN VXLAN/MPLS GW

### Standard based solution

- RFC 9014 & evpn-ipvpn-interworking
- Multiple encap support (VXLAN/MPLS/SR)
- Standards based Multi-homing support

### EVPN GW for Hierarchical scaling

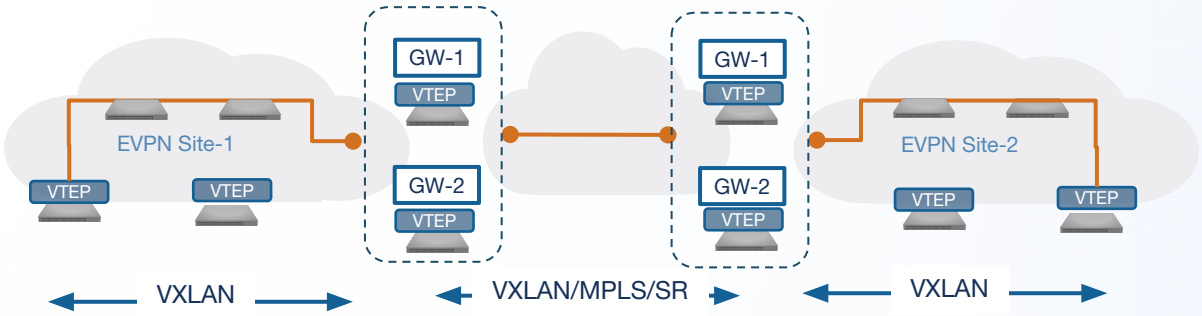
EVPN GW for scaling EVPN-VXLAN deployments inter-POD and intra-site (DCI) by introducing hierarchy

### Layer 3 interconnect

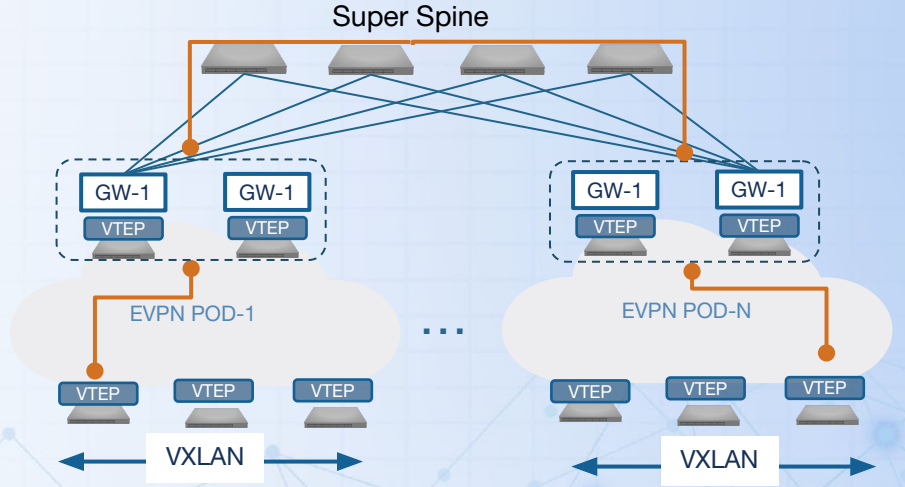
- Layer 3 (type-5) interconnect between domains
- Type-5 routes re-advertised with GW next-hop

### Scalable L2 interconnect

- GW scoping of Type 1,4 and 3 routes
- Flood-list scale with split-horizon forwarding of BUM traffic on GW
- Type-2 re-originated with GW next-hop




DCI for interconnecting Datacenters



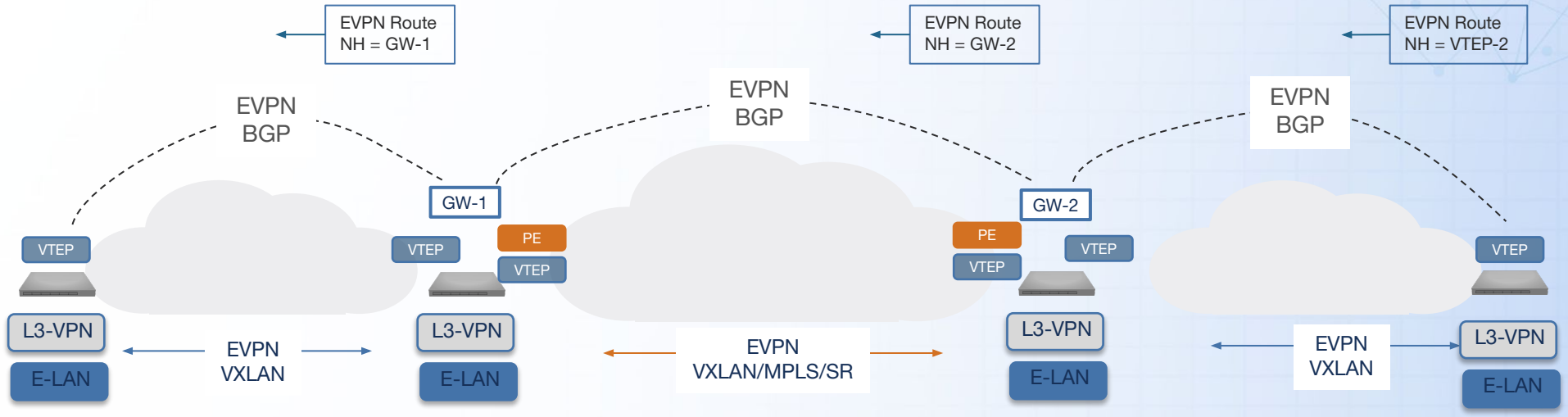

Interconnecting PODs for hierarchical scaling

modus one ARISTA

# EVPN GW – EVPN VXLAN/MPLS GW

## EVPN Gateway Solution




### EVPN GW behavior

- PE/VTEP nodes EVPN peer with their local GW node via eBGP or iBGP
- GW node EVPN peer with the GW nodes in the remote domain via eBGP or iBGP
- Import received type-2 & 5 routes based on RT policy
- Export type-2 & 5 routes between domains based on RT policy
- When exporting between domains, new Next-hop, encap and label

### Benefits

- End-to-End Layer 2 and 3 connectivity regardless of interdomain encap
- Support L2 and L3 VPN between VXLAN VTEPs and MPLS PE nodes
- EVPN A-A for GW redundancy for L2 interconnect across domains
- Hierarchical flood-list for BUM traffic forwarding
- Reduction in EVPN state churn across domains

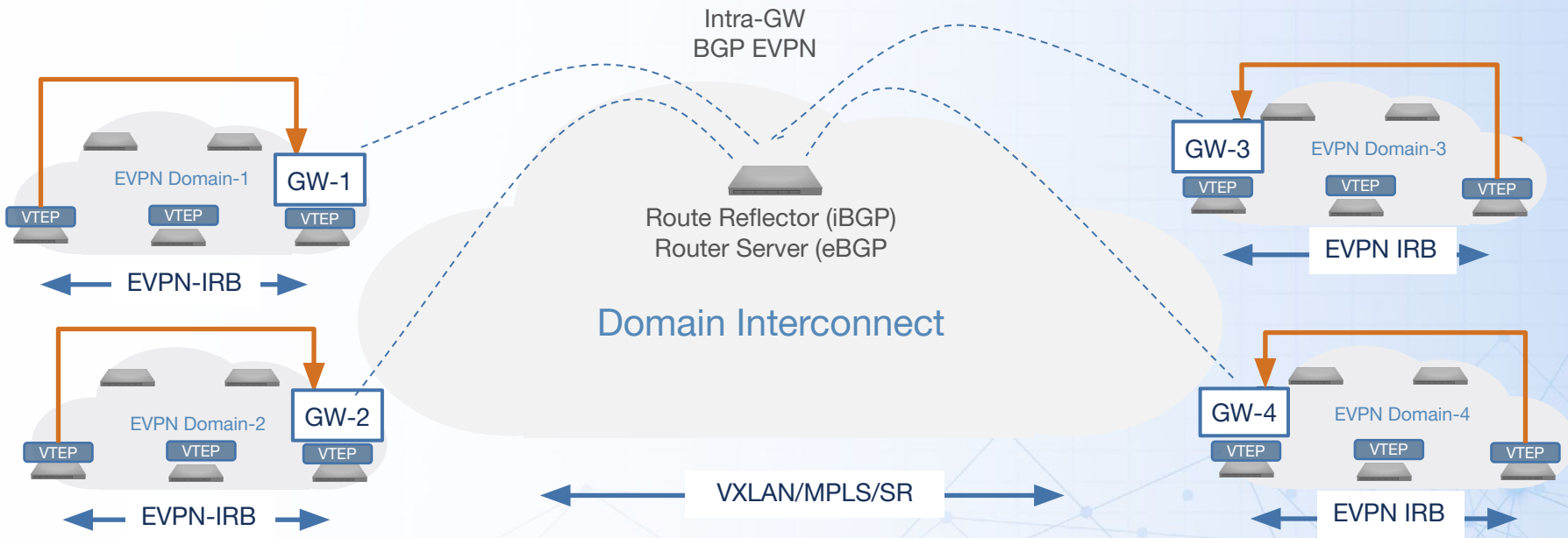
 

modus one ARISTA

# EVPN GW – EVPN VXLAN/MPLS GW

## EVPN Gateway Solution

- Not just a point-to-point DCI solution
- Support for multiple domains, improved DC scale with inter-POD(s) or inter-Site(s) connectivity
- Support for all BGP topology variants in the Local and remote domain (ebGP and iBGP)



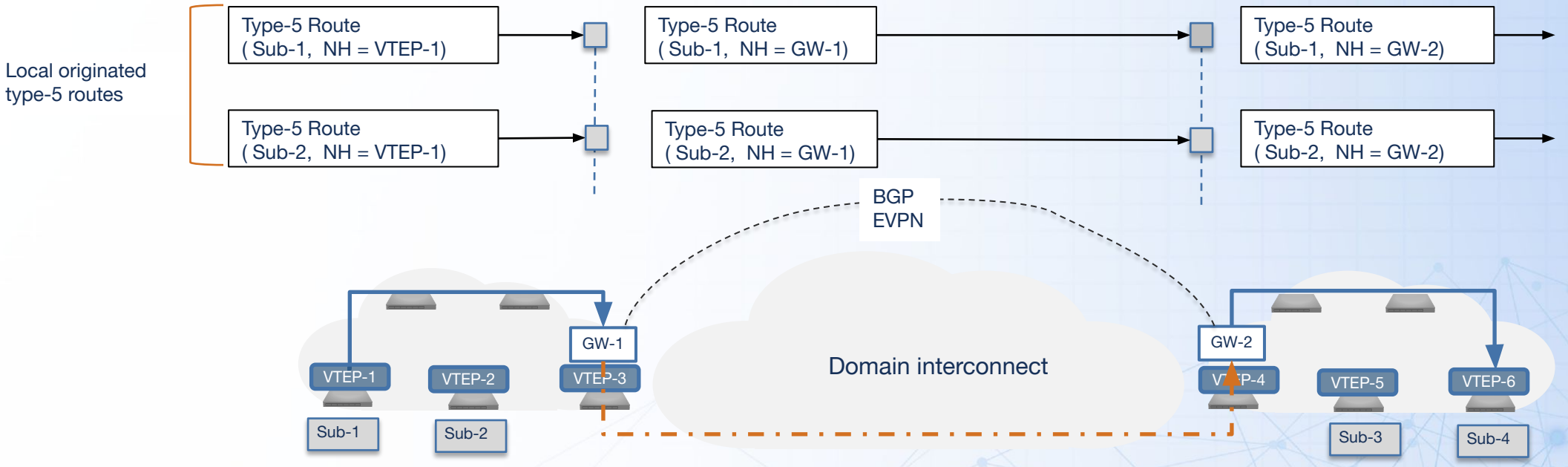
# EVPN Layer 3 GW Model

# EVPN L3 GW – Type-5 route behavior

EVPN GW Layer 3 forwarding model

- Defined in *ietf-bess-evpn-ipvpn-interworking* draft
- GW Re-advertises local type-5 routes, with next-hop changed to the GW
- Hierarchical as the loopback IPs (NH) of the local domain VTEPs are hidden by the GW

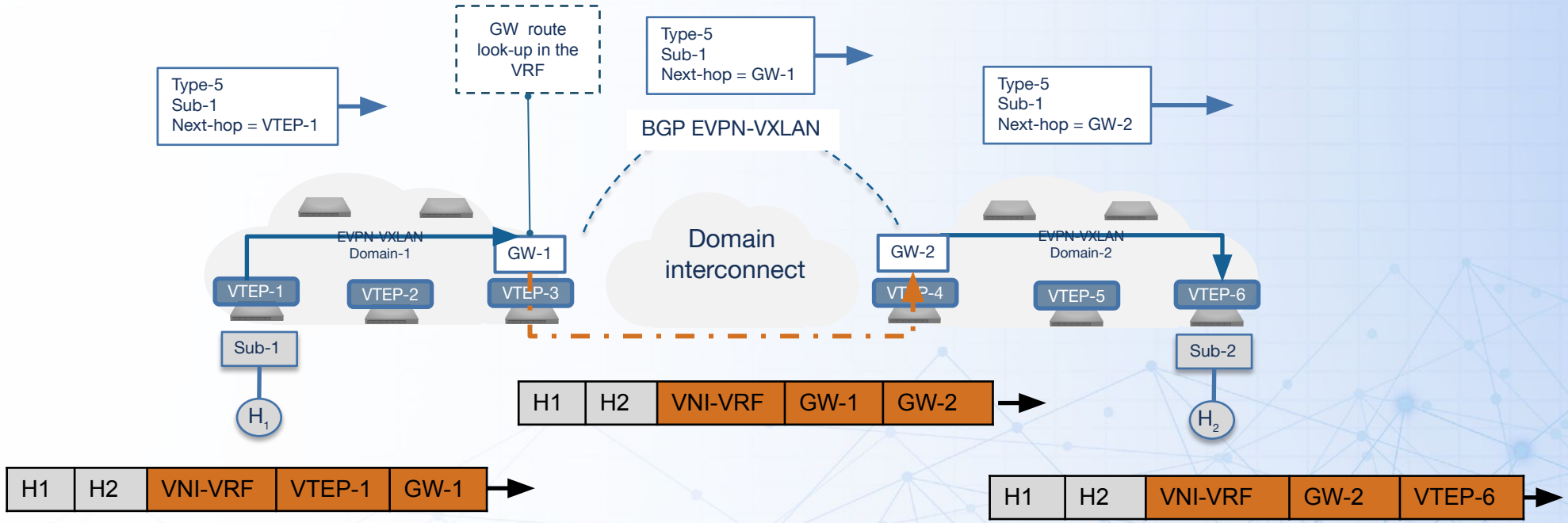


# EVPN L3 GW – Layer 3 forwarding behavior

The EVPN GW node is in the L3 VXLAN forwarding plane

- Traffic routed to the GW, as remote subnet (Sub-2) learnt with a next-hop of the GW
- GW performs VXLAN decap, with a route lookup in the VRF based on the L3
- GW performs VXLAN encap, as route lookup resolved to a next-hop of GW-2
- Remote GW performs a similar VXLAN decap/encap action before routing to the local VTEP

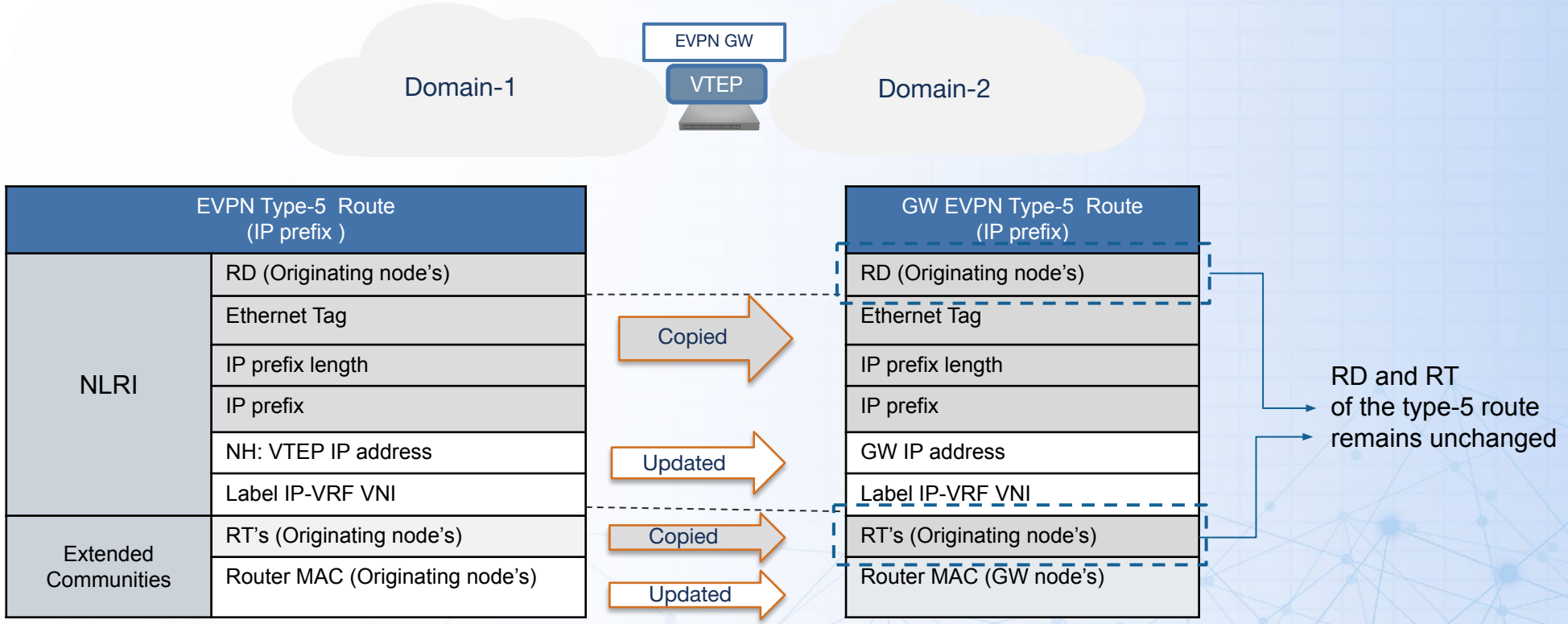


# EVPN L3 GW – Arista implementation detail

Arista EVPN layer 3 GW Layer 3 implementation

- In the model, the neither RT or the RD are changed on the route
- There is no best-path calculation, all type-5 routes are re-advertised to the EVPN peer

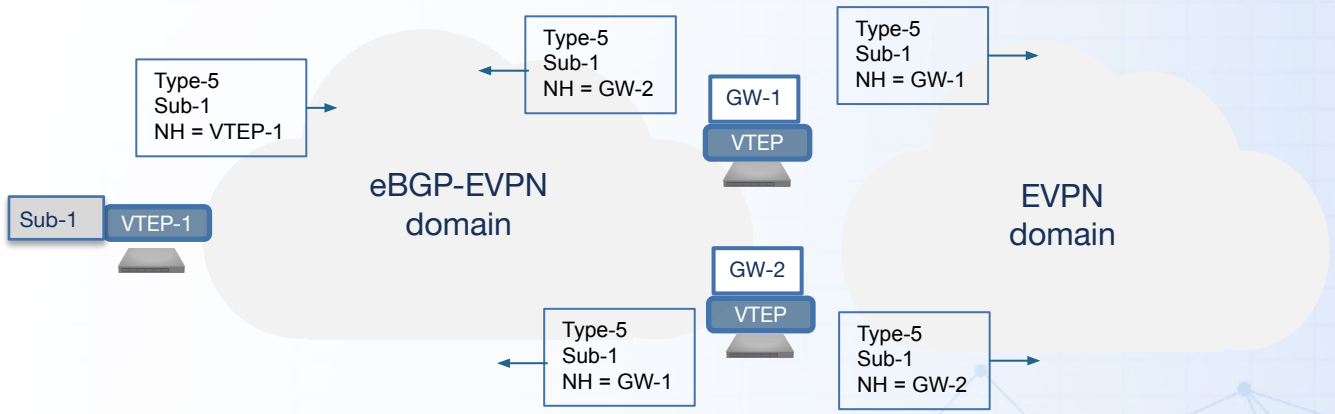


# EVPN L3 GW – Redundancy

## EVPN L3 GW redundancy

- Typically two GW deployed for connectivity to a domain for redundancy reasons
- Potential for a routing loop, as Type-5 routes could be advertised back into the domain by the peer GW
- Requirement to configure RCF policy on the GW to mark Site-of-Origin or use **D-PATH community**



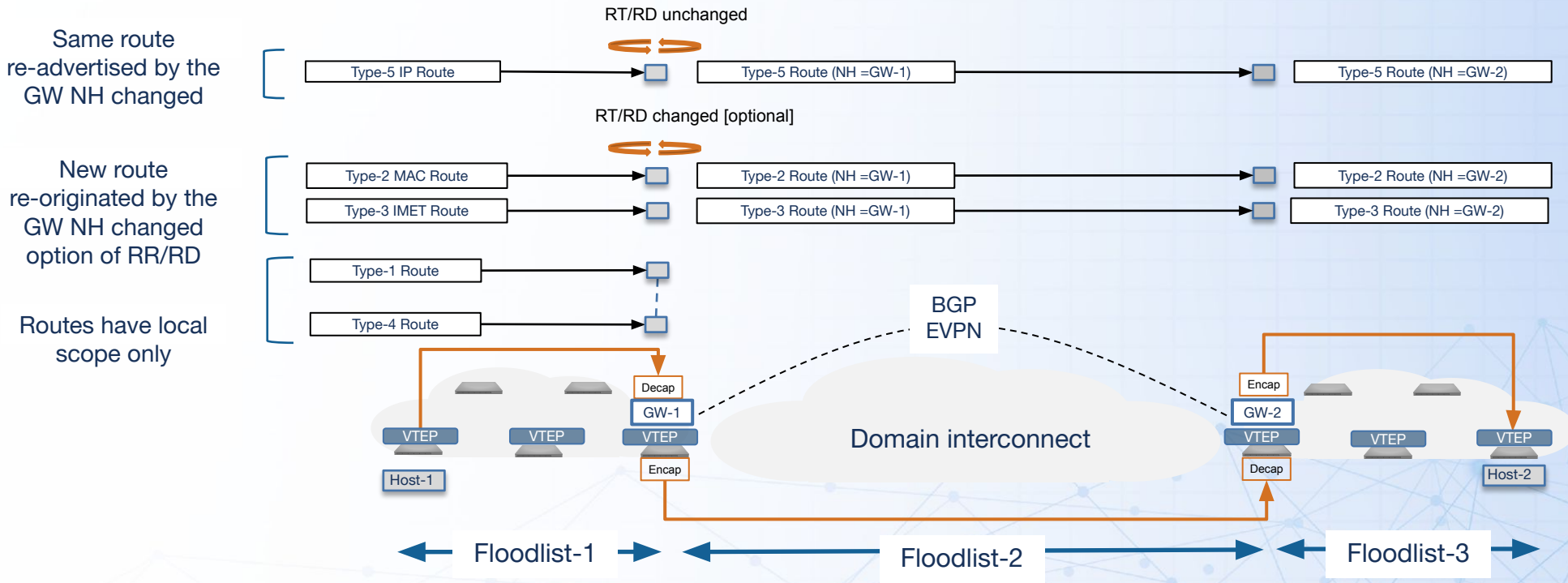
# EVPN Layer 2 GW Model

# EVPN L2 GW – RFC 9014

EVPN Layer 2 GW control-plane

- Does NOT re-advertise locally received Type-1, 3 or 4 routes for any BD stretched between domains
- Re-originates any locally received type-2, with the NH changed, RT/RD changed [optional]
- Advertises a new type-3 route, between GWs with NH changed, RT/RD changed [optional]

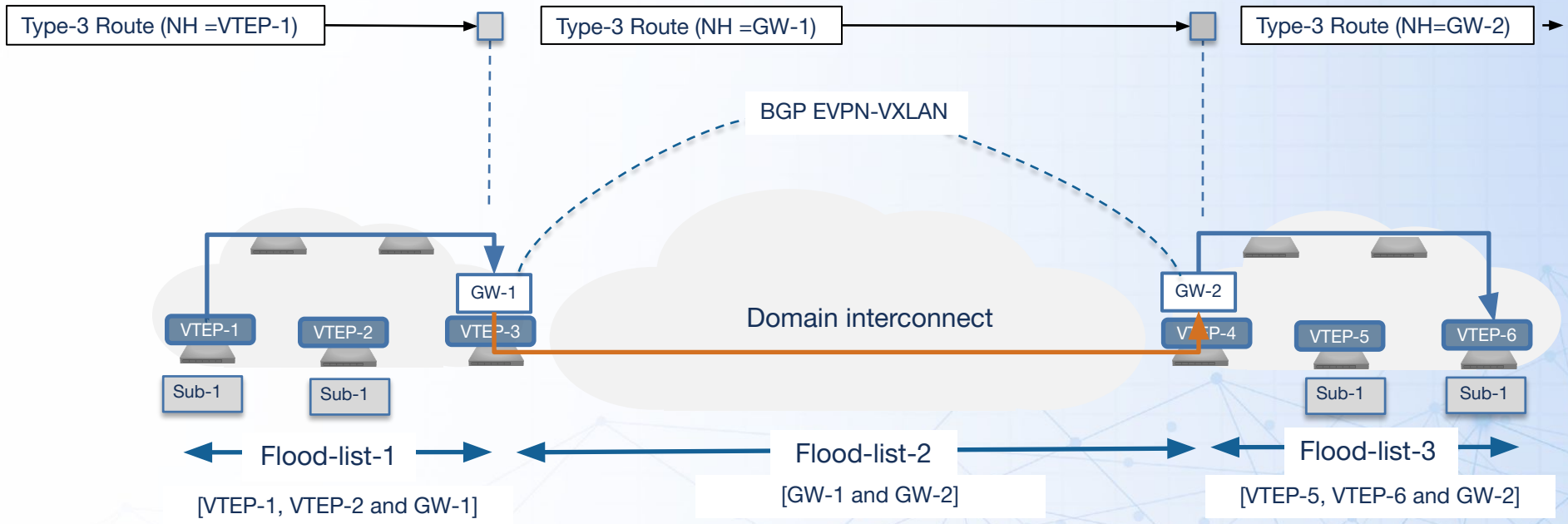


# EVPN L2 GW – RFC 9014

## EVPN L2 GW BUM traffic handling

- GW advertises it's own type-3 IMET route, resulting in flood-sets for each stretched bridge-domain
- Local flood-list containing the local VTEPs in the stretched bridge-domain
- Remote flood-list containing only the GW nodes in the stretched bridge-domain
- Hierarchical scaling, flood-list of the local VTEPs contain only the local VTEPs and the GW

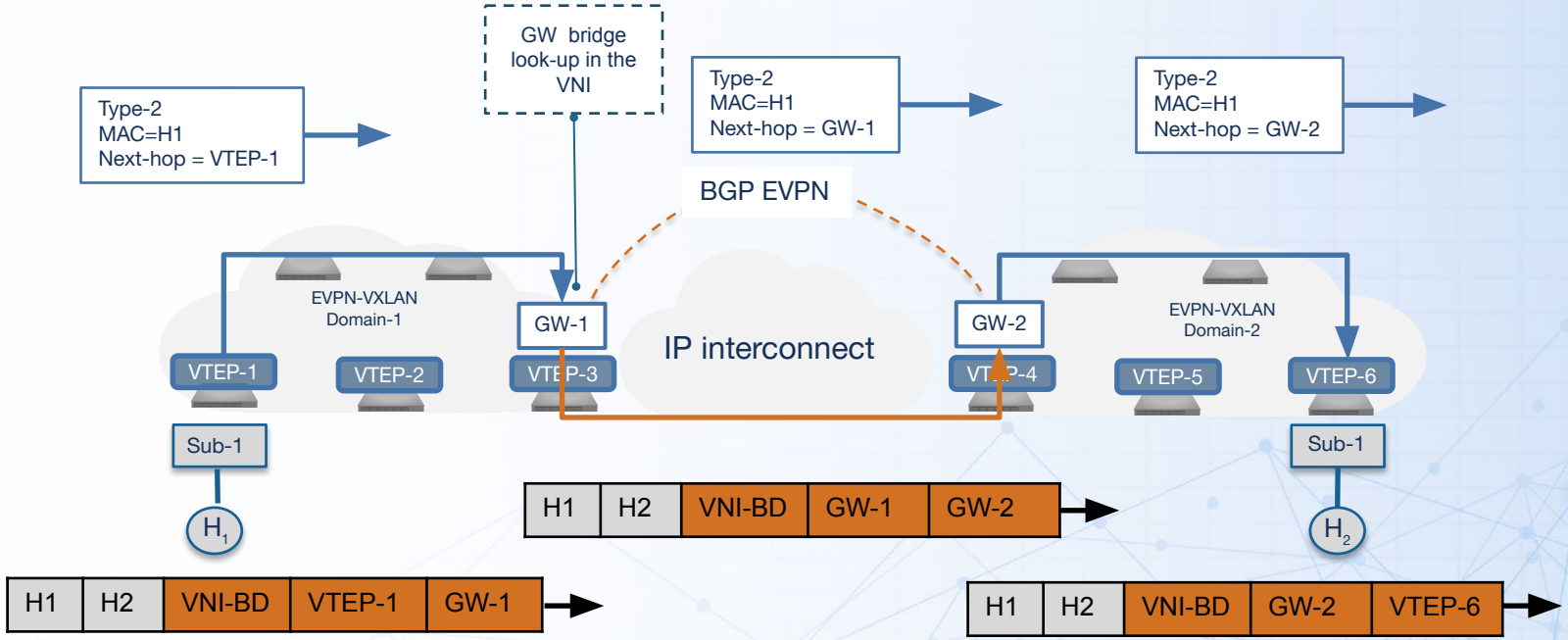


# EVPN L2 GW – RFC 9014

## EVPN L2 GW forwarding behavior

- Traffic bridged to the GW, as remote host (Sub-2) learnt with a next-hop of the GW
- GW performs VXLAN decap, with a MAC lookup in the BD of associated VNI
- GW performs VXLAN encap, as MAC lookup resolved to a next-hop of GW-2
- Remote GW performs a similar VXLAN decap/encap action before bridging to the local VTEP



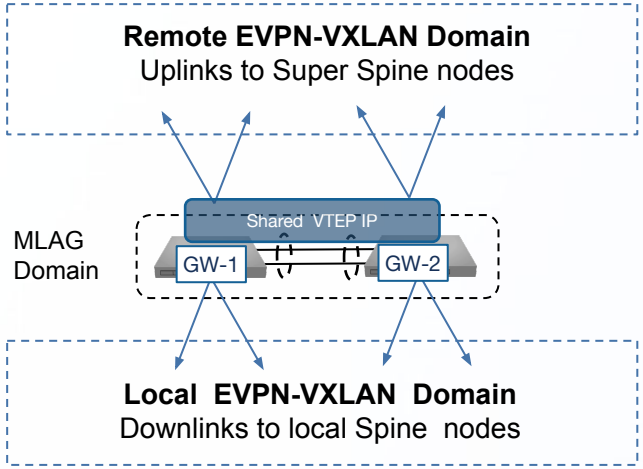

 

modus one ARISTA

# EVPN L2 GW – Resiliency models

## EVPN-GW with MLAG

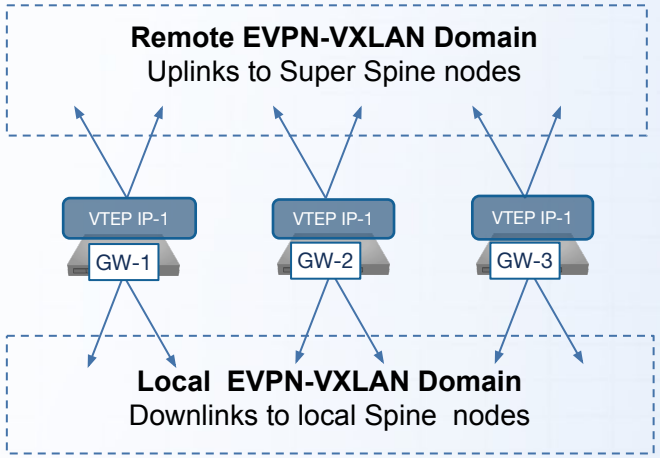
(VXLAN to VXLAN)




- Logical VTEP IP next-hop of EVPN routes
- ECMP load-balancing in the underlay
- Peer-link required between the nodes
- Support for directly attached hosts
- Limit single MLAG domain per-site

## EVPN-GW with Anycast IP

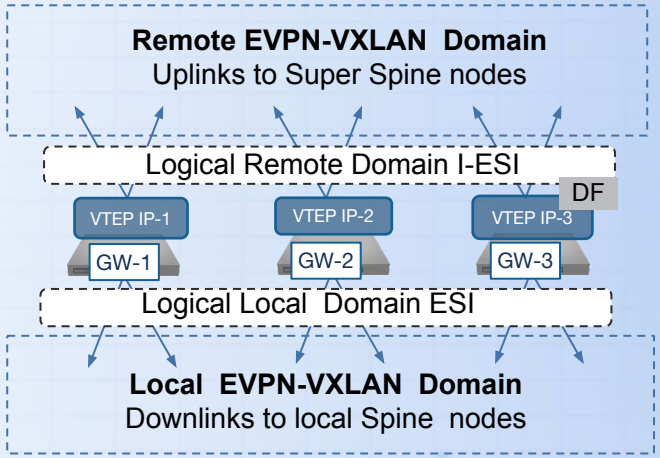
(VXLAN to VXLAN)




- Anycast IP next-hop of EVPN routes
- ECMP load-balancing in the underlay
- No peer-link required between the nodes
- No support for directly attached hosts
- Support for 16 GW nodes per-site

## EVPN-GW with All-Active

(VXLAN to VXLAN)




- GW nodes configured in an I-ESI
- ECMP load-balancing in the overlay
- No peer-link required between the nodes
- Support for directly attached hosts
- Support for 16 GW nodes per-site

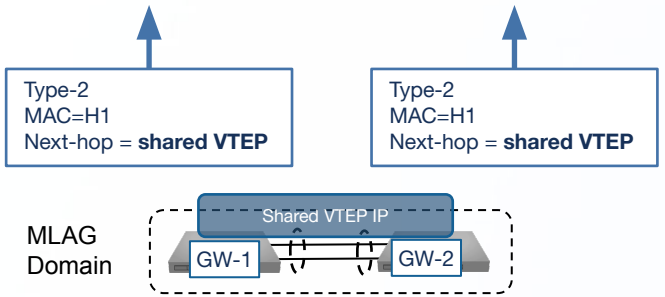
 

modus one ARISTA

# EVPN L2 GW – Resiliency models, loop prevention

## EVPN-GW with MLAG

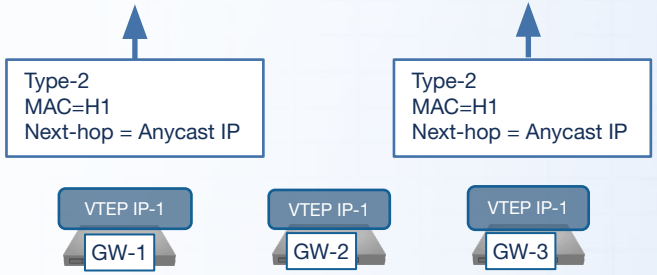
**Loop prevention**




- GWs advertises Type-2 with NH of shared VTEP IP
- Both GWs own the shared VTEP IP
- Route received from MLAG peer thus marked **invalid**
- Thus route is not re-advertised back into the domain.
- Only a single MLAG domain to avoid L2 loops.

## EVPN-GW with Anycast IP

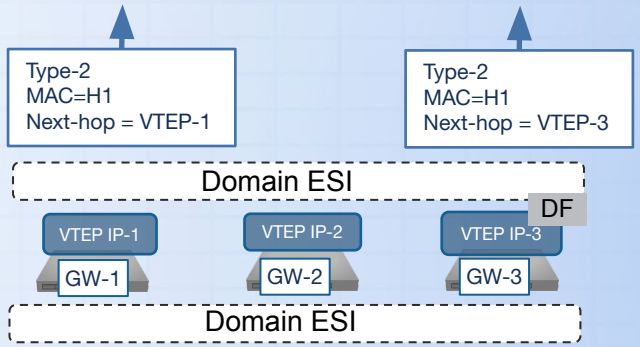
**Loop prevention**




- GWs advertises Type-2 with NH of Anycast IP
- All GWs own the Anycast VTEP IP
- Route received from a GW peer thus marked **invalid**
- Thus route is not re-advertised back into the domain

## EVPN-GW with All-Active

**Loop prevention**




- GWs advertises Type-2 with a unique NH
- Route received from a GW peer thus marked **valid**
- Route could be re-advertised back into the domain.
- Additional procedures required to avoid routing loops

modus one ARISTA

# EVPN L2 GW – Arista implementation detail

Arista EVPN layer 2 GW implementation

- Local domain is the default domain, implied on the peering without any additional configuration
- Remote domain needs to be explicitly defined in the peering with "*domain remote*"
- Configuration of the domains, automatically determine the scope of the type-1,3 and 4 routes

```
!
router bgp 64512
   router-id 1.1.0.7
   address-family evpn
      neighbor EVPN_Spine activate
      neighbor EVPN_RR activate
      neighbor EVPN_RR domain remote
       !
   vlan 10
      rd 3.3.0.1:1010

      rd evpn domain remote 3.3.0.1:64512

      route-target import 1010:1010
      route-target export 1010:1010

      route-target import evpn domain remote 64512:1010
      route-target export evpn domain remote 64512:1010
     !
```

EVPN peering in the **local** (default) domain for the EVPN GW

EVPN peering in the **Remote** domain for the EVPN GW

RD for the **local** (default) domain on the EVPN GW

RD for the **Remote** (default) domain on the EVPN GW

RT import and export policy for the **local** (default) domain on the EVPN GW

RT import and export policy for the **Remote** domain on the EVPN GW

modus one ARISTA

# EVPN Gateway Summary

# EVPN Routes and GW Behaviour

| EVPN Route | EVPN GW Behavior | Comments |
|---|---|---|
| **Type-1**<br>(AD per ES/AD per ESI) | Used for EVPN multihoming and are domain scoped, this means Type-1 route received within a domain are not re-advertised by the GW across domains. This reduces state churn across the EVPN domains during ESI failures | Domain level scope |
| **Type-2**<br>(MAC/MAC-IP) | The GW will re-advertise type-2 routes when the BD is stretched across domains. The the next-hop and the RD is changed to the GW's RD, RT change is optional, sequence number is retained. For MAC-IP routes the router-mac will be changed to the system MAC of the GW. | Advertised between domains when the MAC-VRF is stretched |
| **Type-3**<br>(IMET) | For BDs stretched between domains, the GW will originate its own Type-3 route for the bridge-domain; it will not re-advertise IMET routes between domains. | Domain level scope, the GW originates its own IMET route. |
| **Type-4**<br>(Ethernet Segment) | Used for EVPN multihoming and are domain scoped, this means Type-4 route received within a domain are not re-advertised by the GW across domains. | Domain level scope |
| **Type-5**<br>(IP-prefix) | GW providing L3 connectivity between domains, Type-5 routes are advertised between domains. The next-hop is changed to the GWs IP, the router-mac is changed to the system-mac of the GW, the (RD) and RTs are **not** changed. | Advertised between domains when L3 GW enabled |

ARISTA

# Thank You

www.arista.com

modus one

smart networks