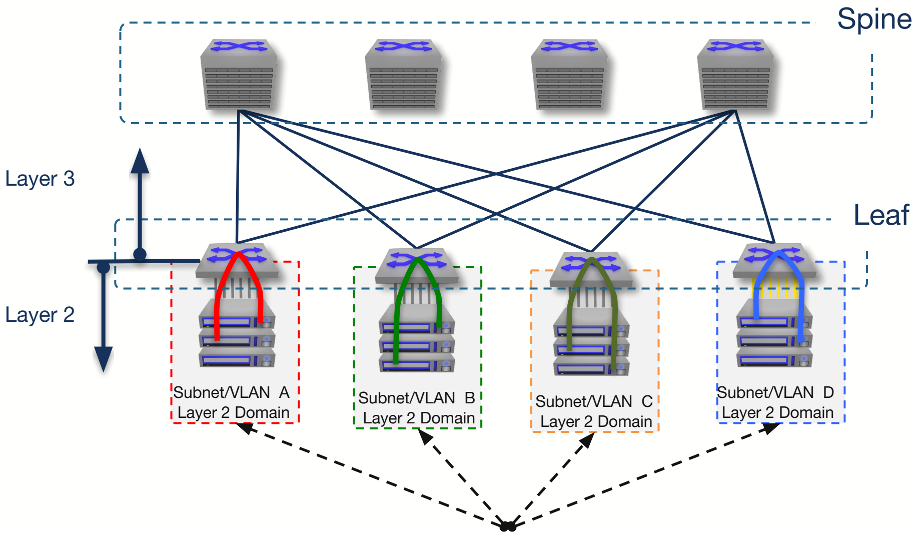
VXLAN 101

DENOG10 - November 21, 2018

Florian Hibler <florian@arista.com>

Data Center – Layer 3 Underlay Architectures



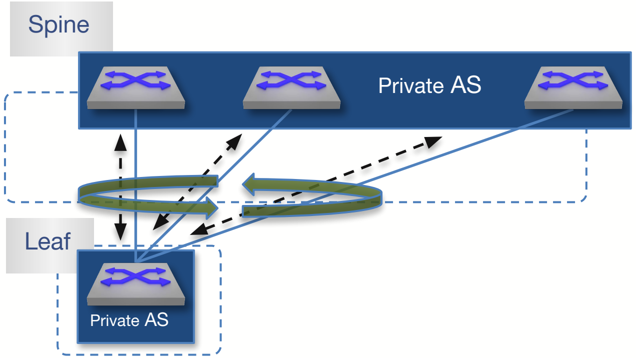
Scope of VM Mobility restricted to within the rack

- For scale and control evolution to Layer 3 architecture
  - Routed traffic at the top of the rack
  - Utilize proven and trusted standard Layer 3 protocols
  - Mature Open standards for interoperability
  - Minimize the size of the Layer 2 domain
  - Reducing the size of the fault & broadcast domains
  - Standard scalable model

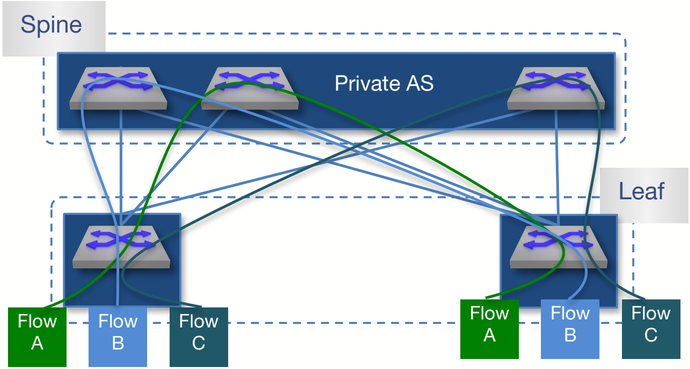
Utilize tried and proven protocols and management tools

DC IP Fabric – Equal Cost MultiPathing (ECMP)

- Each leaf node has multiple paths of equal “cost” to each Spine
  - ECMP used to load balance flows across the multiple spine node
  - For each prefix, routing table has next-hop (path) to each spine
  - Load-balancing algorithm is configurable based on L3/L4 info for granularity
  - Seed hash support to avoid polarization, but not required in a two tier design

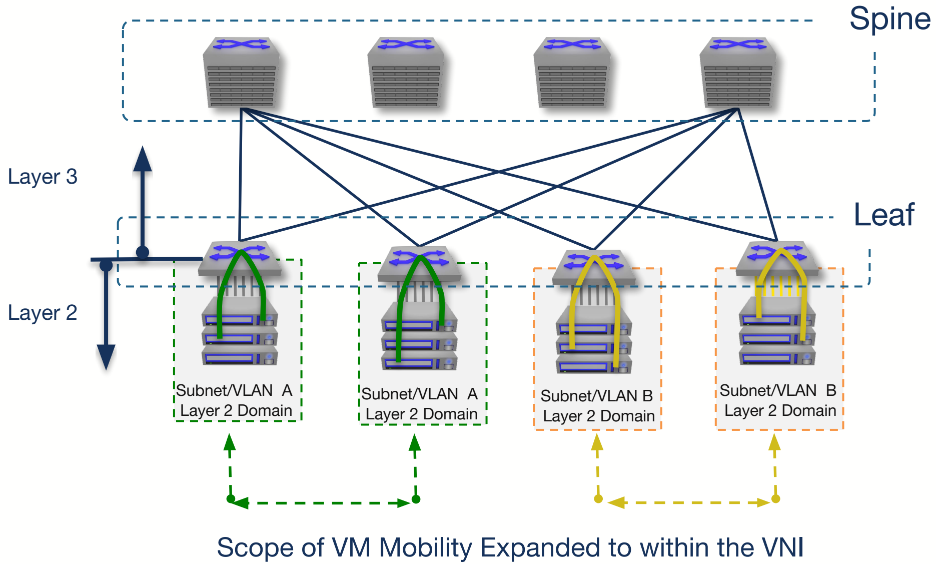


ECMP Load-balancing across all active paths



Flow based load-balancing across the active paths

Data Center – Layer 3 Overlay Architectures

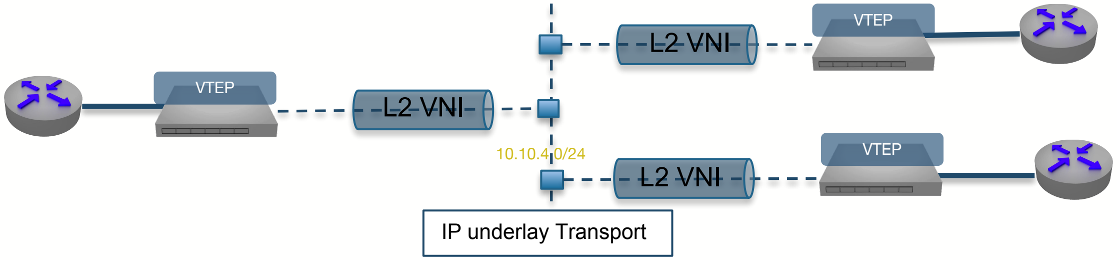


- Virtual eXtensible LAN (VXLAN)
  - Filed as RFC7348
  - Framework co-authored by Arista, Broadcom, Cisco, Citrix, Red Hat, VMware
  - Enables Layer 2 interconnection across Layer 3 boundaries
  - Transparent to the physical IP network
  - Provides Layer 2 scale across the Layer 3 IP fabric
  - Abstracts the Virtual connectivity from the physical IP infrastructure

VXLAN Basics

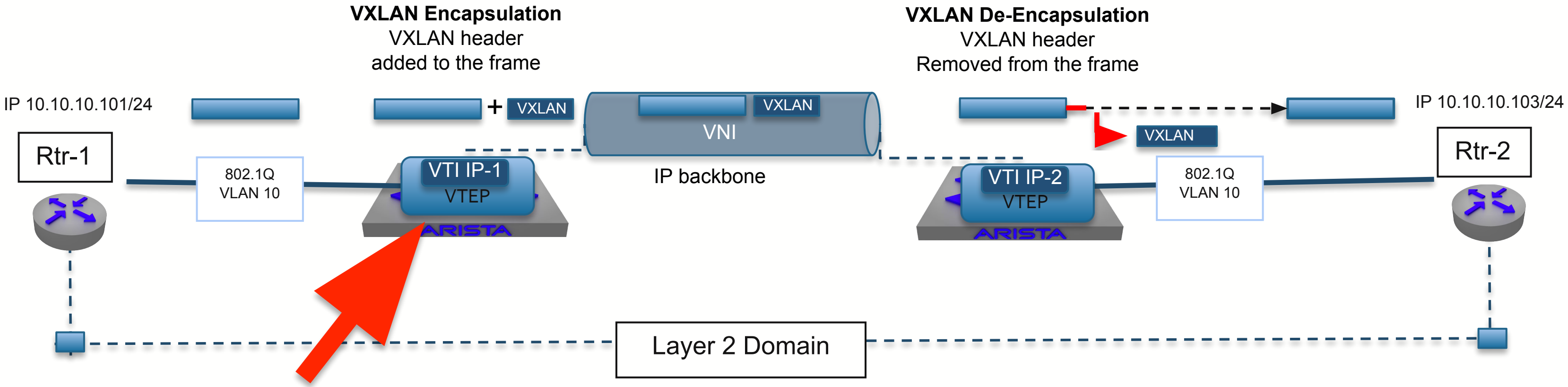
Introducing VXLAN

- 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



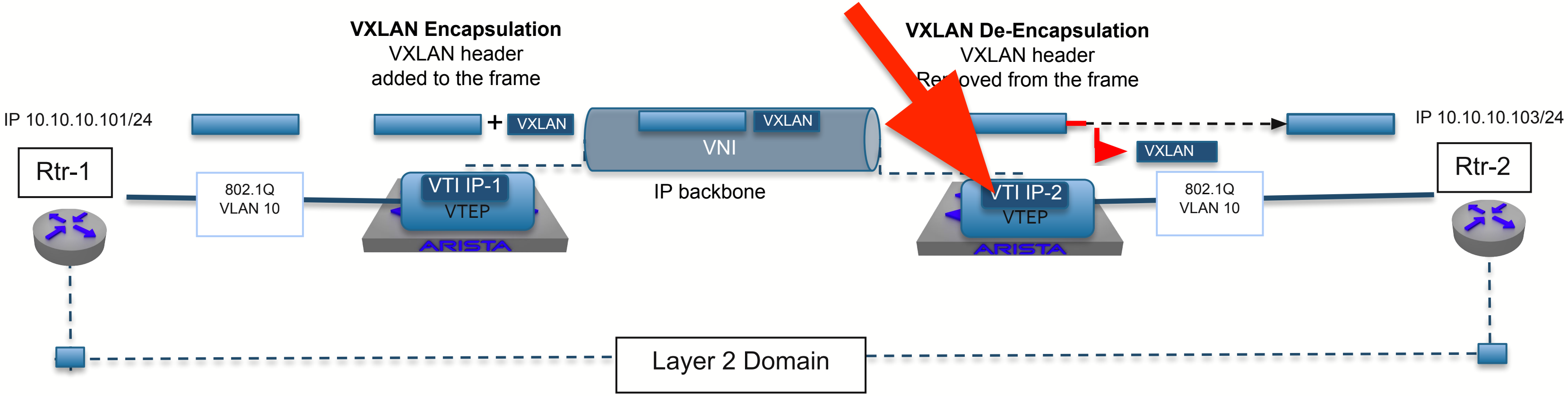
VXLAN Terminology

- **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



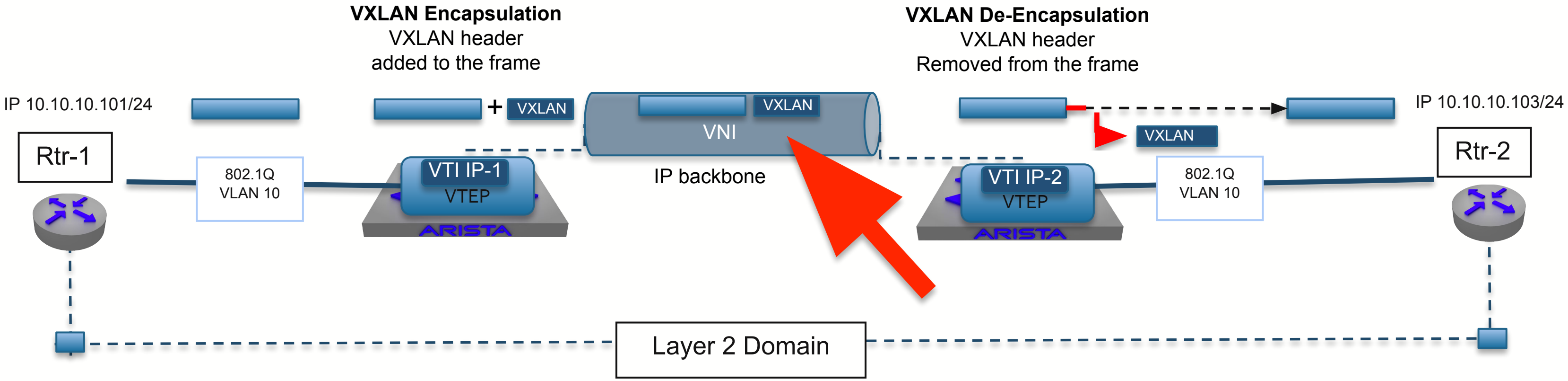
VXLAN Terminology

- **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



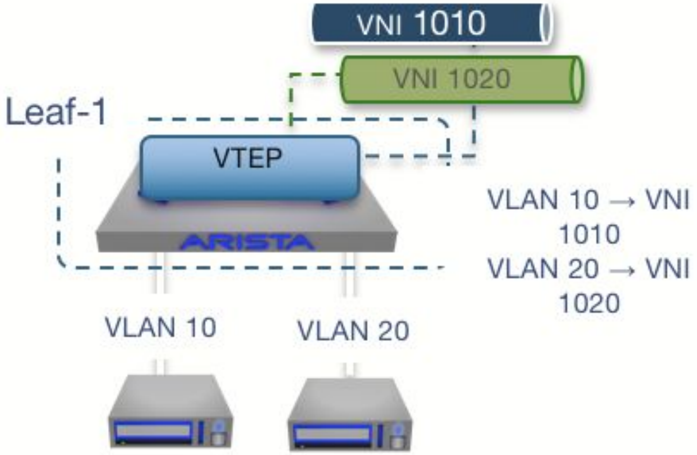
VXLAN Terminology

- **Virtual Network Identifier (VNI)**
  - A 24-bit field added within the VXLAN header
  - Identifies the Layer 2 segment of the encapsulated Ethernet frame

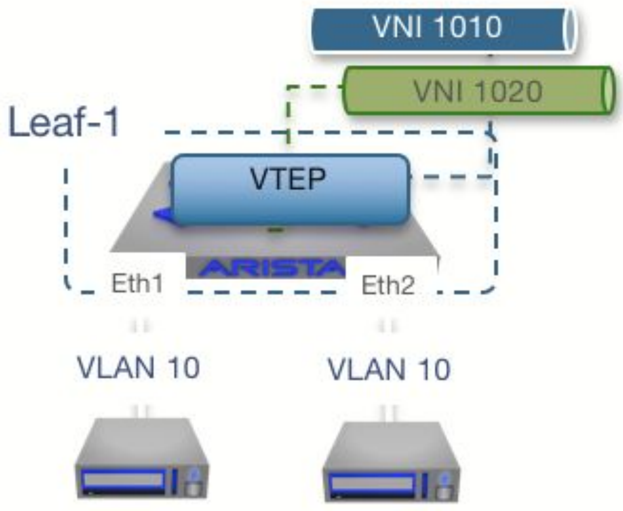


VXLAN Terminology - VLAN service interfaces

- **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



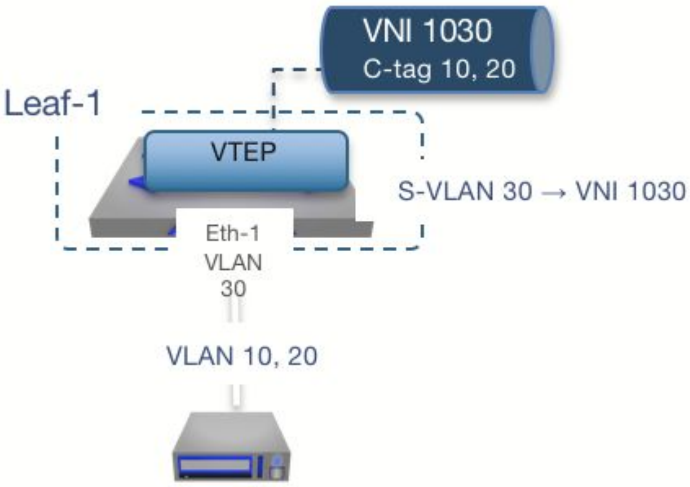
VXLAN Terminology - VLAN service interfaces



- **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

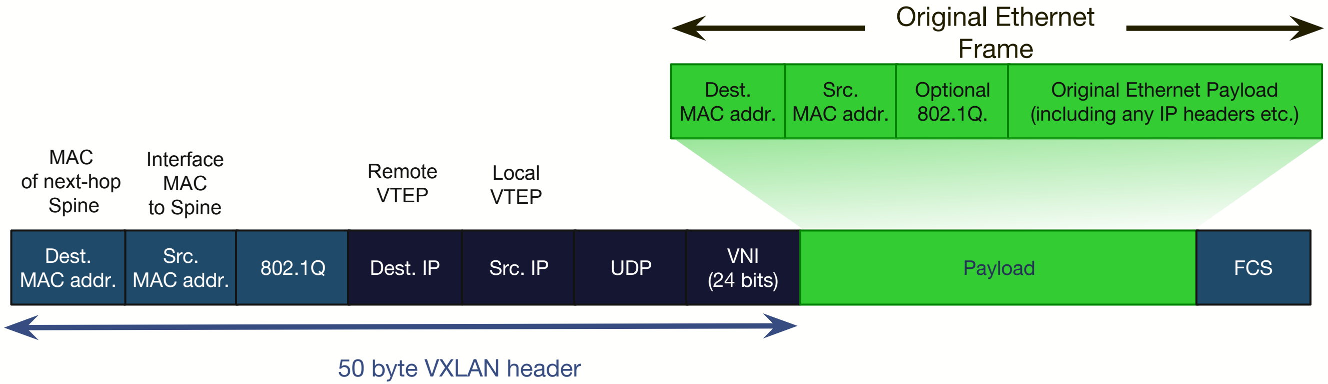
VXLAN Terminology - VLAN service interfaces

- **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



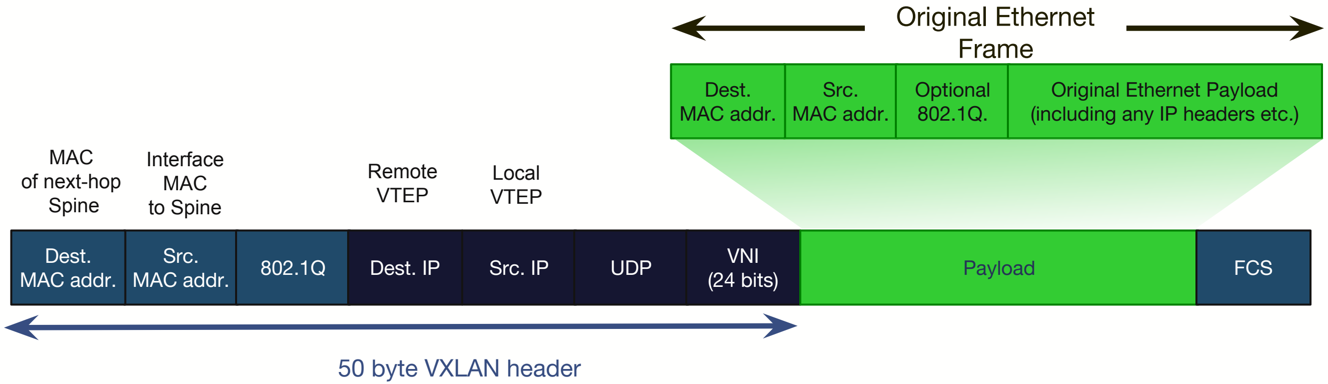
VXLAN Encapsulated Frame Format

- 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)



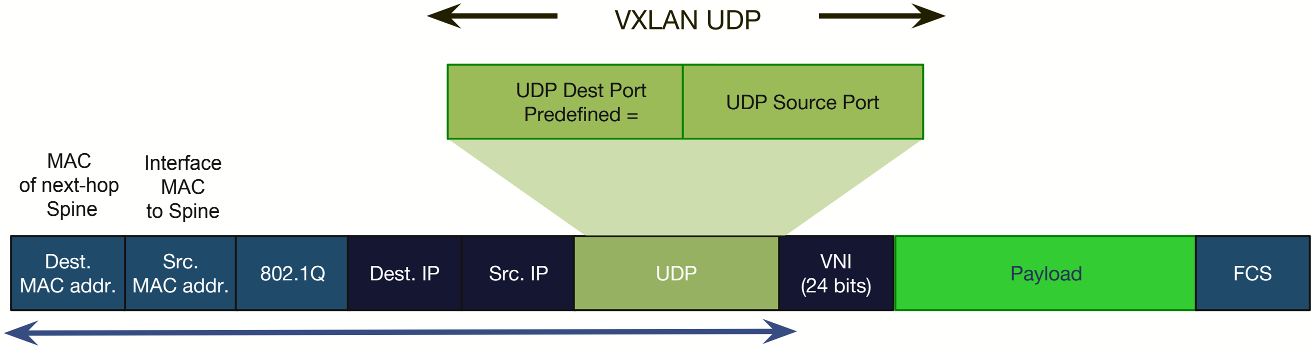
VXLAN Encapsulated Frame Format

- 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)



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 RFC as 4789



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

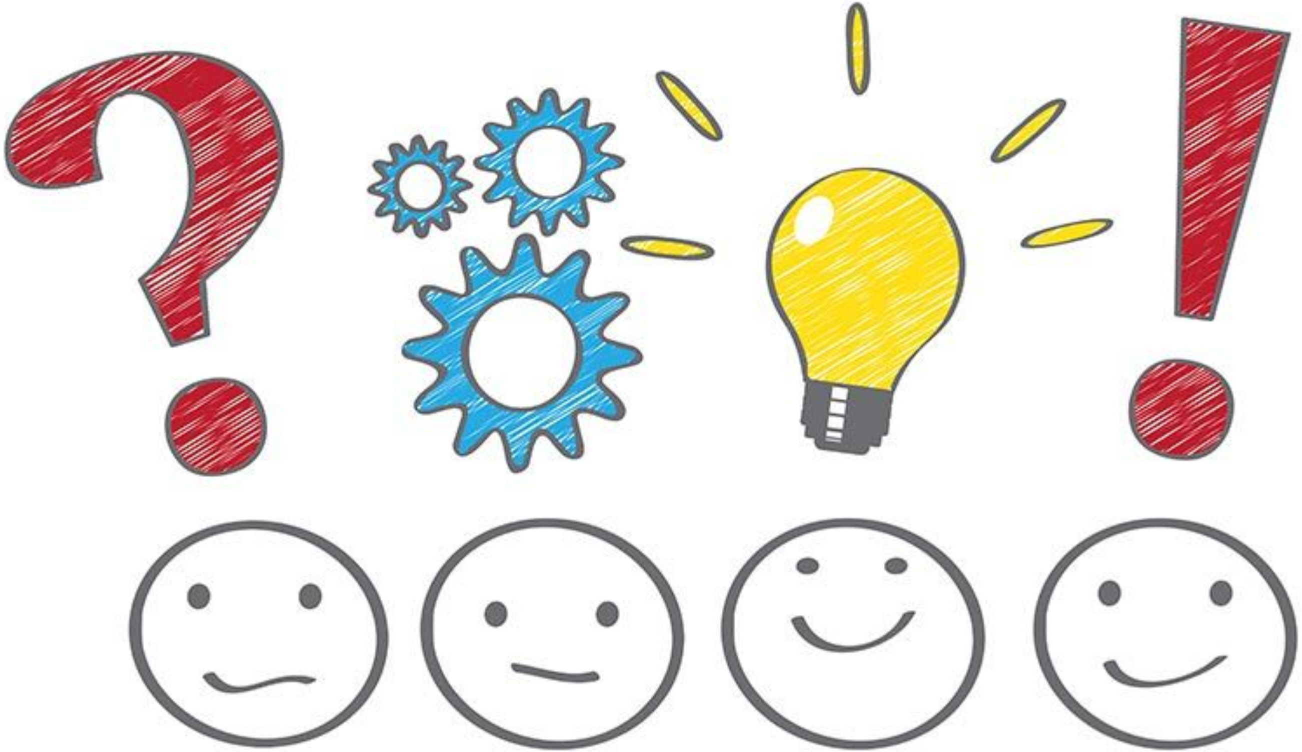
Head-End Replication (HER)

- 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

Questions?

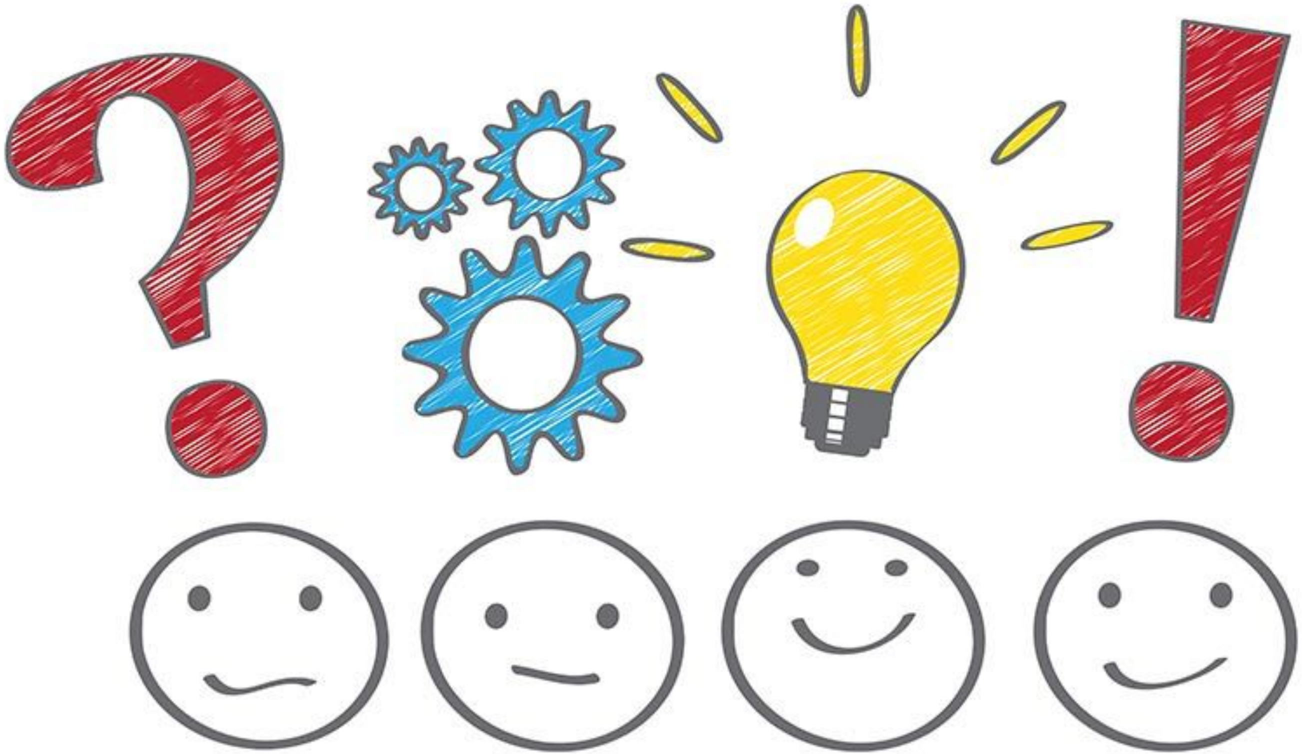


Thank You For Your Attention

Florian Hibler
Systems Engineer
Arista Networks, Inc.

(e) florian@arista.com
(m) +49 171 7576089
(w) http://www.arista.com

Questions?



Thank You For Your Attention

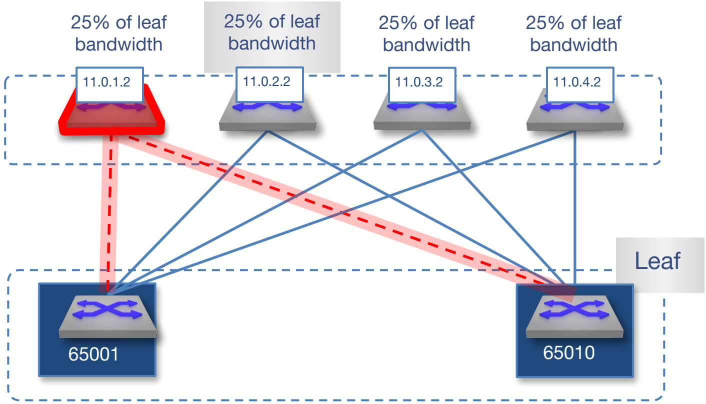
Florian Hibler
Systems Engineer
Arista Networks, Inc.

(e) florian@arista.com
(m) +49 171 7576089
(w) http://www.arista.com

Backup

DC IP Fabric – Equal Cost MultiPathing (ECMP)

- Resilient ECMP, minimize flow disruption during a failure
  - With 4-way ECMP, loss of a single node/link only reduces bandwidth by 25%
  - Resilient ECMP ensures only traffic traversing the failed path is re-distributed.
  - Flows on the remaining active paths are not re-distributed, thus unaffected by the outage



ip hardware fib ecmp capacity 4 redundancy 2

next-hop table

~~1- 11.0.1.2~~ -Fail
2- 11.0.2.2
~~3~~- 11.0.3.2
4- 11.0.4.2
~~5- 11.0.1.2~~ -Fail
6- 11.0.2.2
7- 11.0.3.2
8- 11.0.4.2

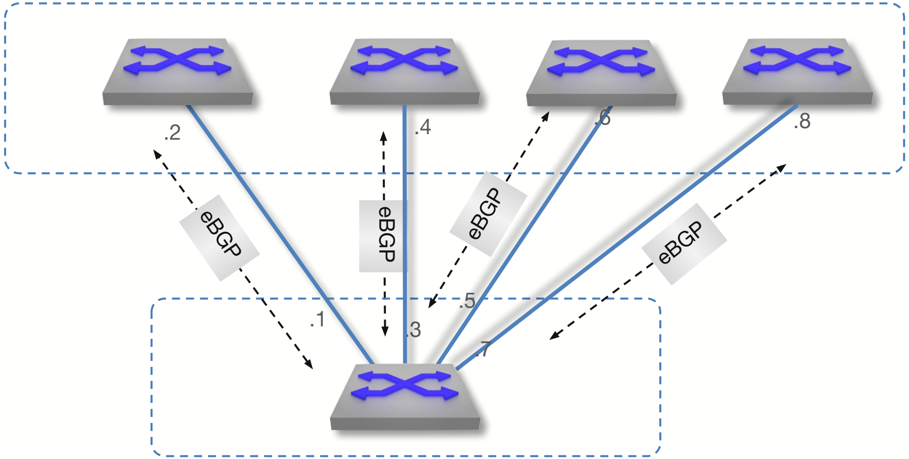
New next-hop table

1- 11.0.2.2 - NEW
2- 11.0.2.2 – no change
3- 11.0.3.2 – no change
4- 11.0.4.2 – no change
5- 11.0.3.2 - NEW
6- 11.0.2.2 – no change
7- 11.0.1.2 – no change
8- 11.0.2.2 – no change

Number of Next-hop (N) remains the same regardless of the number active next-hops

ECMP with four unique next-hop with 1+1 redundancy giving a total of 8 next-hops

DC IP Fabric – iBGP vs eBGP



eBGP session on the physical interface of the nodes using /31 addressing

- eBGP between Nodes
  - Route reflector design not required
  - Easy to determine the source of BGP paths in the RIB
  - Paths can be advertised from one eBGP peer to another

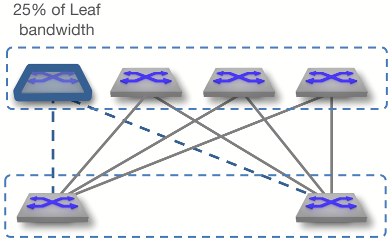
DC IP Fabric – Graceful Maintenance

Seamless automated Spine Upgrade with Open Standards

- Gracefully drain the traffic away from the switch, via BGP route maps and GSHUT communities
- Upgrade switch with no code dependency concern
- Reinsert switch into forwarding and gracefully start forwarding traffic

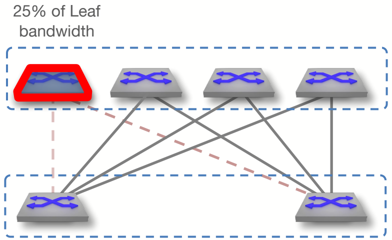
1 **Production**

- Neighbors (BGP & LLDP)
- Routes



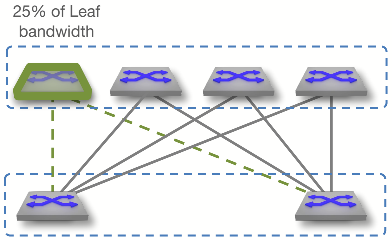
2 **Graceful Removal + upgrade**

- AS prepend, local-pref lowered, GSHUT set



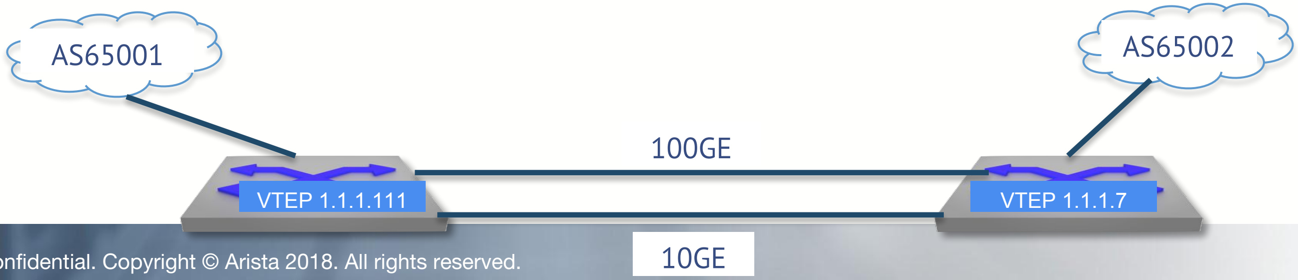
3 **Graceful Insertion**

- Re-establishing forwarding paths
- Remove maintenance mode



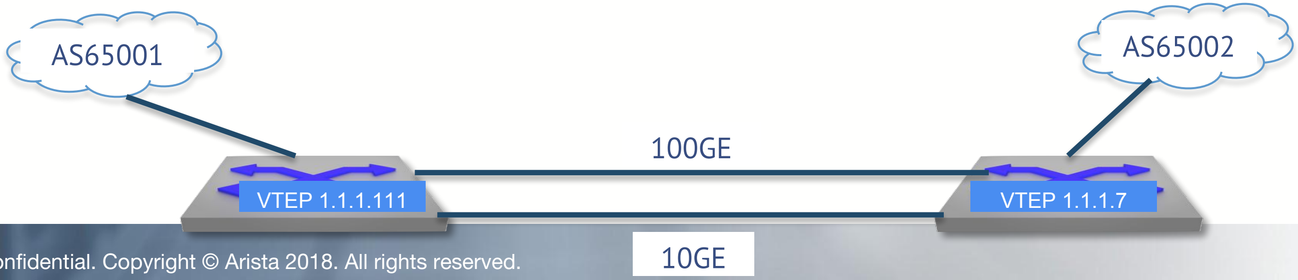
BGP UCMP

```
!
router bgp 64515
   router-id 1.1.1.7
   maximum-paths 64 ecmp 64
   ucmp mode 1 128 0.01
   neighbor 1.1.1.40 peer-group 64512
   neighbor 1.1.1.40 link-bandwidth default 10G
   neighbor 1.1.1.42 peer-group 64512
   neighbor 1.1.1.42 link-bandwidth default 100G
 --More--
```

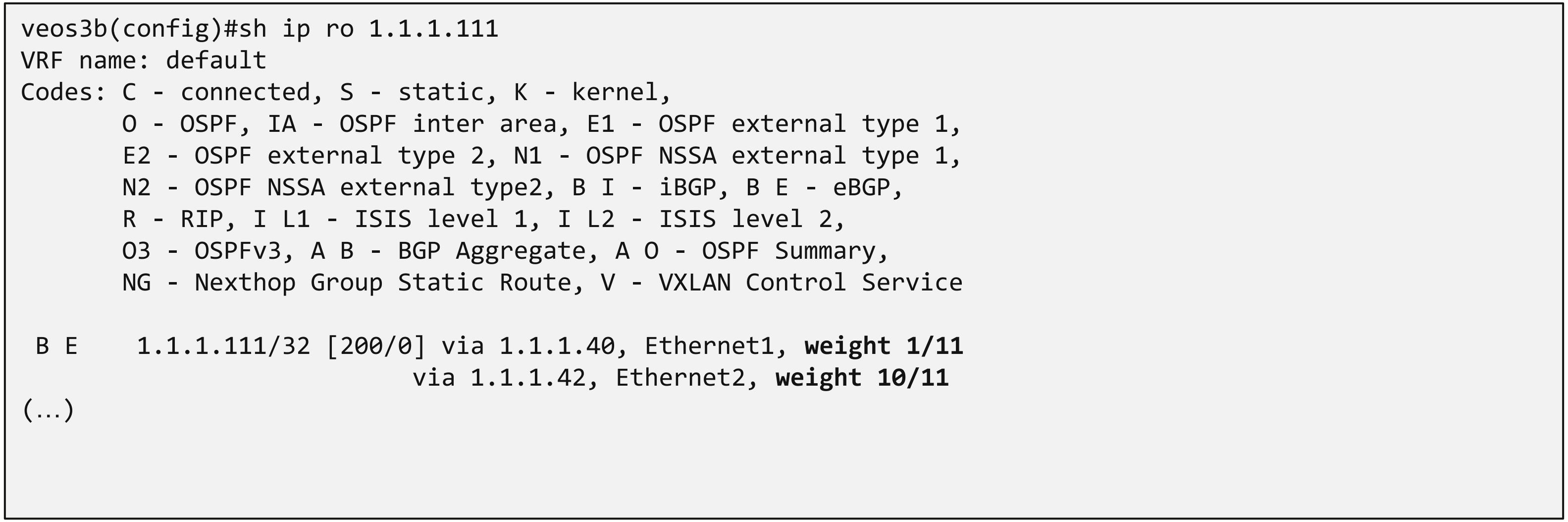


BGP UCMP

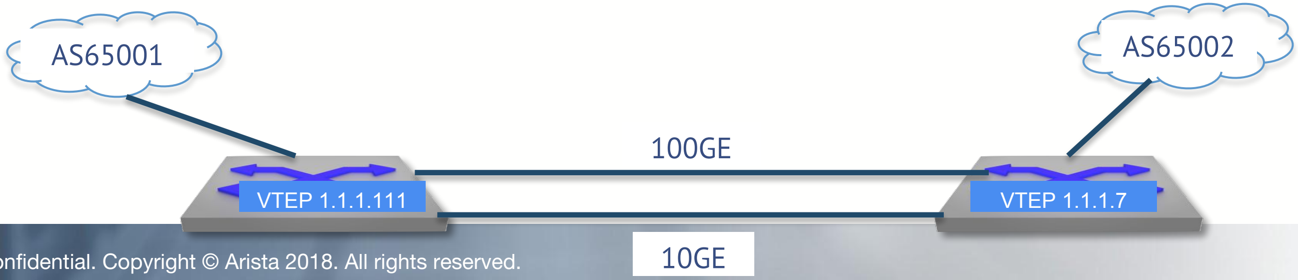
```
veos3b(config)#sh ip bgp 1.1.1.111
...
BGP routing table entry for 1.1.1.1/32
 Paths: 2 available
  64512 64513
    1.1.1.42 from 1.1.1.42 (1.1.1.5)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, ECMP head, ECMP, UCMP, best, ECMP contributor
      Community: 64513:1
      Extended Community: Link-Bandwidth-AS:64512:12499999744.000000(Bps)
      Rx SAFI: Unicast
  64512 64513
    1.1.1.40 from 1.1.1.40 (1.1.1.4)
      Origin IGP, metric 0, localpref 100, weight 0, valid, external, ECMP, UCMP, ECMP contributor
      Community: 64513:1
      Extended Community: Link-Bandwidth-AS:64512:1250000000.000000(Bps)
      Rx SAFI: Unicast
...
```



BGP UCMP



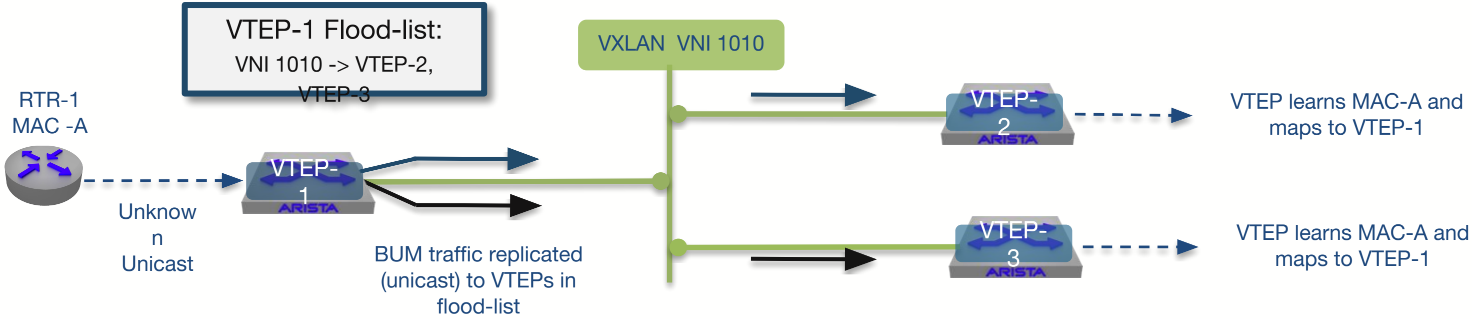
```
veos3b(config)#sh ip ro 1.1.1.111
VRF name: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP,
       R - RIP, I L1 - ISIS level 1, I L2 - ISIS level 2,
       O3 - OSPFv3, A B - BGP Aggregate, A O - OSPF Summary,
       NG - Nexthop Group Static Route, V - VXLAN Control Service

 B E    1.1.1.111/32 [200/0] via 1.1.1.40, Ethernet1, weight 1/11
                           via 1.1.1.42, Ethernet2, weight 10/11
(…)
```



VXLAN Control Plane Options

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


VTEP-1 Flood-list:
VNI 1010 -> VTEP-2, VTEP-3
VXLAN VNI 1010
RTR-1
MAC -A
VTEP-1
VTEP-2
VTEP-3
ARISTA
Unknown Unicast
BUM traffic replicated (unicast) to VTEPs in flood-list
VTEP learns MAC-A and maps to VTEP-1
VTEP learns MAC-A and maps to VTEP-1

Flood list requires provisioning, MAC learning via flood and learn

MAC Security

- Finite set of MAC addresses permitted per edge port
- L2 Access Control List (ACL) - restrict traffic from approved members

```
interface Ethernet3
   description member-A
   switchport access vlan 5
   mac access-group member-A in
```

- Could automatically-generate policy from a database (e.g. IXP-Manager)
- N.B. VXLAN does not change any of this!

Broadcast Control (ARPs, etc.)

- L2 ACLs already limit traffic to only approved speakers
- Storm-control to Broadcast (ARP), Multicast (v6ND)

```
interface Ethernet3
   storm-control broadcast level pps 5000
   storm-control multicast level pps 5000
```

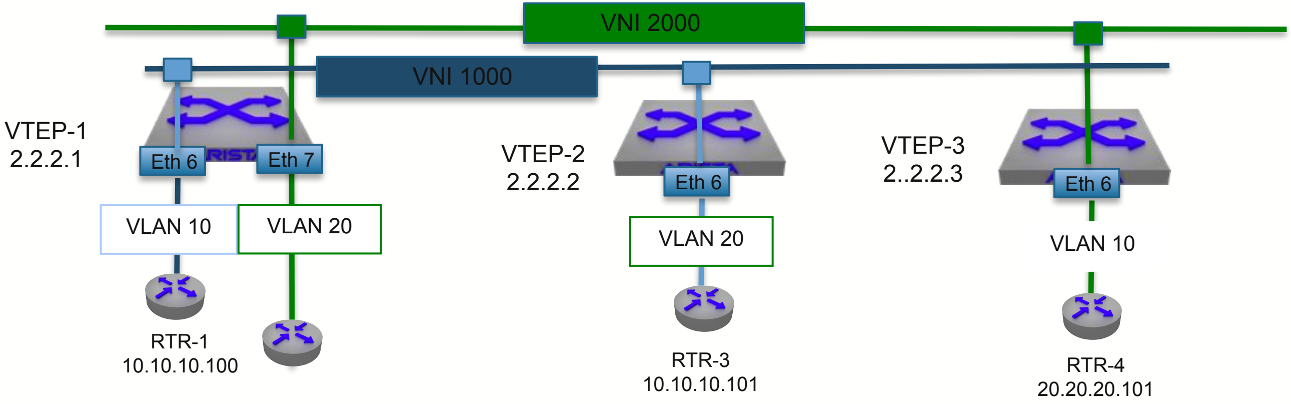
- Statistics can be retrieved via APIs, etc. for automated behaviours
- EVPN provides support for snooping and suppression (see later slides)

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
!
```



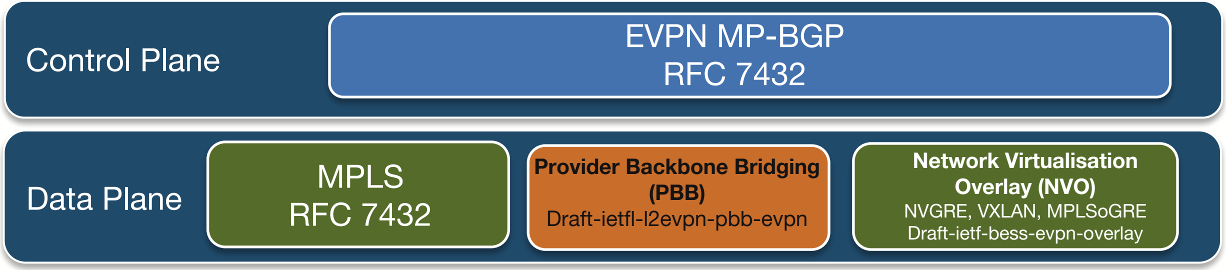
EVPN

What is 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 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)

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 NVO – Network Virtualisation Overlay
  - 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)

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

- 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

EVPN Operation

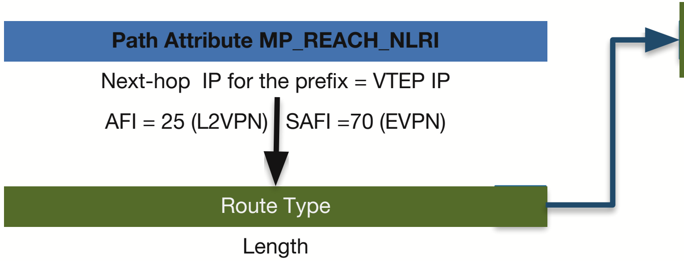
Deploying VXLAN with EVPN

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,

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



Path Attribute MP_REACH_NLRI

Next-hop IP for the prefix = VTEP IP

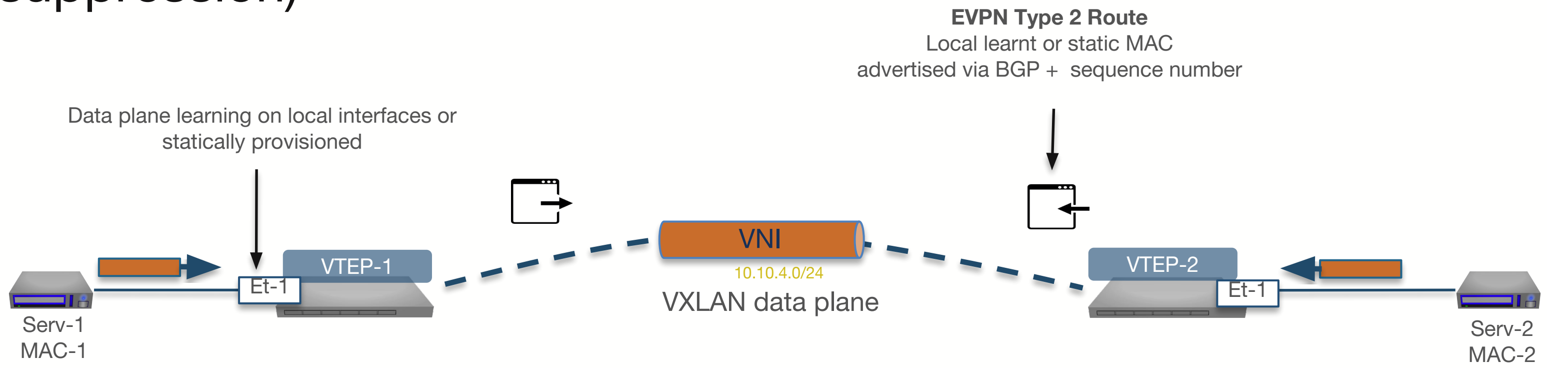
AFI = 25 (L2VPN) | SAFI =70 (EVPN)

Route Type

Length

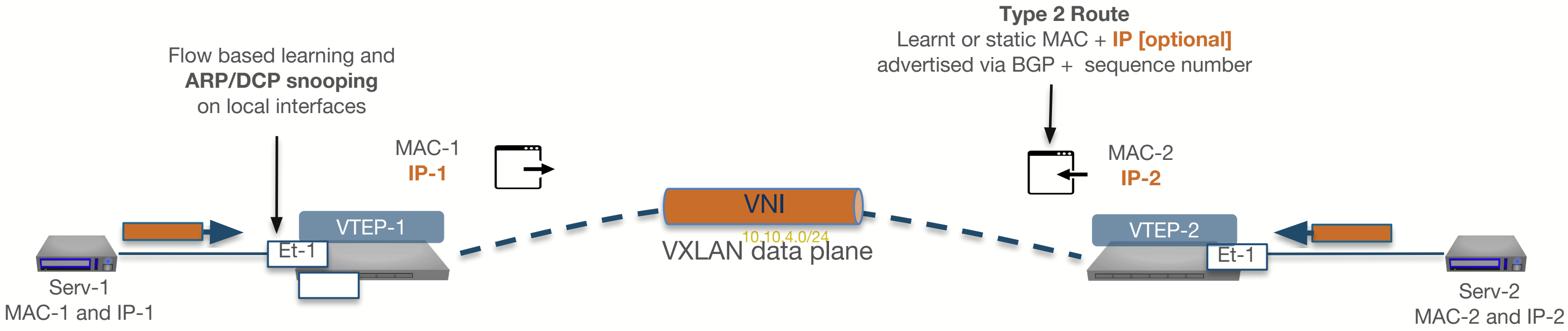
Route Type	Description
1	**Auto-Discover Segment route** – Used to support EVPNs multi-homing deployment models
2	**MAC address Route** - Advertisement of locally learnt/provisioned MAC address and optionally IP addresses.
3	**Inclusive Multicast Ethernet Route** - used to advertise VTEPs VNI membership 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.

EVPN Operation – Type 2 routes (MAC learning, mobility and ARP suppression)



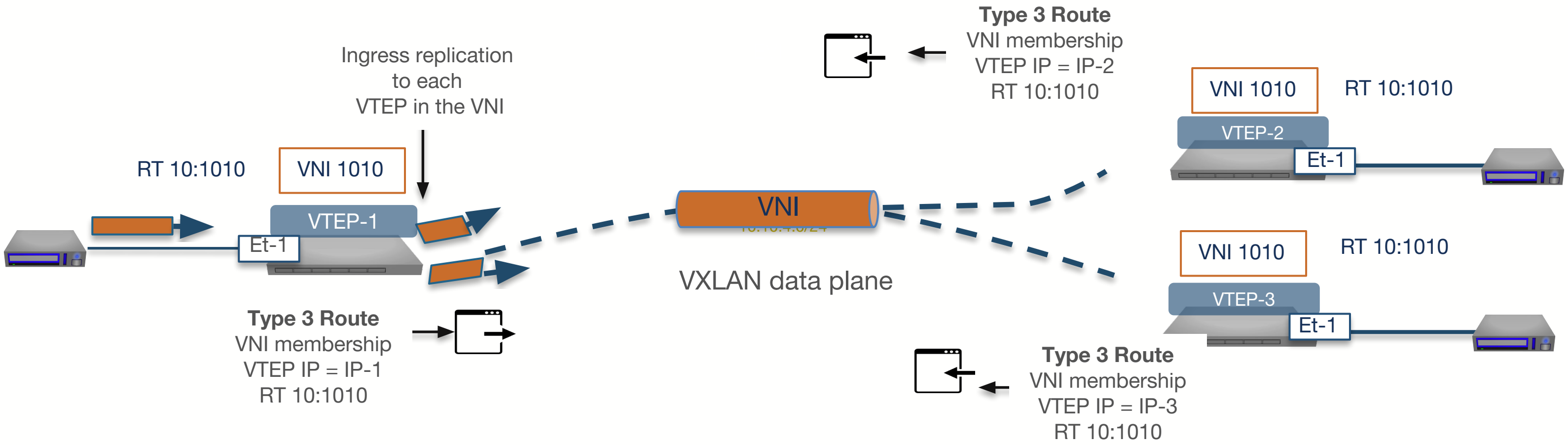
- Flow based MAC learning on the local interfaces of VTEP
  - Locally learnt MACs advertised to BGP peers via EVPN route update
  - Next-hop of the route advertisement set to the IP of the advertising NVE (VTEP) and sequence number for mobility
  - Advertised label in the update, VNI of the MAC-VRF/ L2 domain of the learnt MAC address.
- EVPN Type 2 (MAC route) used to advertise the MAC address
  - MAC and IP [optional] address advertised within the type 2 route
  - Host IP address advertisement can be used for ARP suppression, to reduce flooding in the VNI
  - Arista phase 1 implementation, only MAC addresses are advertised

EVPN Operation – Type 2 route (IP snooping + ARP Suppression)



- Mac Address advertisement can optionally carry the IP address
  - IP to MAC binding learnt via ARP/DHCP snooping on the local VTEP
  - Advertisement of the IP to MAC binding via BGP to remote VTEPs
  - ARP proxy on the remote VTEP to reduce the flooding in the VNI
- EVPN Type 2 (MAC route) used to advertise the [Optional] IP address
  - Can be advertised in a separate update to avoid removal after ARP timeout
  - Arista phase 1 implementation, only MAC address are advertised.

EVPN Operation – Type 3 routes (Ingress replication)



Ingress replication
to each
VTEP in the VNI

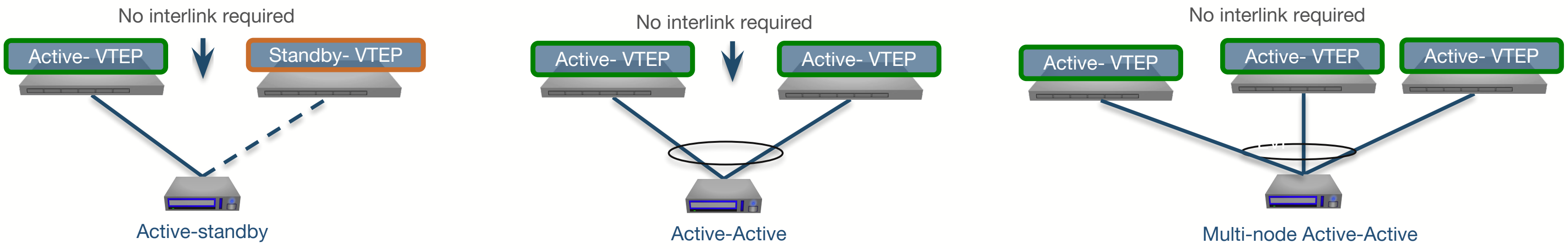
Type 3 Route
VNI membership
VTEP IP = IP-1
RT 10:1010

Type 3 Route
VNI membership
VTEP IP = IP-2
RT 10:1010

Type 3 Route
VNI membership
VTEP IP = IP-3
RT 10:1010

- BUM traffic handled via ingress replication
  - VTEP replicates BUM traffic to each VTEP in the same VNI
  - Each VTEP nodes advertises their local VNI membership status
  - Flood-list of the VTEP dynamically populated based on the advertisement
- EVPN Type 3 (IMET Route) used to advertise EVI/VNI membership
  - Arista implementation will utilize ingress replication (HER), supporting Type 3 routes
  - Multicast forwarding model is an option, not supported in Arista implementation

EVPN Operation – Type 1 and 4 routes (Multi-homing)



- EVPN provides support for Multi-homing hosts and CPE nodes
  - Dual-homing end nodes/hosts to multiple VTEP nodes
  - Support for active-active and active-standby forwarding model plus multi-node solutions
  - VTEP nodes in the model operate independently, they’re not interconnected via a “peer” link
- Arista implementation, will utilize MLAG for multi-homing
  - Support for interoperating with third-party EVPN multi-homing models
  - EVPN multi-homing utilises type 1 and 4 routes