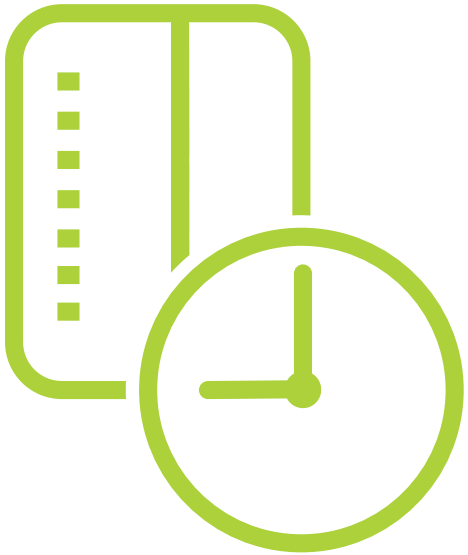
VXLAN/EVPN in a Nuttshell

Ralf Korschner, Systems Engineer EMEA
ralf@arista.com

ARISTA

Agenda

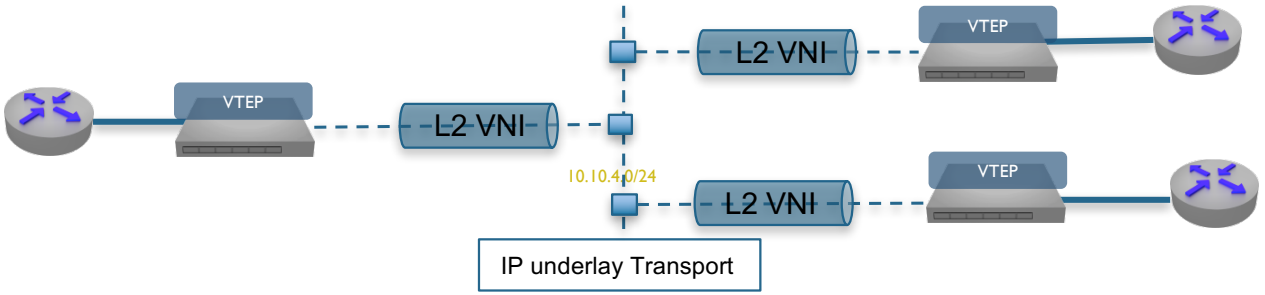
- VXLAN Introduction
- VXLAN for IXPs
- Control Plane Options
  - Head-end Replication
  - EVPN



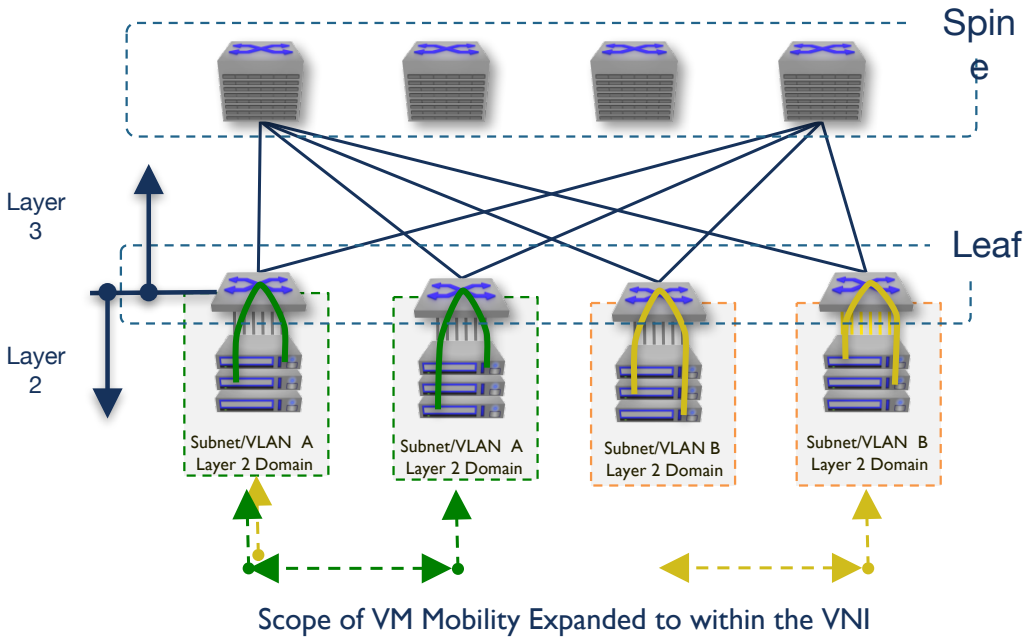
VXLAN Basics

Introducing VXLAN

- Layer 2 “Overlay Networks” on top of a Layer 3 network
  - “MAC in IP” Encapsulation
  - Layer 2 multi-point tunneling over IP UDP
  - Transparent to the physical IP underlay network
  - Provides Layer 2 scale across the Layer 3 IP fabric



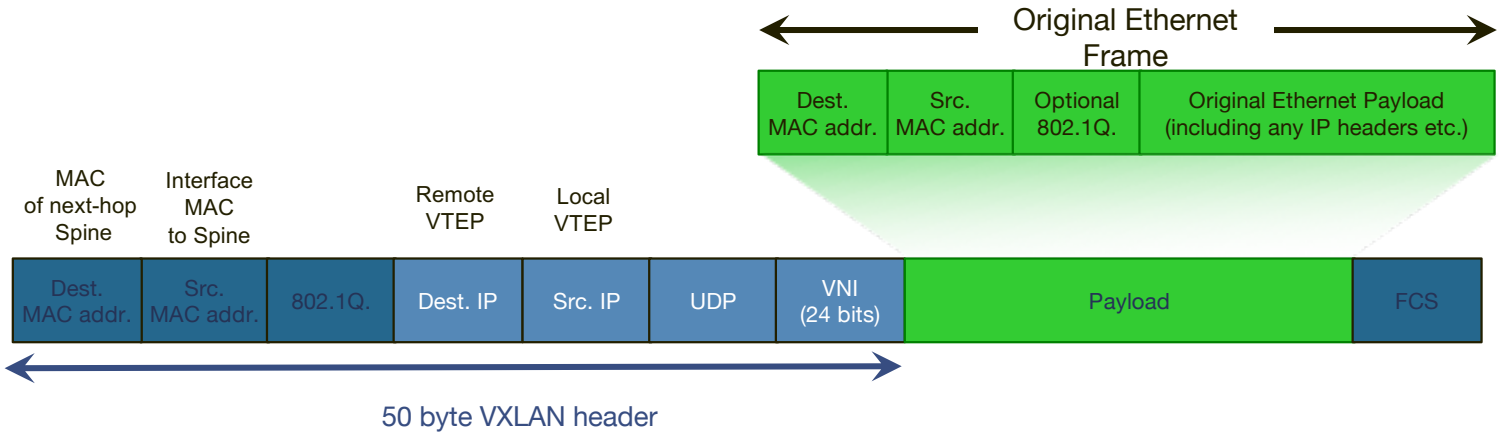
Data Center – Layer 3 Overlay Architectures



- Virtual eXtensible LAN (VXLAN)
  - IETF framework proposal, co-authored by:
    - Arista
    - VMware
    - Cisco
    - Citrix
    - Red Hat
    - Broadcom
- Enables Layer 2 interconnection across Layer 3 boundaries
  - Transparent to the physical IP network
  - Provides Layer 2 scale across the Layer 3 IP fabric
  - Abstracts the Virtual connectivity from the physical IP infrastructure
  - Enables Vmotion, etc. across IP fabrics

VXLAN Encapsulated Frame Format

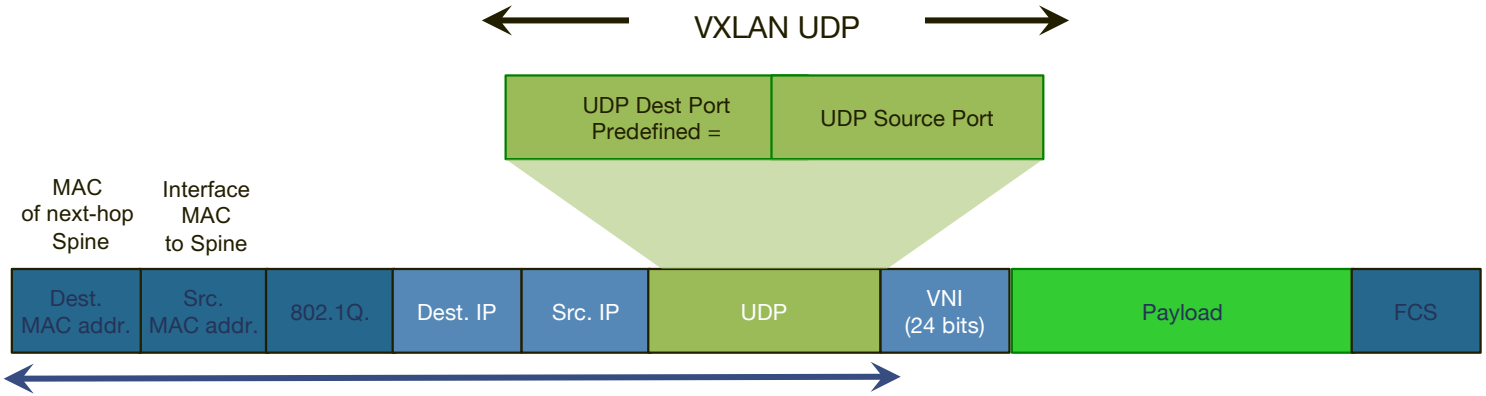
- Ethernet header uses local VTEP MAC and default router MAC (14 bytes plus 4 optional 802.1Q header)
- The VXLAN encapsulation source/destination IP addresses are those of local/remote VTI (20 bytes)
- UDP header, with SRC port hash of the inner Ethernets header, destination port IANA defined (8 bytes)
  - Allows for ECMP load-balancing across the network core which is VXLAN unaware.
- 24-bit VNI to scale up to 16 million for the Layer 2 domain/ vWires (8 bytes)



ARISTA

VXLAN Encapsulated Frame Format

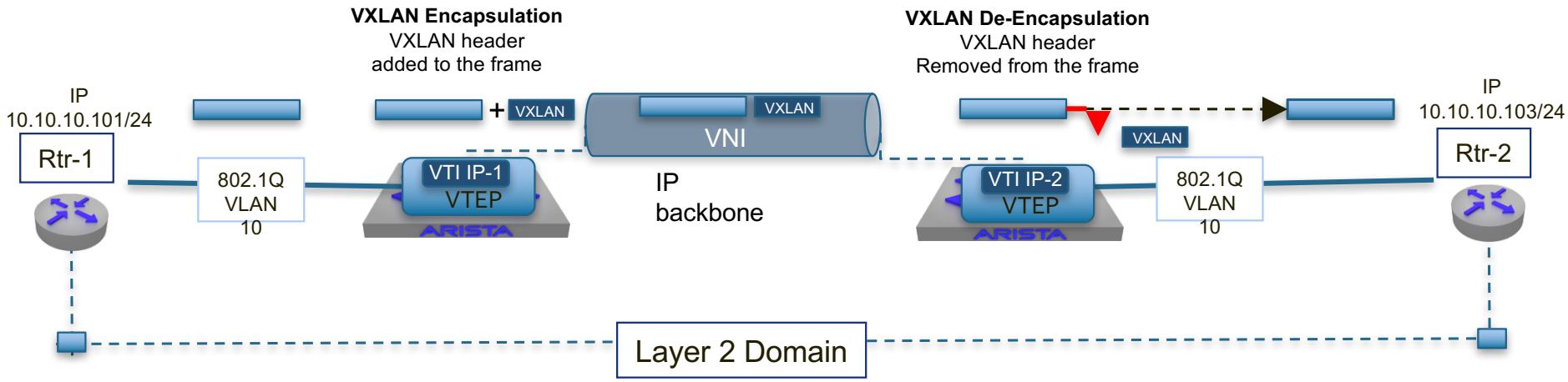
- To provide Entropy across a multi-path ECMP underlay network
  - UDP source port created from a Hash of the inner frame
  - What fields are hashed from the inner is not defined in the standard
  - Silicon vendor, will define the level of Entropy that can be achieved
  - UDP destination port, predefined in the standard as 4789



Source Port: It is recommended that the UDP source port number be calculated using a hash of fields from the inner packet - one example being a hash of the inner Ethernet frame's headers. When calculating the UDP source port number in this manner, it is RECOMMEND that the value be in the dynamic/private port range 49152-65535 [RFC6335].

VXLAN Terminology

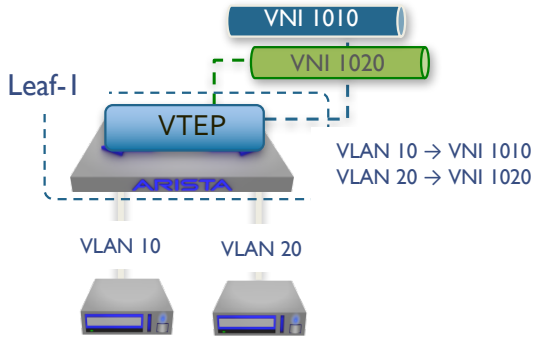
- **Virtual Tunnel End-point (VTEP).**
  - Entry point for connecting nodes into the VXLAN overlay network.
  - Responsible for the encap/decap with the appropriate VXLAN header.
- **Virtual Tunnel Identifier (VTI)**
  - An IP interface used as the Source IP address for the encapsulated VXLAN traffic
  - IP address residing in the underlay network
- **Virtual Network Identifier (VNI)**
  - A 24-bit field added within the VXLAN header.
  - Identifies the Layer 2 segment of the encapsulated Ethernet frame



VXLAN Terminology - VLAN service interfaces

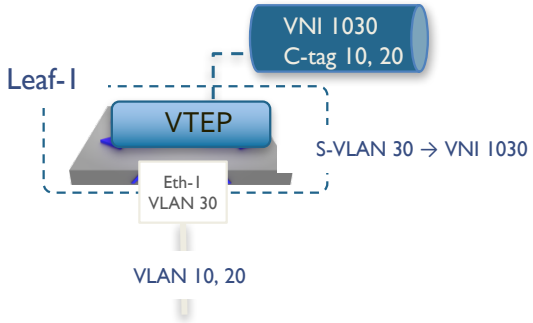
VLAN to VNI mapping

- One to One mapping between VLAN ID and the VNI
- Mapping is only locally significant,
- VLAN ID not carried on VXLAN encap frame
- Allows VLAN translation between remote VTEPs



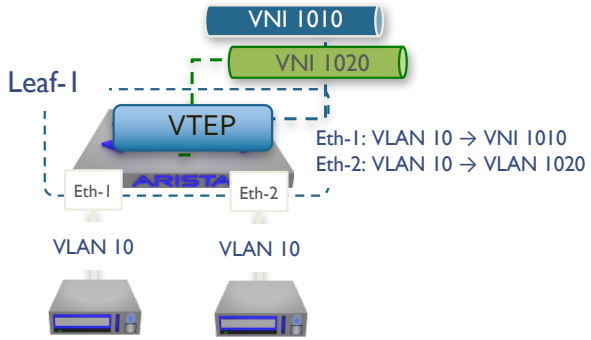
S-VLAN to VNI mapping

- Mapping of the outer S-Tag to a single VNI
- Inner C-Tags are transported within a single VNI
- The inner VLAN ID are carried on VXLAN encap frame
- Ability to transport all customer VLANs across a single VXLAN point to point link

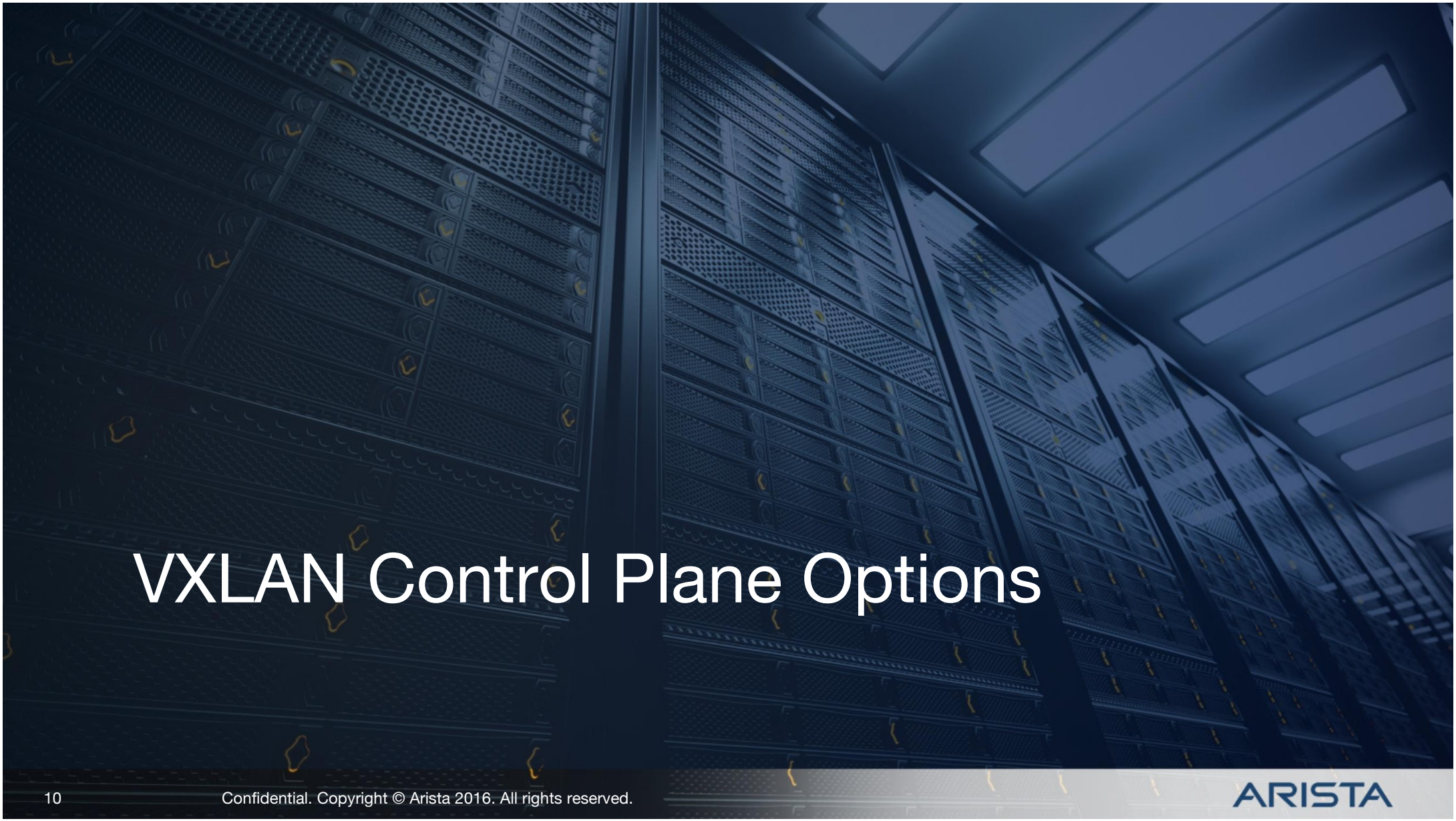


Port + VLAN to VNI mapping

- Mapping traffic to a VNI based on a combination of the ingress port and it VLAN-ID
- The VLAN ID is not carried on VXLAN encap frame
- Provides support for overlapping VLANs within a single VTEP to be mapped to different VNIs



ARISTA



VXLAN Control Plane Options

VXLAN Control Plane Options

- **The VXLAN control plane is used for MAC learning and packet flooding**
  - Learning what remote VTEP a host resides behind
  - Allowing the mapping of remote MACs to their associated remote VTEP
  - Mechanism for forwarding of the Broadcast and multicast traffic within the Layer 2 segment (VNI)

Controller Model

- State learning driven by third-party controller
- OVSDB or OpenStack ML2 plugin for orchestration
- Data Center virtualization and Orchestration focus

IP Multicast Control Plane

- VTEP join an associated IP multicast group(s) for the VNI(s)
- Unknown unicasts forwarded to VTEPs in the VNIs via IP multicast
- Flood and learn and requires IP multicast support in the underlay
- Limited deployments

Head-End Replication (HER)

- BUM traffic replicated to each remote VTEPs in the VNIs
- Unicast Replication carried out on the ingress VTEP
- MAC learning still via flood and learn, but no requirement for IP multicast

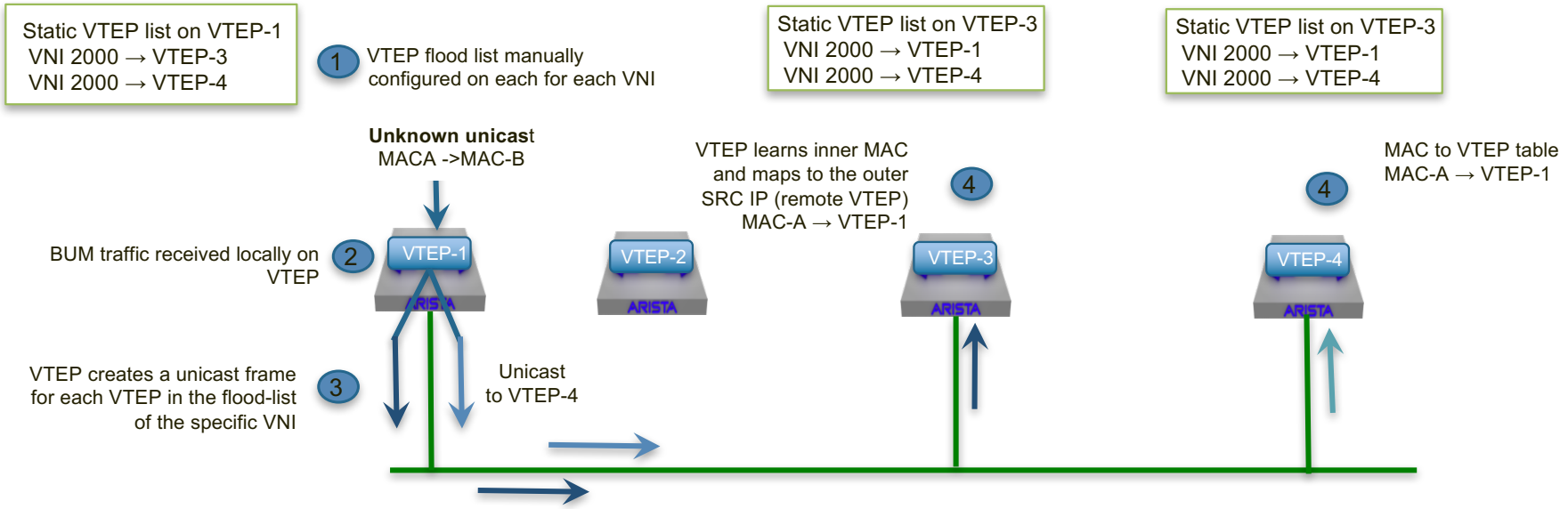
EVPN Model

- BGP used to distribute local MAC to IP bindings between VTEPs
- Broadcast traffic handled via IP multicast or HER models
- Dynamic MAC distribution and VNI learning, configuration can be BGP intensive

ARISTA

VXLAN Control plane – HER

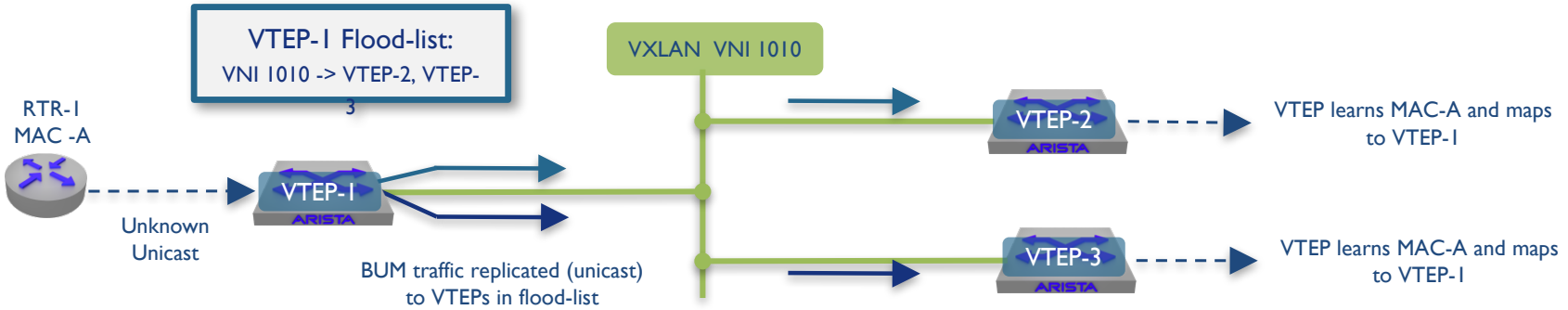
- Head-end Replication operation
  - Each VTEP is configured with an IP address "flood list" of the remote VTEPs within the VNI
  - Any Broadcast/Multicast or Unknown traffic is then replicated to the configured VTEPs in the list
  - Remote VTEPs receiving the flooded traffic learn inner source MAC from the received frame
  - VTEP's creating a remote MAC to outer SRC IP (VTEP) mapping for the entry



ARISTA

VXLAN Control Plane - HER

- Head-end Replication operation
  - Each VTEP is configured with an IP address "flood list" of the remote VTEPs within the VNI
  - Any Broadcast/Multicast or Unknown traffic is then replicated to the configured VTEPs in the list
  - Remote VTEPs receiving the flooded traffic learn inner source MAC from the received frame
  - Creating a remote MAC to outer SRC IP (VTEP) mapping for the entry



Flood list requires provisioning, MAC learning via flood and learn

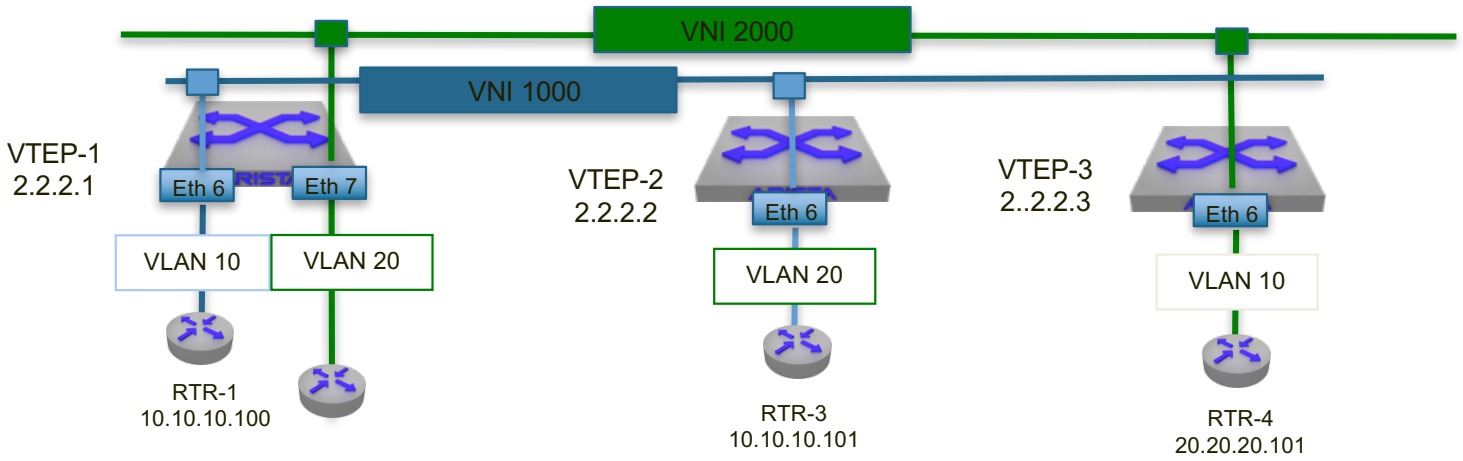
ARISTA

VXLAN Control Plane – HER, simple config

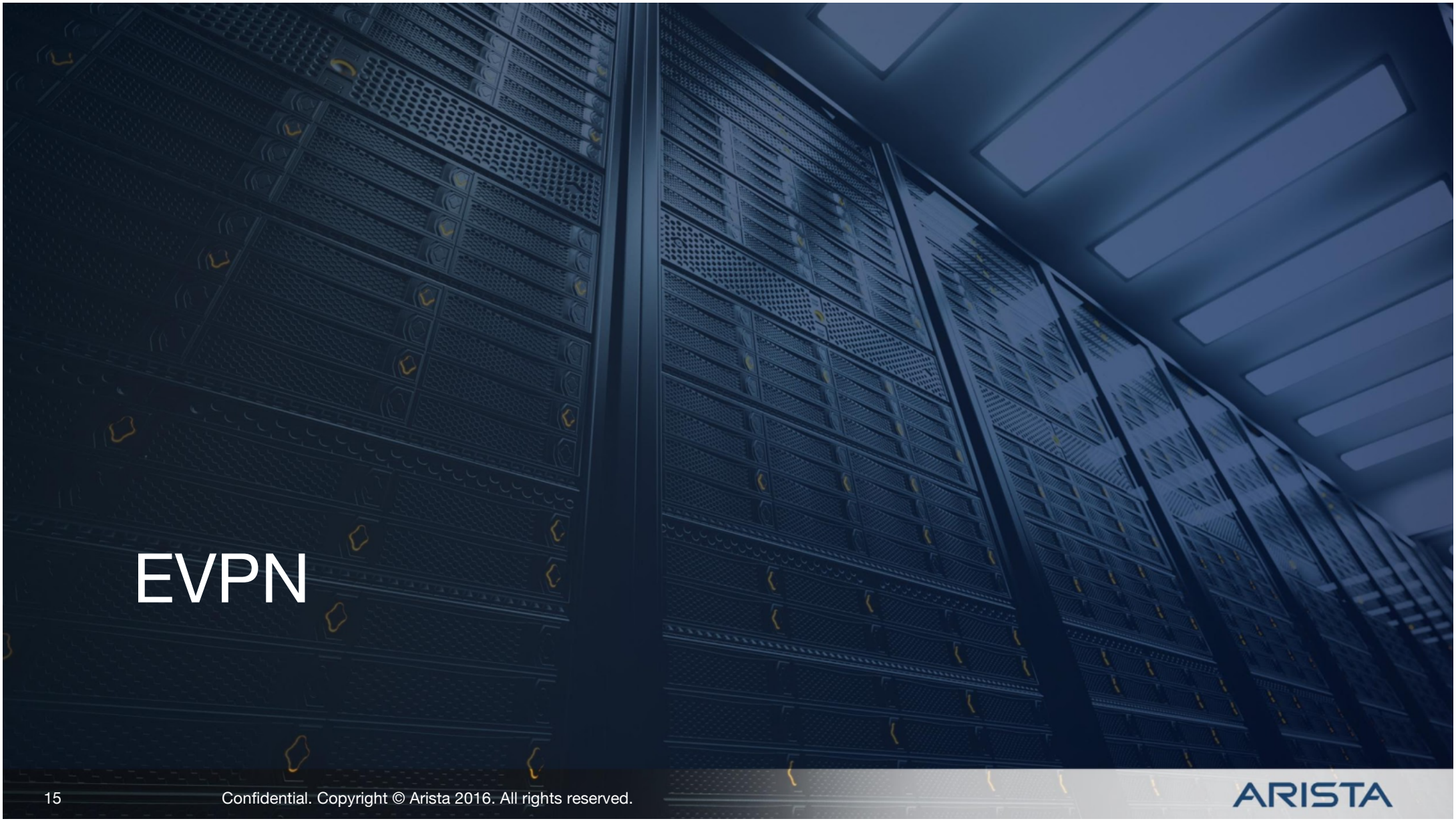
```
!
interface Loopback2
   ip address 2.2.2.1/32
!
Interface ethernet 6
   switchport mode access
   switchport access vlan 10
!
Interface ethernet 7
   switchport mode access
   switchport access vlan 10
!
interface Vxlan1
   vxlan source-interface Loopback2
   vxlan udp-port 4789
   vxlan vlan 10 vni 1000
   vxlan vlan 20 vni 2000
   vxlan vlan 10 flood vtep 2.2.2.3
   vxlan vlan 20 flood vtep 2.2.2.2
!
```

```
!
interface Loopback2
   ip address 2.2.2.2/32
!
Interface ethernet 6
   switchport mode access
   switchport access vlan 10
!
interface Vxlan1
   vxlan source-interface Loopback2
   vxlan udp-port 4789
   vxlan vlan 20 vni 2000
   vxlan vlan 20 flood vtep 2.2.2.1
!
```

```
!
interface Loopback2
   ip address 2.2.2.3/32
!
Interface ethernet 6
   switchport mode access
   switchport access vlan 20
!
interface Vxlan1
   vxlan source-interface Loopback2
   vxlan udp-port 4789
   vxlan vlan 10 vni 2000
   vxlan vlan 10 flood vtep 2.2.2.1
!
```



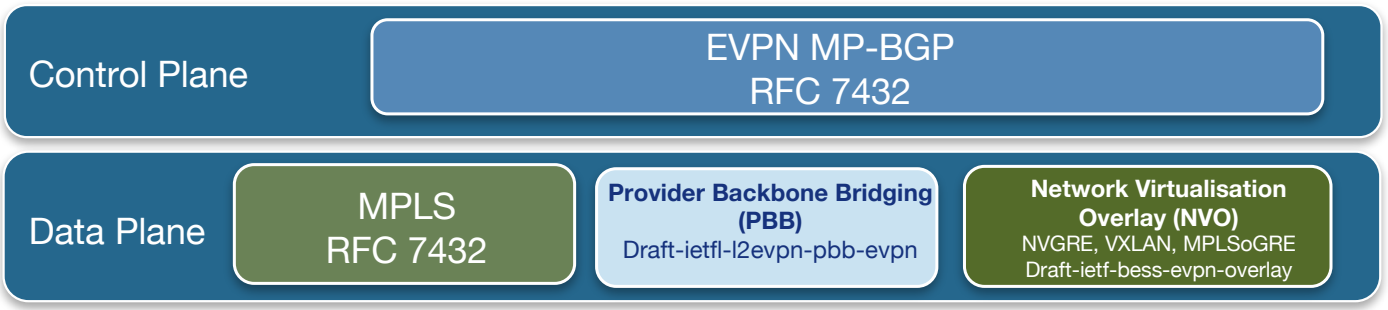
ARISTA



EVPN

What is Ethernet VPN (EVPN) - Standard body for EVPN

- EVPN Standard RFC 7432
  - Specifics an BGP EVPN control plane with a MPLS data plane
  - BGP control plane, new address family to advertise MAC/IP and IP prefixes.
  - Previously known as draft-ietf-l2vpn-evpn
  - Multi-vendor authors involving vendors and operators : ALU, Cisco, Juniper, AT&T, Bloomberg and Verizon
- Proposal for EVPN with Network Virtualisation Overlay (NVO)
  - Same EVPN control plane with a VXLAN Data plane (NGRE, MPLSoGRE)
  - Draft-ietf-bess-evpn-overlay



For the EVPN Data Plane, currently 1 standard (MPLS) and 2 proposals (NVO and PBB)

ARISTA

EVPN Protocol Summary

- **EVPN Protocol recap**
  - BGP control plane for the advertisement of MAC +IP binding and IP-prefixes
  - Support for multiple encapsulations VXLAN, NVGRE (NVO draft) and MPLS (RFC 7342)
  - New BGP address family , AFI = 25 (L2VPN) and SAFI =70 (EVPN)
  - IPVPN concepts for the multi-tenancy
    - Route Distinguishers to provide support for overlapping IP between tenant's
    - Route-Targets to allow the control of the export and import of route between VRFs

Route Type	Description
1	**Auto-Discover Segment route** - Used in EVPN's multi-homing deployments to allow the advertisement of Nodes sharing the same Ethernet Segment. Arista is supporting MLAG for multi-homing, support interpreting type-1 routes
2	**MAC address Route** - Advertisement of locally learnt/provisioned MAC address and optionally IP addresses. Can be advertised with a single label (asymmetric IRB) or dual label (symmetric IRB)
3	**Inclusive Multicast Ethernet Route** - Used to advertise EVI/VNI membership for the creation of ingress replication list.
4	**Ethernet Segment Route** – Used in multi-homing deployments to allow the dynamic discovery of shared Ethernet segments. . Arista is supporting MLAG for multi-homing no need to support this route
5	**IP prefix Route,** advertisement of a IP prefix and next-hop, no MAC address for the route is advertised.

Ethernet VPN

- EVPN, MP-BGP control-plane for delivering L2 and L3 VPN services with VXLAN
  - Evolution from the flood-learn mechanism of traditional L2 VPN (VPLS) service
  - Abstracts the (MP-BGP) control-plane from the (VXLAN/MPLS/PBB) forwarding plane
  - MP-BGP control plane to advertise host MAC and IP addresses and IP prefixes
  - Allows within a single MP-BGP control, L2 VPNs (hosts addresses) and L3 VPNs (IP prefixes).
- Potential use cases
  - Network virtualisation (overlay) services for stretching Layer 2 connectivity
  - Integration of Layer 2 and Layer 3 VPN services in the overlay
  - Data Center Interconnect (DCI)
  - Internet Exchange Points (IXPs)

What is Ethernet VPN (EVPN) -- Standard body for EVPN & EANTC Interop Testing

- Standards and Draft documents
  - RFC 7432 – BGP MPLS-Based Ethernet VPNs
    - https://tools.ietf.org/html/rfc7432
  - Network Virtualisation Overlay solutions using EVPN – VXLAN/NVGRE forwarding model
    - https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-04
  - Integrated Routing and Bridging within EVPN
    - https://www.ietf.org/archive/id/draft-sajassi-l2vpn-evpn-inter-subnet-forwarding-05.txt
  - IP prefix advertisement in EVPN
    - https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-02
- EANTC MPLS+ SDN+ NFV World Congress
  - http://www.eantc.de/fileadmin/eantc/downloads/events/2011-2015/MPLSSDNNFV_2017/EANTC-MPLSSDNNFV2017-WhitePaper-Final_v2.pdf

Deploying VXLAN with EVPN

EVPN Operation

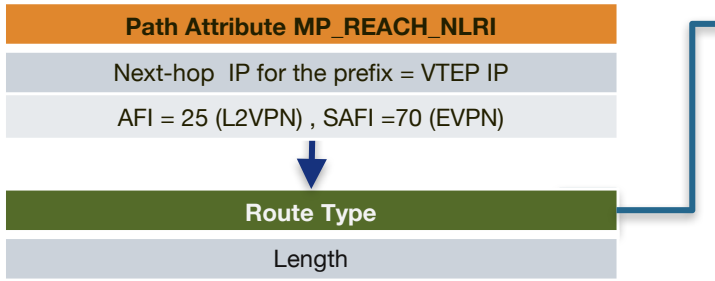
ARISTA

EVPN Operation

- EVPN is built on Multi Protocol BGP
  - Introduction of a new EVPN address family
    - Address Family Identifier 25 (Layer 2 VPN) subsequent AFI 70 (EVPN)
    - Advertisement of host MAC/IP binding and IP prefixes
    - Distribution of Layer 2/3 information allows support for integrated bridging and routing in VXLAN overlay networks.
  - Utilises Layer 3 VPN concepts of Route-distinguishers and Route Targets
    - Providing support for multi-tenant VXLAN overlays
    - Support for over-lapping IP address spaces between tenants
  - Multiple tenant's NLRI information carried within a single shared BGP session,
    - NOT BGP session per tenant

EVPN Operation – Route Types

- The new EVPN NLRI defines five route types
  - Not all route type are mandatory, specific support will be based on the vendors implementation
  - Next hop (VTEP IP address) for the route is contained in the MP_REACH_NLRI path attribute



Route Type	Description
1	**Auto-Discover Segment route** – Used to support EVPNs multi-homing deployment models
2	**MAC address Route** - Advertisement of locally learnt/provisioned MAC address and optionally IP addresses.
3	**Inclusive Multicast Ethernet Route** - used to advertise VTEPs VNI membership for the creation of ingress replication lists
4	**Ethernet Segment Route** – used in multi-homing deployments to allow the dynamic discovery of shared Ethernet segments
5	**IP prefix Route,** advertisement of a IP prefix and next-hop, no MAC address for the route is advertised.

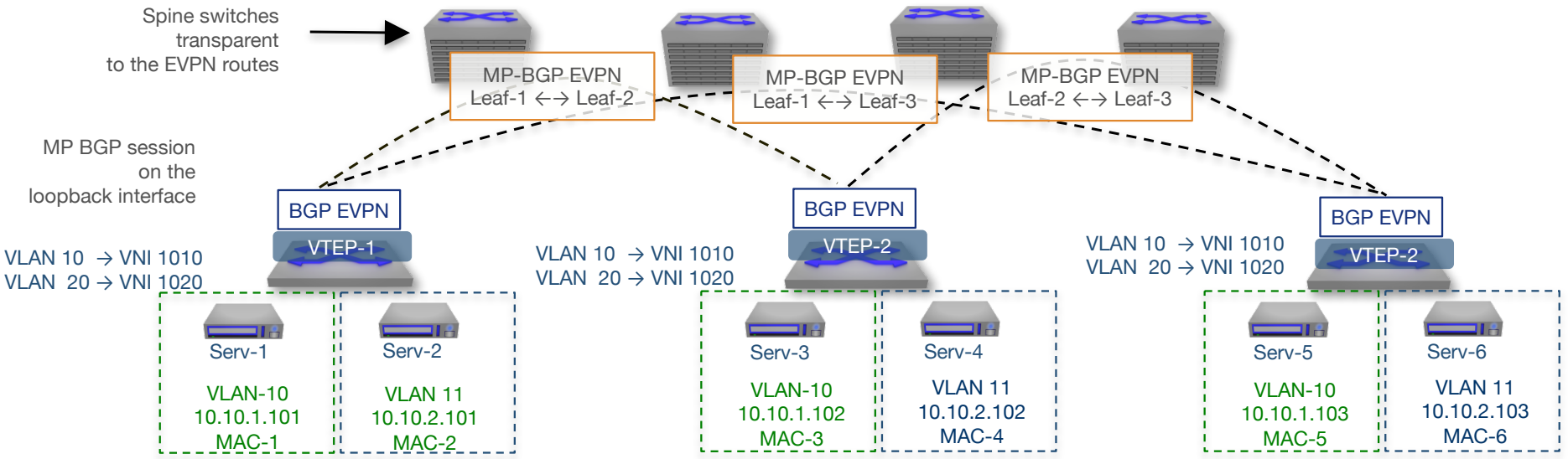
Deploying VXLAN with EVPN

Layer 2 VPN deployment model

ARISTA

Layer 2 EVPN deployment model – eBGP topology

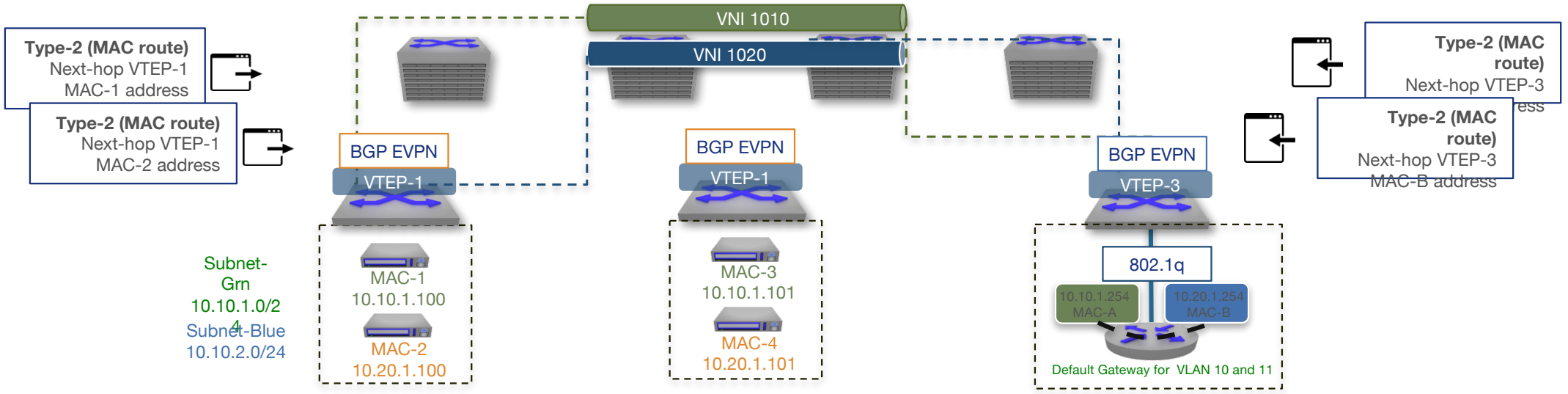
- Layer 2 EVPN model, multi-hop eBGP between leaf switches
  - Spine transparent to the EVPN sessions, unless acting as an EVPN/VTEP
  - Full-mesh multi-hop eBGP between Leafs switches sharing a VNI
  - Advertise EVPN type 2 and 3 routes via MP-BGP EVPN session



ARISTA

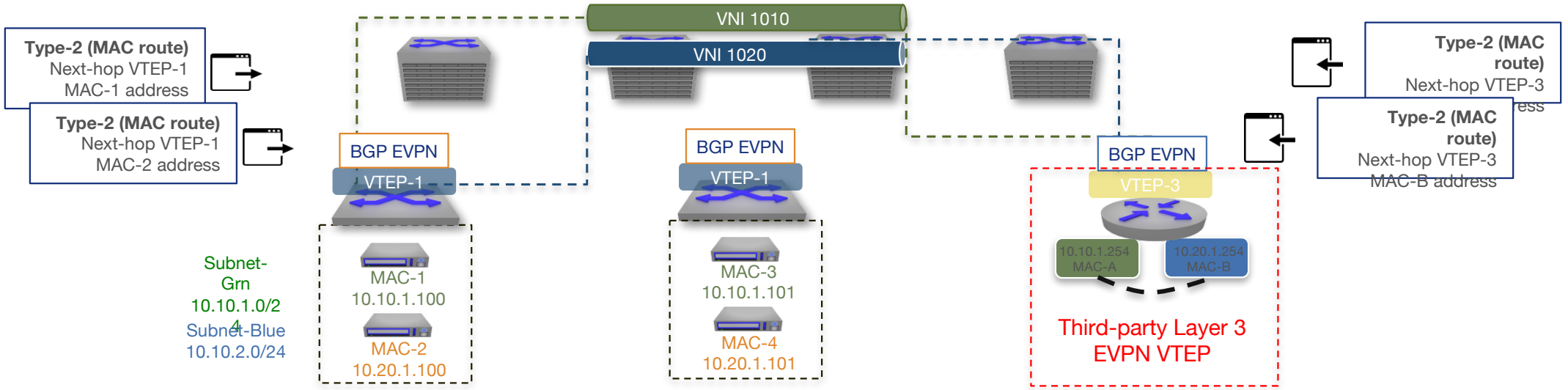
Layer 2 EVPN – Use case 1

- Layer 2 EVPN model – Layer 3 connectivity via external L3 node
  - Layer 2 EVPN Model only announcing MAC routes between VTEPs (type 2 mac route)
  - Providing layer 2 connectivity between leaf across the L3 infrastructure
  - No current support for VXLAN routing on the Arista VTEP nodes in this model
  - Inter-VLAN routing between the layer 2 domains via an external non EVPN aware node



Layer 2 EVPN – Use case 2

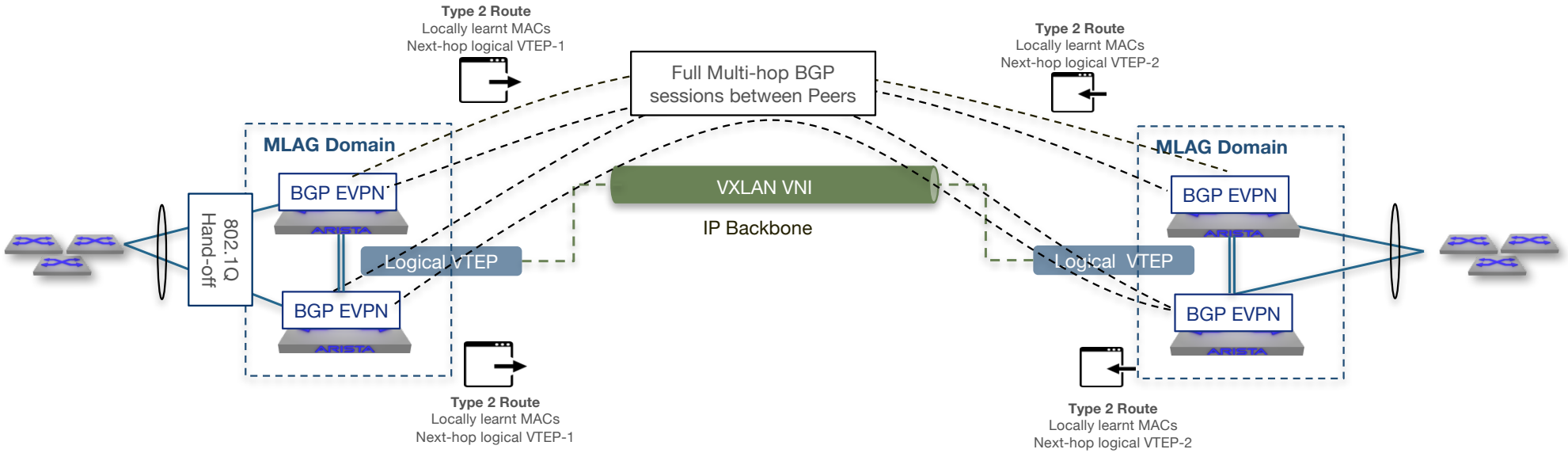
- Layer 2 EVPN model – Layer 3 connectivity via third-party Layer3 VTEP inter-op
  - Support for inter-operability with third-party Layer 3 EVPN VTEPs
  - VXLAN Routing achieved on the third-party Layer 3 VTEP, running EVPN
  - MAC addresses exchanged with the third-party via EVPN
  - Support for the third-party in a active-active and active-standby (reception of type 1 route)



ARISTA

Layer 2 EVPN – Use case 3

- Layer 2 VPN functionality can be used for Standalone DCI solution
  - MLAG Domain at each site for resiliency, with VLAN hand-off to the MLAG nodes
  - BGP control plane to advertise MAC address across the WAN
  - Multi-hop eBGP sessions with the DCI peers at the remote site
  - Advertisement of MAC addresses



ARISTA

ARISTA