Evpn/vxlan outside the DataCenter

How to create scalable campus and wan solutions with evpn/vxlan

ITNOG 9 19/5/2025

Nicola Modena - CCIE #19119 / JNCIE-SP #986
nicola@modena.to - http://tierzero.it

What I will talk about

- The growing support of evpn/vxlan on devices such as access switches, but also low-end routers and firewalls, enables the use of this technology to create distributed layer-2 and layer-3 connectivity solutions. In this introductory talk, let's understand how the technology works and how to apply it in the best way

About me

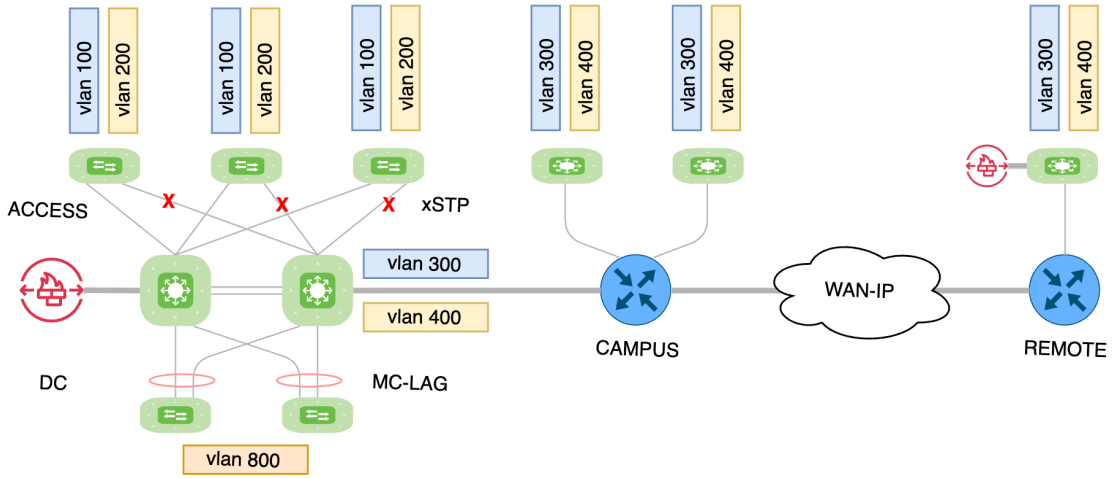
- **Nicola Modena** - CCIE #19119 / JNCIE-SP #986 Emeritus
  Independent Network Architect

  More than 25 years experience designing and implementing
  service provider and large enterprise networks.
  https://tierzero.it | nicola@modena.to

What problems do we want to solve ?

Is it finally time to abandon layer-2?

Problem #1 - Multi Tenancy

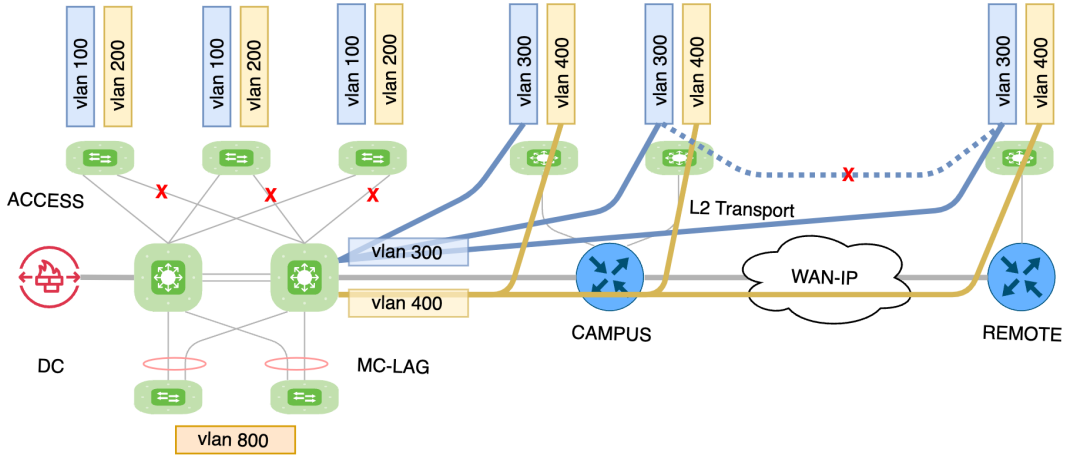


Layer-2 with VLAN it's still the predominant solution for segmentation

Layer-3 with VRF-LITE but it's difficult to scale and propagate in campus and wan

Multitenancy on remote site usually performed replicating central infrastructure

Problem #2 – Optimal Path

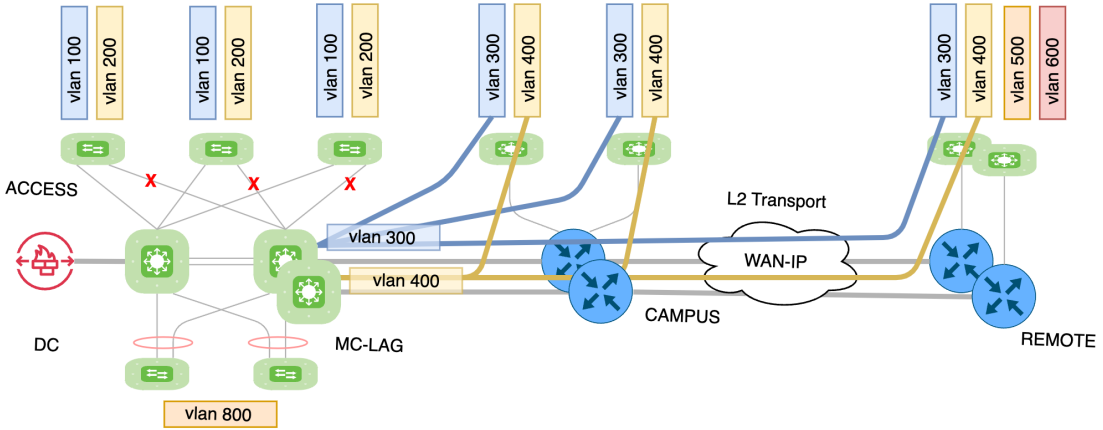


Traditional Layer-2 solution does not provide multi-path and optimal any-to-any path

Also Layer-3 must be centralized and usually with Active/Standby solution (VRRP/HSRP)

All the intra-vlan traffic must transit on the central site for Switching and Routing

Problem #3 – HA and Scalability



Layer-2 solutions lack scalability and have limited solutions for redundancy (Stack,MC-LAG,etc)

We need a reliable solution to handle fault and load distribution

We need to be able to **scale both infrastructure and services**

Problem #4 – Lack of Knowledge

The only solution to layer-2 disaster is awareness, knowledge
….and moving to a higher layer!

In the AI age, do you still use Layer-2 and Spanning-Tree?

Introducing EVPN-VXLAN in Campus and WAN

It’s this technology usable outside the DC / DCI context as an MPLS alternative ?

EVPN/VXLAN – Ethernet VPN (rfc 8365, and more..)

Control Plane

Build and control packet forwarding

BGP with dedicated AFI/SAFI

Advertise :

- Topology
- MAC/IP Address
- IP Prefix
- MCAST groups

Data Plane

Forwards packets

Encapsulated in IP / VXLAN (UDP)

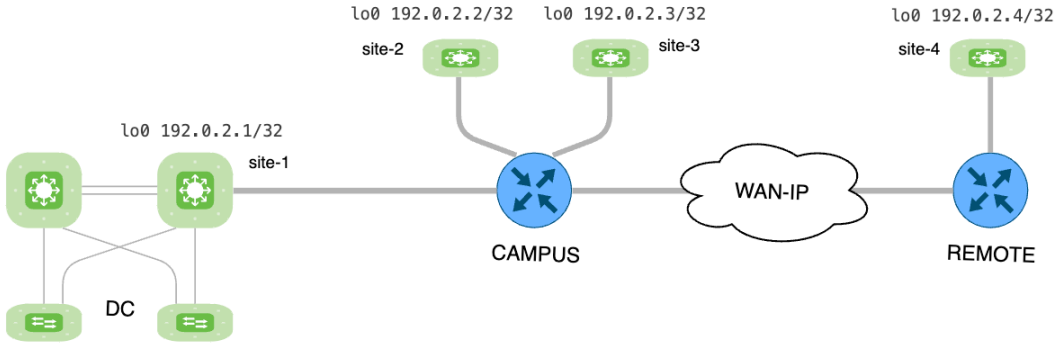
Alternate encap MPLS/MGRE/GENEVE/SRv6

- Use optimal path
- Multipath if it's available
- Handle broadcast and multicast

Underlay: IP connectivity

frrouting.org

```
RR# sh ip route ospf
O>* 192.0.2.1/32 [110/20] via 10.0.0.1,
O>* 192.0.2.2/32 [110/20] via 10.0.0.2,
O>* 192.0.2.3/32 [110/10] via 10.0.0.3,
O>* 192.0.2.4/32 [110/11] via 10.0.0.4,
O   192.0.2.254/32 [110/0] is directly
```

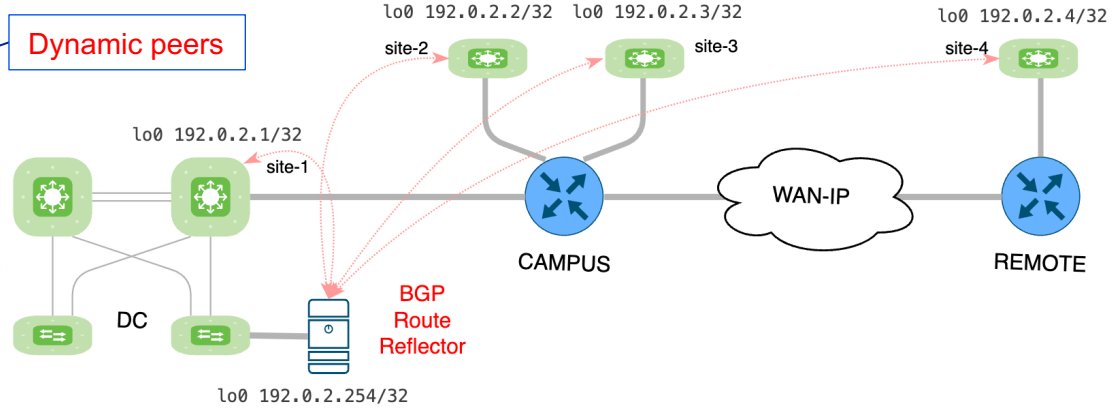


- Just IP connectivity it's required between sites, use loopbacks for device compatibility
- VXLAN add 50 byes for encapsulation, an MTU of 1550 on wan links take away any hassles
- vxlan packet fragmentation it's not supported (RFC 7348 - 4.3)
- mss-adjust, path-mtu-discovery etc, work **only for Layer-3**
- No need for Leaf & Spine fabric, as well as eBGP, for EVPN/VXLAN underlay!

Control Plane : BGP RR

frrouting.org

```
! RR config
!
router bgp 65000
 neighbor WAN peer-group
 neighbor WAN remote-as 65000
 bgp listen range 192.0.2.0/24 peer-group WAN
 !
 address-family l2vpn evpn
  neighbor WAN activate
  neighbor WAN route-reflector-client
 exit-address-family
exit
!
```

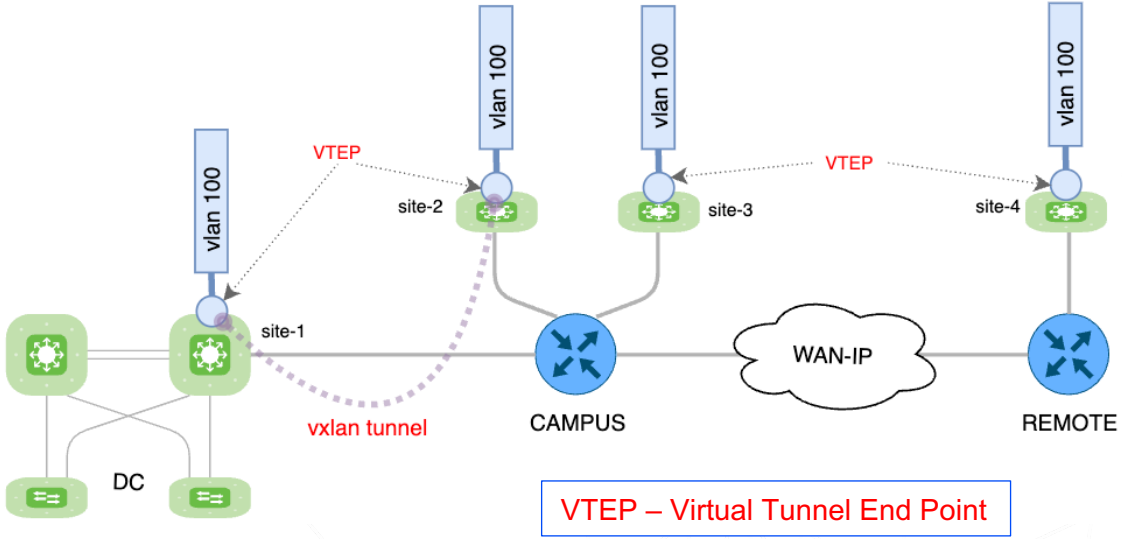


- EVPN uses a dedicated BGP AFI/SAFI, just edge devices must speak BGP, not the transport one
- Although BGP full-mesh is possible, it is preferable to **use a dedicated route-reflector**
- **RR is not involved in routing**, it can be placed anywhere, es: VM with FRR in the Data Center
- You must have two RR for high availability (in different sites) and avoid circular dependency
- Remote devices can peer just with RRs

Configure the VTEP

ARISTA EOS

```
--< site-2 >--
!
interface Loopback0
   ip address 192.0.2.2/32
!
interface Vxlan1
   vxlan source-interface Loopback0
   vxlan vlan 100 vni 100100
!
interface Ethernet2
   switchport access vlan 100
!
interface Ethernet3
   switchport access vlan 100
!
```



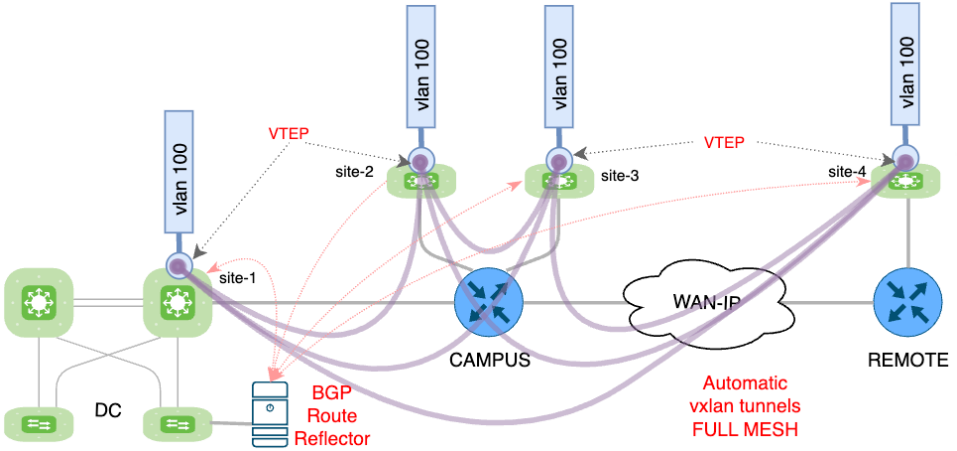
- VNI (Virtual Network Identifier) is equivalent of VLAN ID, just bigger
- Map VLAN to VNI to bridge traffic into VXLAN point-to-multipoint tunnels
- In this example **VLAN 100 is assigned a VNI 100100 and bridged over vxlan tunnel**
- BGP/EVPN distribute this information using **Type-3 advertisement**

Topology : evpn Type-3

frrouting.org

IOS XE

```
RR# sh bgp l2vpn evpn route type 3
…
EVPN type-3 prefix: [3]:[EthTag]:[IPlen]:[OrigIP]
*>i[3]:[0]:[32]:[192.0.2.1]
*>i[3]:[0]:[32]:[192.0.2.2]
*>i[3]:[0]:[32]:[192.0.2.3]
*>i[3]:[0]:[32]:[192.0.2.4]
```

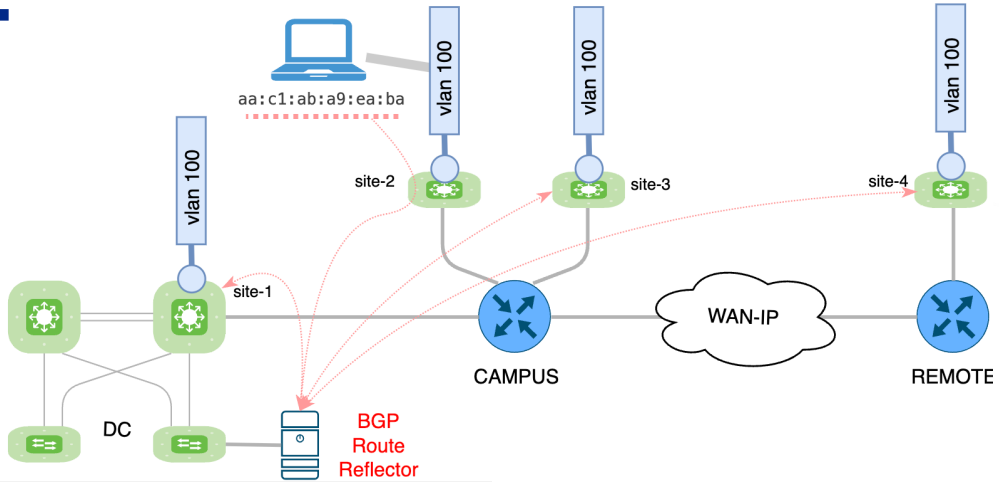



```
site-4#show nve peers
Interface  VNI      Type Peer-IP          RMAC/Num_RTs   eVNI     state flags UP time
nve1       100100   L2CP 192.0.2.1        3              100100     UP   N/A  01:51:30
nve1       100100   L2CP 192.0.2.2        2              100100     UP   N/A  01:51:31
nve1       100100   L2CP 192.0.2.3        2              100100     UP   N/A  01:51:31
```

- The Route Reflector distribute this information to all evpn devices
- A **full-mesh of vxlan tunnels it's automatically created** on all devices with this VNI (&RT)

MAC Addr : Type-2

frrouting.org ARISTA EOS




```
site-2#sh mac address-table
Vlan    Mac Address       Type        Ports
----    -----------       ----        -----
 100    aac1.aba9.eaba    DYNAMIC     Et2
 100    aac1.abb1.a1b3    DYNAMIC     Et3
```

```
RR# sh bgp l2vpn evpn route detail type 2
BGP routing table entry for 192.0.2.2:100:[2]:[0]:[48]:[aa:c1:ab:a9:ea:ba]
Paths: (1 available, best #1)
  Advertised to non peer-group peers:
  192.0.2.1 192.0.2.3 192.0.2.4
  Route [2]:[0]:[48]:[aa:c1:ab:a9:ea:ba] VNI 100100
```

MAC ADDRESS

VNI

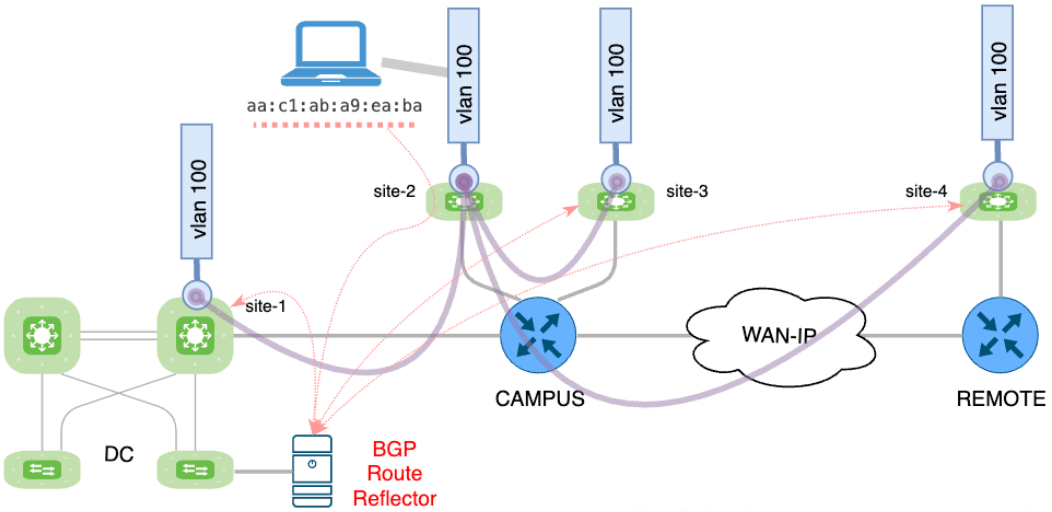
As soon as a new local mac-address is discovered:

- Is advertised as **Type-2 advertisement** with the corresponding VNI
- The Route Reflector distributes this information to all the devices

MAC Addr : Type-2

ARISTA EOS

```
site-1#sh mac address-table
Vlan    Mac Address       Type        Ports
----    -----------       ----        -----
 100    aac1.aba9.eaba    DYNAMIC     Vx1
 100    aac1.abb1.acac    DYNAMIC     Et2
 100    aac1.ab3e.43af    DYNAMIC     Et2
```



As soon as the type-2 mac advertisement is received on all the remote devices:

- The bridging table is pre-loaded with mac address and remote vtep address.
- Traffic can be bridged using the specific VXLAN tunnel

--> No more “flood and learn” or “unknow unicast” with the evpn control-plane <--

Layer-2 ++

Some Layer-2 enhancements provided by the evpn control-plane:

Multihoming :: ESI Ethernet Segment Identifier Type-1/Type-4 for multihoming A/A & A/S
Mac Mobility :: community & operations to speed-up L2 convergence and prevent L2 loops
Massive Withdraw :: withdraw all the mac-addresses with a single BGP message
Unknown Mac Route :: “L2 default gateway” useful for hub&spoke and mac scale: new rfc9014

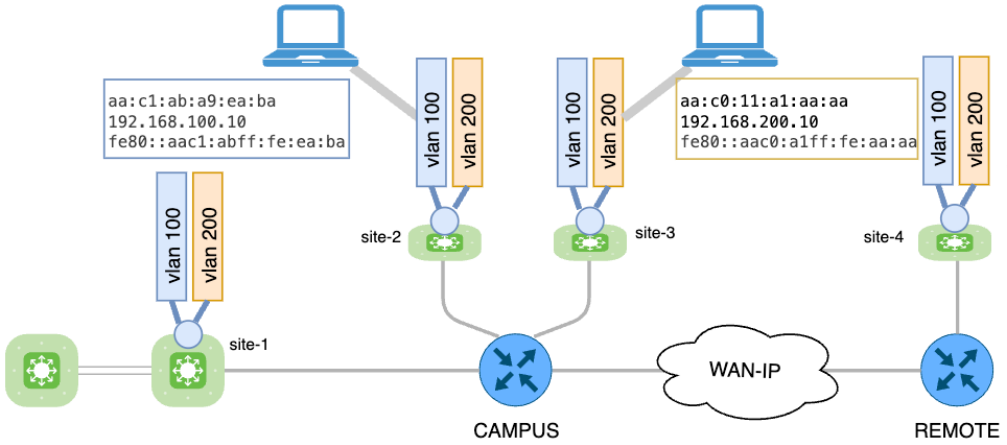
Introducing Layer-3

Because evpn is much more than just mac-vrf

Discover Layer-3 info

frrouting.org

VTEP discovers IP from ARP,DHCP,NDP
- Advertises into evpn **Type-2 (MAC/IP)**
- May respond locally to ARP & ND &
Suppresses broadcast traffic over tunnels
wait? does it also support ipv6? Yes..




```
RR# sh bgp l2vpn evpn route detail type 2
BGP routing table entry for 192.0.2.2:100:[2]:[0]:[48]:[aa:c1:ab:a9:ea:ba]:[32]:[192.168.100.10]
Paths: (1 available, best #1)
  Advertised to non peer-group peers:
  192.0.2.1 192.0.2.3 192.0.2.4
  Route [2]:[0]:[48]:[aa:c1:ab:a9:ea:ba]:[32]:[192.168.100.10] VNI 100100
…
BGP routing table entry for 192.0.2.5:100:[2]:[0]:[48]:[aa:c1:ab:a9:ea:ba]:[128]:[fe80::aac1:abff:fea9:eaba]
Paths: (1 available, best #1)
  Advertised to non peer-group peers:
 192.0.2.1 192.0.2.3 192.0.2.4
  Route [2]:[0]:[48]:[aa:c1:ab:a9:ea:ba]:[128]:[fe80::aac1:abff:fea9:eaba] VNI 100100
…
```

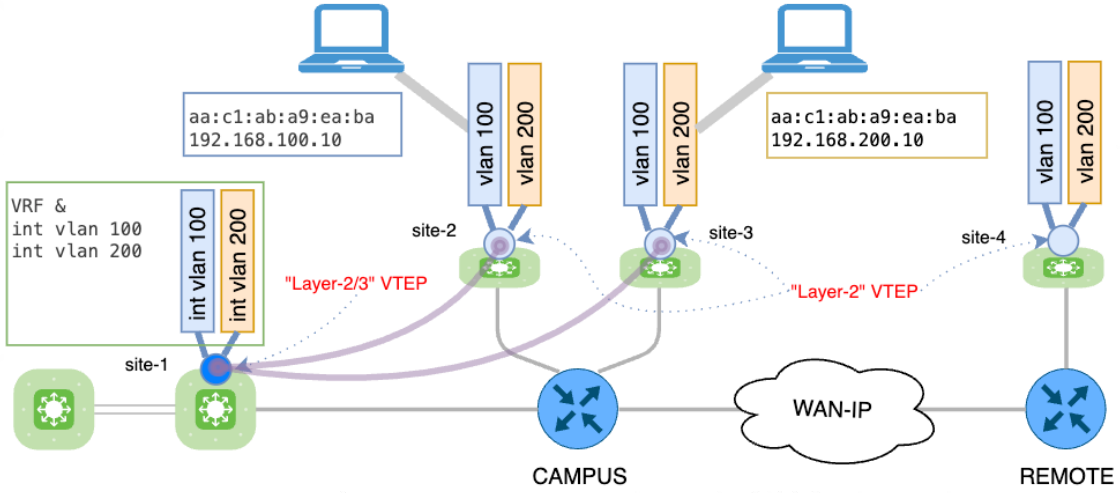
MAC ADDRESS
IP ADDRESS
VNI
MAC ADDRESS
IPv6 ADDRESS
VNI

Centralized vxlan routing

ARISTA EOS

```
--< SITE-1 >-- VRF and Layer 3 SVI --

interface Vlan100
   vrf VRF-1
   ip address 192.168.100.1/24
!
interface Vlan200
   vrf VRF-1
   ip address 192.168.200.1/24
!
interface Vxlan1
   vxlan vlan 100 vni 100100
   vxlan vlan 200 vni 100200
!
```



CRB – Centrally-Route Bridging
CGW – Centralized Default Gateway

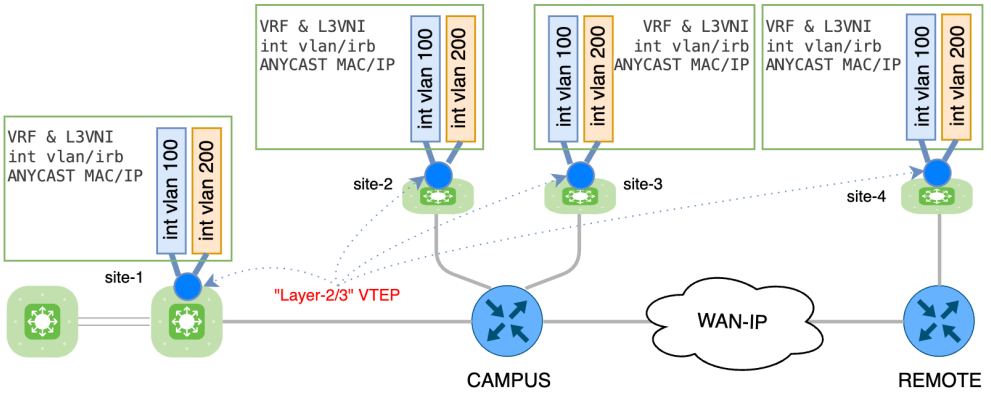
A central VTEP advertises default-gateway MAC/IP

- Source L2VTEP bridges traffic to L3VTEP advertising GW IP/MAC (type-2)
- Central VTEP performs full vxlan routing (using a form of asymmetric vxlan routing)
- Central VTEP bridges the packet using the VXLAN tunnel to the destination L2VTEP

Layer-3 VNI for distributed routing

```
[edit routing-instances]
 VRF-1 {
    instance-type vrf;
    protocols {
        evpn {
            irb-symmetric-routing {
                vni 300300;
    }}}
    interface irb.100;
    interface irb.200;
}
```

L3 VNI




```
RR# sh bgp l2vpn evpn route detail type 2
BGP routing table entry for 192.0.2.3:100:[2]:[0]:[48]:[aa:c1:ab:a9:ea:ba]:[32]:[192.168.100.10]
…
  Route [2]:[100100]:[48]:[aa:c1:ab:a9:ea:ba]:[32]:[192.168.100.10] VNI 100100/300300
      Extended Community: RT:65000:100 RT:65000:300 ET:8 Default Gateway Rmac:2c:6b:f5:17:05:f0
```

L2 VNI

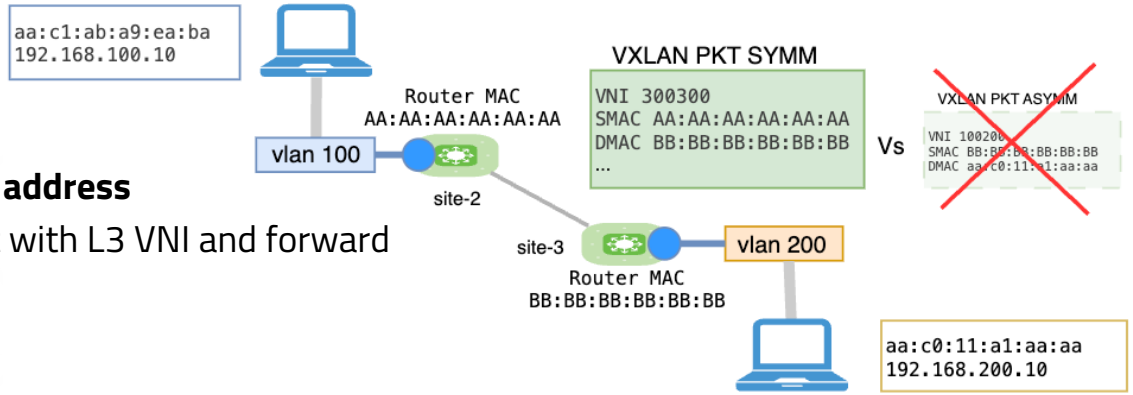
L3 VNI

RMAC

A unique "L3" VNI can be assigned to the VRF for all routing operations **L3VNI <-> VRF**

- Must be defined on all L3 enabled VTEPs with all SVI/IRB layer-3 interfaces
- All the L3 gateways share the same MAC/IP -> **Anycast Gateway** for optimal operations
- Type-2 has now 2 VNI: for switching and for routing (and vtep RMAC community)

Symmetric vxlan routing



Source L3VTEP:

- **Use dest VTEP RMAC as dest-mac address**
- Decrement TTL,encapsulate packet with L3 VNI and forward

Destination L3VTEP:

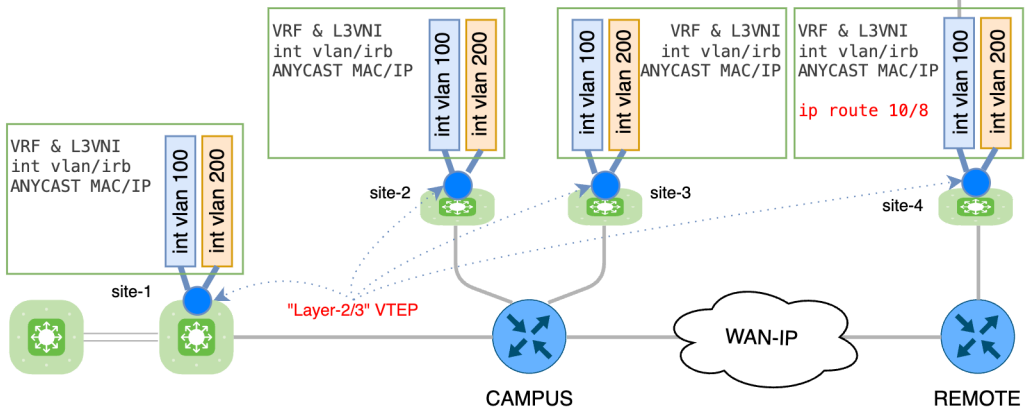
- Identify required cooperation in routing from L3VNI and dest RMAC
- **Perform full-blown L3 forwarding**

The source VTEP does not require to use destination mac-address to perform L3 operation

- **Increased scalability:** less information needed (and lower tcam usage)
- It's not required to configure/import all the L2 vlan/information on all devices
- L3VNI is really the Virtual Ethernet Segment between edge routers

IP prefix: Type-5

```
--< Site-4 >--
!
router bgp 65000
!
 address-family ipv4 vrf VRF-1
  redistribute connected
  redistribute static
 exit-address-family
!
ip route vrf vrf-1 10.0.0.0 255.0.0.0 192.168.
```



MASK PREFIX L3 VNI

```
RR# sh bgp l2vpn evpn route detail type 5
BGP routing table entry for 192.0.2.3:9:[5]:[0]:[8]:[10.0.0.0]
Advertised to non peer-group peers:
  192.0.2.1 192.0.2.2 192.0.2.3
  Route [5]:[0]:[8]:[10.0.0.0] VNI 300300
      Extended Community: RT:65000:300 ET:8 Rmac:2c:6b:f5:17:05:f0
```

RMAC

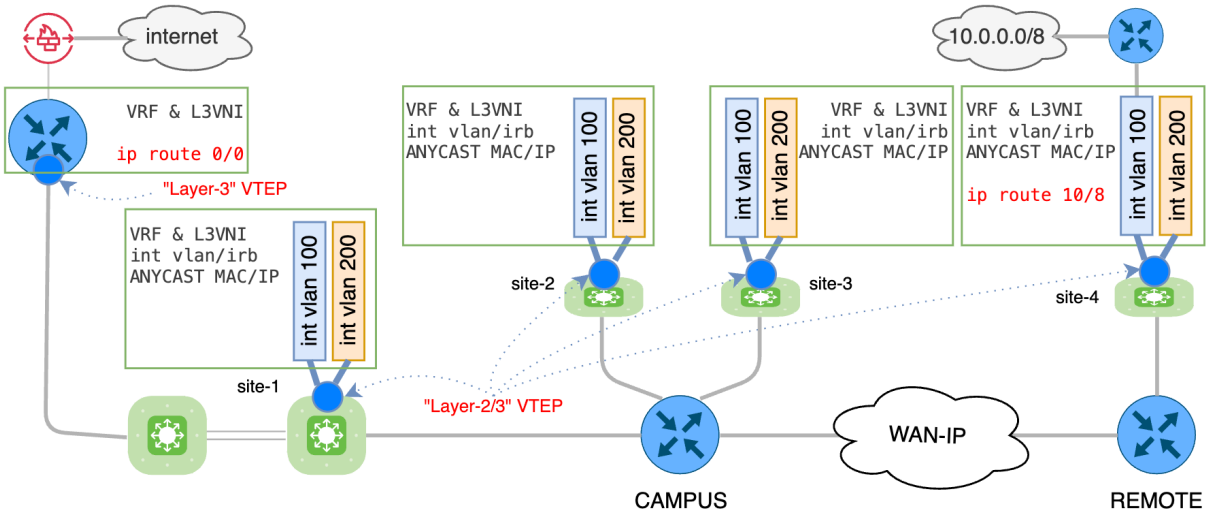
To perform routing, it's required ip prefix, L3VNI and RMAC regardless it's a HOST or NETWORK

EVPN type-5 advertises prefixes with L3VNI (no more mac address) for integrated routing

IP prefix: Type-5

```
--< Internet Router >--

[edit routing-instances]
VRF-1 {
  instance-type vrf;
  routing-options {
    static {
      route 0.0.0.0/0 next-hop 1.2.3.4;
  }}
  protocols {
    evpn {
      ip-prefix-routes {
        advertise direct-nexthop;
        encapsulation vxlan;
        vni 300300;
}}}}
```



External connectivity can be injected from router EVPN/VXLAN enabled with just Type-5 prefix

Layer-3 only VTEP does not require mac knowledge, and just perform routing

Route leaking between different VRF is also supported, but with platform/vendor constrains

There is also “pure type-5 route”, VMTO, and so on, but we do not have time…

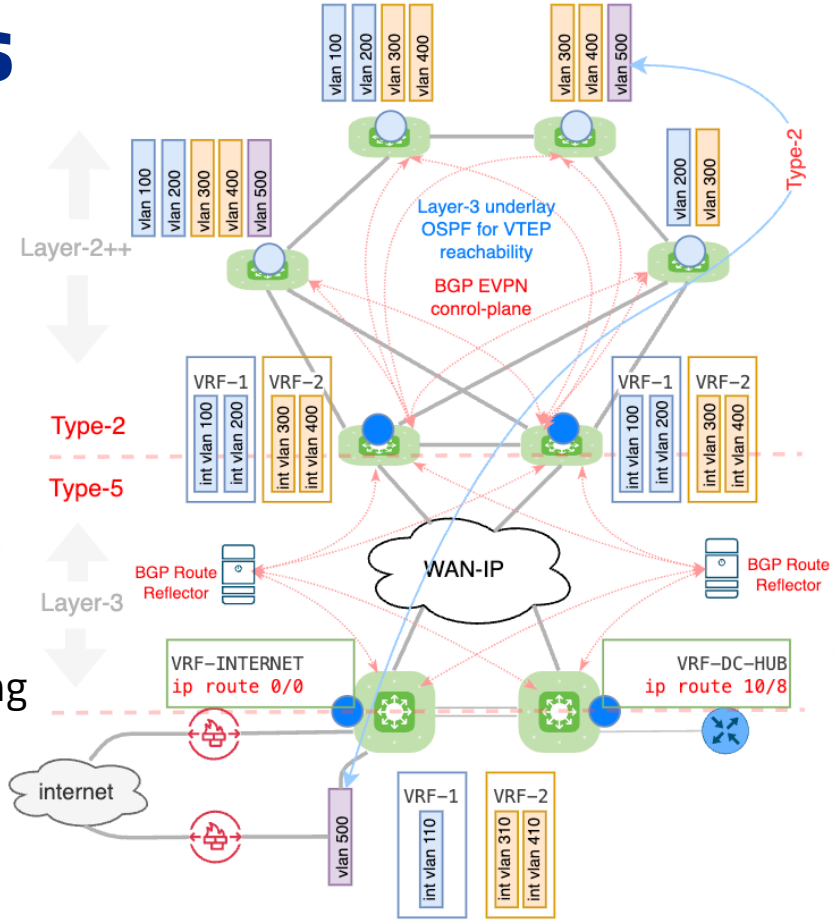
Let’s try to put it all together

because all the features must be used in the right way

EVPN/VXLAN solutions

Use EVPN/VXLAN end-to-end for uniform solution

- Routed underlay with IGP -> OSPF
- ECMP & LFA are always welcome!
- RR hierarchy to maximize service continuity
- L2 VTEP & CRB for simplicity, compatibility & cost
- Place L3 VTEP carefully for optimal underlay routing
- Prefer Layer-3 & Type-5 for remote sites
- Stretched L2 only for specific cases (MTU attn!)
- Route leaking between VRF is platform dependent



A glimpse on some other topics

ESI LAG – Ethernet Segment Identifier – multi-homing active/standby or active/active
VPWS – point-to-point L2 connection without mac learning (pseudowire)
GPO – Group Policy Options – add security tags/function into VXLAN encoding
IPv6 – as underlay protocol

SDN Integration:

Proxomx has native support for evpn-vxlan trough integrated FRR
VmWare NSX has EVPN/VXLAN support with type-5 host route
Full Linux kernel support, static & dynamic with FRR as control-plane

Conclusion

let's try to recap

Conclusion

If you get:

“VXLAN extend layer-2 over IP, this solves all my problems”

I have completely failed on problem #4, the right message is:

“EVPN/VXLAN seamlessly integrates L3 capabilities, making it a first choice for multitenancy and multivendor CAMPUS and WAN solutions”

or

When in doubt, route.

(better with a bike)

Any questions ?

you can find me:

nicola@modena.to

linkedin.com/in/nmodena

Itnog telegram group

This presentation (and future updates) at https://github.com/nmodena/blog

Special thank to Christian,Ivan,Tiziano and Davide for the revision