LET’S BUILD TOMORROW TODAY
VXLAN Deployment Models
A practical perspective
Victor Moreno, Distinguished Engineer

BRKDCT-2404
Agenda

• Why VXLAN?
• VXLAN Fundamentals
• Overlay Deployment Considerations
• Underlay Deployment Considerations
• Summary and Conclusion
Trend: Flexible Data Center Fabrics

Use VXLAN to Create DC Fabrics

Mobility
Segmentation + Policy
Scale
Automated & Programmable
Full Cross Sectional BW
L2 + L3 Connectivity
Physical + Virtual
VXLAN Fundamentals
Why Overlays?

Seek well integrated best in class Overlays and Underlays

Robust Underlay/Fabric
- High Capacity Resilient Fabric
- Intelligent Packet Handling
- Programmable & Manageable

Flexible Overlay Virtual Network
- Mobility – Track end-point attach at edges
- Segmentation
- Scale – Reduce core state
  - Distribute and partition state to network edge
- Flexibility/Programmability
  - Reduced number of touch points
Overlay Taxonomy

Service = Virtual Network Instance (VNI)
Identifier = VN Identifier (VNID)
NVE = Network Virtualization Edge
VTEP = VXLAN Tunnel End-Point
VXLAN is an Overlay Encapsulation

**Data Plane Learning**
Flood and Learn over a multidestination distribution tree joined by all edge devices

**Protocol Learning**
Advertise hosts in a protocol amongst edge devices
VXLAN Packet Structure

Ethernet in IP with a shim for scalable segmentation

- Outer MAC Header
- Outer IP Header
- Outer UDP Header
- VXLAN Header
- Original Layer 2 Frame
- FCS

- Dest. MAC Address
- Src. MAC Address
- VLAN Type (0x8100)
- VLAN Tag
- Ether Type (0x0800)
- 48
- 48
- 16
- 16
- 16

- Next-Hop MAC Address

- IP Header
  - IPv4 Type (0x0800)
  - VLAN Type (0x8100)
  - VLAN Tag
- 72
- 8
- 16
- 32
- 32

- IPv4 Header
  - Protocol (0x011)
  - Header Checksum
  - Source IP
  - Dest. IP
- 8
- 16
- 16
- 16

- UDP Header
  - UDP Port
  - Destination Port
  - UDP Length
  - Checksum
- 8
- 16
- 16
- 8

- VXLAN Header
  - VXLAN Flags (RRRRIRRR)
  - VXLAN Port
  - Source Port
  - VXLAN Length
  - Checksum 0x0000
- 8
- 8
- 16
- 16

- Source Port
- Dest. Port
- VXLAN Length
- Checksum 0x0000
- 8
- 8
- 16
- 16

- Dest. IP
- Source IP
- Header
- 32
- 32
- 16
- 8

- Tunnel Entropy

- Ethernet Payload

- Large scale segmentation

- 50 (54) Bytes of overhead

- Source Port
- Dest. Port
- VXLAN Length
- Checksum 0x0000
- 8
- 8
- 16
- 16

- VNI
- Reserved

- 8
- 24
- 24
- 8

- Hash of the inner L2/L3/L4 headers of the original frame. Enables entropy for ECMP Load balancing in the Network.

- Tunnel Entropy

- Allows for 16M possible segments
Data Plane Learning

Dedicated Multicast Distribution Tree per VNI

VTEP

PIM Join for Multicast Group 239.1.1.1

PIM Join for Multicast Group 239.2.2.2

Web VM

DB VM

DB VM

Web VM

L3 Core

Multicast-enabled Transport

Web VM

DB VM

DB VM

Web VM
Data Plane Learning
Learning on Broadcast Source - ARP Request Example

VM 1
VTEP 1
1.1.1.1

VM 2
VTEP 2
2.2.2.2

VM 3
VTEP 3
3.3.3.3

IP A ➔ G

Multicast-enabled Transport
Data Plane Learning
Learning on Unicast Source - ARP Response Example

VM 1
VTEP 1 1.1.1.1

VM 2
VTEP 2 2.2.2.2

VM 3
VTEP 3 3.3.3.3

MAC IP Addr
VM 2 VTEP 2

MAC IP Addr
VM 1 VTEP 1

ARP Resp

ARP Resp

ARP Resp

VTEP 2 ➔ VTEP 1

Multicast-enabled Transport
-sharing multicast groups across VNIs
Overlay Network Evolution: Edge Devices

**Network Overlays**
- Router/switch end-points
- Protocols for resiliency/loops
- Traditional VPNs
- OTV, VPLS, LISP, FP

**Host Overlays**
- Virtual end-points only
- Single admin domain
- VXLAN, NVGRE, STT

**Hybrid Overlays**
- Physical and Virtual
- Resiliency + Scale
- x-organisations/federation
- Open Standards
VXLAN Evolution

Multicast Independent
- Head-end replication enables unicast-only mode
- Control Plane provides dynamic VTEP discovery

Protocol Learning prevents floods
- Workload MAC addresses learnt by VXLAN NVEs
- Advertise L2/L3 address-to-VTEP association information in a protocol

External Connectivity
- VXLAN HW Gateways to other encaps/networks
- VXLAN HW Gateway redundancy
- Enable hybrid overlays

IP Services
- VXLAN Routing
- Distributed IP Gateways
VXLAN Evolution: Using a Control Protocol

VTEP Discovery

VTEPs advertise their VNI membership in BGP

1. BGP consolidates and propagates VTEP list for VNI

2. VTEP obtains list of VTEP neighbors for each VNI

3. Overlay Neighbors
   - POD3, IP C
   - POD2, IP B

4. VTEP can perform Head-End Replication
VXLAN Unicast Mode

**Head-end replication**

A host sends a L2 BUM* frame

1. **BUM Frame**

2. **VTEP retrieves the list of Overlay Neighbors**

3. **VTEP performs Head-End Replication**

4. **VXLAN Encap**

5. **Frames are unicast to the neighbors**

*Broadcast, Unknown Unicast or Multicast

**Information statically configured or dynamically retrieved via control plane (VTEP discovery)**
VXLAN Evolution: Using a Control Protocol

Protocol Learning

VTEPs advertise host routes (IP+MAC) to local hosts

1. VTEPs advertise host routes (IP+MAC) to local hosts

2. BGP propagates routes for the host to all other VTEPs

3. VTEPs obtain host routes for remote hosts and install in RIB/FIB

Overlay Forwarding Table
Host1 <MAC,IP>, VTEP IP A
Host2 <MAC,IP>, VTEP IP B

Overlay Forwarding Table
Host1 <MAC,IP>, VTEP IP A

BGP Route Reflector

BGP propagates routes for the host to all other VTEPs

VTEPs obtain host routes for remote hosts and install in RIB/FIB
Host Route Distribution decoupled from the Underlay protocol

Use MP-BGP on the leaf nodes to distribute internal host/subnet routes and external reachability information
VXLAN Control Plane

Host Advertisement

1. Host Attaches
2. Attachment NVE advertises host’s MAC (+IP) through BGP RR
3. Choice of encapsulation is also advertised

NLRI:
- Host MAC1, IP1
- NVE IP L1/MAC L1
- VNI 5000
- Ext. Community:
  - Encapsulation: VXLAN, NVGRE
  - Sequence 0

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
<th>VNI</th>
<th>Next-Hop</th>
<th>Encap</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5000</td>
<td>IP L1</td>
<td>VXLAN</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MAC L1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VNI 5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host 1</td>
</tr>
<tr>
<td>VLAN 10</td>
</tr>
</tbody>
</table>
VXLAN Control Plane

Host Moves

1. Host Moves to NVE3
2. NVE3 detects Host1 and advertises H1 with seq#1
3. NVE1 sees more recent route and withdraws its advertisement

NLRI:
- Host MAC1, IP1
- NVE IP L3/MAC L3
- VNI 5000

Ext. Community:
- Encapsulation: VXLAN, NVGRE
- Sequence 1

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
<th>VNI</th>
<th>Next-Hop</th>
<th>Encap</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5000</td>
<td></td>
<td>IP L3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MAC L3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VXLAN</td>
<td>1</td>
</tr>
</tbody>
</table>
VXLAN L2 and L3 Gateways

Connecting VXLAN to the broader network

L2 Gateway: VXLAN to VLAN Bridging

Ingress VXLAN packet on Orange segment

Egress interface chosen (bridge may .1Q tag the packet)

Destination is in another segment.
Packet is routed to the new segment

L3 Gateway: VXLAN to X Routing

- VXLAN
- VLAN

<table>
<thead>
<tr>
<th></th>
<th>N1KV</th>
<th>N7K w/F3 LC</th>
<th>Nexus 3K</th>
<th>N5K/6K</th>
<th>N9K</th>
<th>CSR 1000V</th>
<th>ASR1K/ASR 9K</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 Gateway</td>
<td>VXGW on 1110</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>L3 Gateway</td>
<td>CSR 1000V</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The Multi-encapsulation Gateway

- Multi-encapsulation Gateway:
  - VXLAN, NVGRE, MPLS, LISP, VLAN, OTV
- Bridging (L2 Gateway)
- Routing (L3 Gateway)
VXLAN Evolution: L3 Services with VXLAN

- Forward based on IP address (learnt via Control Protocol)
- Make routing decisions at VTEPs
- Leverage L3 Gateway capabilities along with Protocol Information
- Reduce impact of ARP on network
- Reduce exposure to floods
Evolution of the VXLAN Data Plane

Beyond Ethernet in IP ➔ GPE: Generic Protocol Encapsulation

Payload:
- IP
- Ethernet
- other

50 Bytes of overhead

14 Bytes (4 Bytes Optional)

20 Bytes

8 Bytes

8 Bytes

Hash of the inner L2/L3/L4 headers of the original frame. Enables entropy for ECMP Load balancing in the Network.

24 bit Protocol Type field (previously reserved)
Overlay Deployment Considerations
Layer 2 Overlays
- Emulate a LAN segment
- Transport Ethernet Frames (IP and non-IP)
- Single subnet mobility (L2 domain)
- Exposure to open L2 flooding
- Useful in emulating physical topologies

Layer 3 Overlays
- Abstract IP based connectivity
- Transport IP Packets
- Full mobility regardless of subnets
- Contain network related failures (floods)
- Useful in abstracting connectivity and policy

Hybrid L2/L3 Overlays offer the best of both domains
L2 VXLAN Fabric
VXLAN Configuration – Global Configuration

Changing default UDP port for VxLAN:

```
vxlan udp port <number>
```

<table>
<thead>
<tr>
<th>UDP Port</th>
<th>N1KV</th>
<th>Nexus 7K with F3 LC</th>
<th>Nexus 3K</th>
<th>Nexus 5K/6K</th>
<th>Nexus 9K Standalone</th>
<th>CSR 1000V ASR1K</th>
<th>ASR9k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V 3.1 IANA port 4789</td>
<td>IANA port 4789</td>
<td>IANA port 4789</td>
<td>IANA port 4789</td>
<td>IANA port 4789</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(configurable)</td>
<td>(configurable)</td>
<td>(Config Future)</td>
<td>(configurable)</td>
<td>(Config Future)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Default Port</td>
<td>4789</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IANA port 4789 (configurable)
VXLAN Configuration – Overlay Configuration

```plaintext
interface nve1
  no shutdown
  source-interface loopback1
  member vni 6000 mcast-group 235.1.1.1
```

- **VxLAN Identifier**
- **IP Multicast Group for Multi-destination Traffic**
- **Used to Derive Local VTEP IP address**
- **Point to Multi-point tunnel with VxLAN encapsulation**
VXLAN Configuration – Mapping VLANs to VNIs

Layer 2 Gateway

VLAN CLI Model

```
vlan 3002
vn-segment 6000

interface nve1
  no shutdown
  source-interface loopback1
  member vni 6000 mcast-group 235.1.1.1
```

EFP/VSI CLI Model

```
vni 6000

Bridge-domain 100
  member vni 6000

interface nve1
  no shutdown
  source-interface loopback1
  member vni 6000 mcast-group 235.1.1.1
```

Map VNI to VLAN/BD

**VXLAN tunnel interface**

**range**

**range**
## L2 VXLAN Configuration – Edge port configuration

### VLAN CLI Model

```bash
text>
interface <phy if>
  switchport mode access
  switchport access vlan 3002

vlan 3002
  vn-segment 6000
```

### EFP CLI Model

```bash
text>
interface <phy if>
  service instance <id> vni
    encapsulation dot1q 200 vni 6000

bridge domain 200
  member vni 6000
  member interface <phy if> service instance <id>
```
L2 VXLAN Configuration – Edge port configuration

**VLAN CLI Model**

```bash
interface <phy if>
  switchport mode access
  switch port access vlan 3002

vlan 3002
  vn-segment 6000
```

**VSI CLI Model**

```bash
encapsulation profile vni <name>
  dot1q 100,200 vni 5000,6000

interface <phy if>
  service instance 1 vni
  encapsulation profile <name> default

bridge domain 200
  member vni 6000
  member interface <phy if> service instance <id>
```
## L2 VXLAN Configuration – Edge port configuration

### VSI CLI Model

```
encapsulation profile vni <name>
dot1q 100,200 vni 5000,6000

interface <phy if>
  service instance 1 vni
  encapsulation profile <name> default

bridge domain 200
  member vni 6000
```

### EFP CLI Model

```
interface <phy if>
  service instance <id> vni
  encapsulation dot1q 200 vni 6000

bridge domain 200
  member vni 6000
  member interface <phy if> service instance <id>
```
Centralized Routing in a L2 VXLAN Fabric
VXLAN Configuration

Layer 3 Gateway

**VLAN CLI Model**

- `vlan 3002`
- `vn-segment 6000`

- `interface nve1`
  - `no shutdown`
  - `source-interface loopback1`
  - `member vni 6000 mcast-group 235.1.1.1`

- `interface vlan 3002`
  - `ip address x.x.x.x`
  - `hsrp 100`
  - `ip address v.v.v.v`

**EFP/VSI CLI Model**

- `Vni 6000`

- `Bridge-domain 100`
  - `member vni 6000`

- `interface nve1`
  - `no shutdown`
  - `source-interface loopback1`
  - `member vni 6000 mcast-group 235.1.1.1`

- `interface bdi 100`
  - `ip address x.x.x.x`
  - `hsrp 100`
  - `ip address v.v.v.v`

**Map VNI to VLAN/BD**

**VXLAN tunnel interface**

**Routed interface**
VTEP Redundancy in a VXLAN Fabric

vPC provides MAC state synchronization and HSRP peering
Redundant VTEPs share anycast VTEP IP address in the underlay

vMAC → Emulated VTEP

L3 Gateway redundancy based on vPC and HSRP (2 nodes)

L2 Gateway redundancy based on vPC (anycast VTEP address)
VXLAN Configuration

Redundant VTEP Anycast Source-interface

**VLAN, VSI or EFP CLI model**
Create BDs in VLAN, VSI or EFP mode

```plaintext
interface loopback1  
ip address <x.x.x.x>  
   ip address <VTEP-anycast> secondary

interface nve1  
   no shutdown  
   source-interface loopback1  
   member vni 6000 mcast-group 235.1.1.1

interface [vlan|bvi] 3002  
ip address x.x.x.x  
hsrp 100  
   ip address v.v.v.v
```

Map VNI to VLAN/BD

Source Interface

VXLAN Tunnel Interface

Routed interface

Configure a vPC per standard vPC guidelines
HW VTEP Redundancy – L2 Gateway

VXLAN Overlay on IP Fabric

VTEP Anycast IP A

VTEP Anycast IP B

VXLAN L2 Gateway

IP A1

IP A2

VXLAN L2 Gateway

IP B1

IP B2

WAN/DCI
HW VTEP Redundancy – L3 Gateway

**L3 VXLAN Overlay**
- Overlay: No HSRP
- Underlay: Independent VTEPs

- **IP A1**
- **IP A2**

- **VXLAN L3 Gateway**
- **WAN/DCI**

- **p2p IP links**

**L2 VXLAN Overlay**
- Overlay: HSRP VIP
- Underlay: VTEP Anycast IP B

- **IP B1**
- **IP B2**

- **VXLAN L3 Gateway**
- **IP over Port Channels**

- **WAN/DCI**

---

Overlay: HSRP VIP
Underlay: VTEP Anycast IP B

- **IP B1**
- **IP B2**

- **VXLAN L3 Gateway**
- **IP over Port Channels**

- **WAN/DCI**
Traditional L2 - centralised L2/L3 boundary

- Always bridge, route only at an aggregation point
- Large amounts of state converge
- Scale problem for large# of L2 segments
- Traditional L2 and L2 overlays

L2/L3 fabric (or overlay)

- Always route (at the leaves), bridge when necessary
- Distribute and disaggregate necessary state
- Optimal scalability
- Enhanced forwarding and L3 overlays
Distributed IP Anycast Gateway

The same “Anycast” SVI IP/MAC is used at all VTEPs/ToRs. A host will always find its SVI anywhere it moves.

SVI IP Address
MAC: 0000.dead.beef
IP: 10.1.1.1

SVI IP Address
MAC: 0000.dead.beef
IP: 10.1.1.1
Distributed IP Anycast Gateway

Detailed View

Consistent Anycast SVI IP / MAC address at all leaves
VLAN-IDs are locally significant
VXLAN Configuration – Distributed Gateway SVI

```plaintext
interface vlan 10
  no shutdown
  vrf member customer1
  ip address 10.10.10.1/24 tag 12345
  fabric forwarding mode anycast-gateway

vrf context customer1
  vni X
  rd auto
  address-family ipv4 unicast
    route-target import auto vni
    route-target export auto vni
```
VXLAN CP Configuration – Distributed Gateway

```
router bgp 65000
  address-family l2vpn evpn
neighbor x.x.x.x
e   remote-as 65000
  address-family l2vpn evpn
vrf customer1
  address-family ipv4 unicast
```

Activate EVPN in iBGP
VXLAN Bridging

802.1Q Tagged Traffic to VNI Mapping

- VLANs are stretched over L2 VNIs
- VLANs (VLAN A) mapped to VNI (VNI A) at each VTEP: VLAN A’ ⇔ VNI A ⇔ VLAN A
- Bridged traffic forwarded over the L2 VNIs
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Same Subnet aka Bridging (ARP)

1. PM1 sends an ARP request for Default Gateway – 10.10.10.1
2. The ARP request is suppressed at TOR and punted to the Supervisor, where MAC and IP is learned and distributed
3. TOR response with Gateway MAC to PM1

Standard behavior of End-Host (virtual or physical) to ARP for the Default Gateway
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Same Subnet aka Bridging (ARP)

4. VM1 sends an ARP request for PM1 – 10.10.10.20

5. The ARP request is suppressed at TOR and punted to the Supervisor, where MAC and IP is learned and distributed

6. Assuming PM1 is known and a valid route does exist in the Unicast RIB, TOR responds to ARP with PM1 MAC as Source MAC. VM1 can build its ARP cache

If there is Unicast RIB miss on TOR, ARP request will be forwarded to all ports except the original sending port (ARP snooping). ARP response will be punted to Supervisor of destination TOR for Unicast RIB population (learn) and subsequently forwarded to source TOR.
Optimizing ARP behavior

Minimizing Flooding in the Fabric with ARP suppression

- IP and MAC addresses host information distributed by control protocol
- NVEs (Leaf Switches) create an ARP cache for remote hosts
- NVEs reply locally to ARP requests for remote hosts
  - Avoid ARP request broadcast flooding

```
switch# sh ip arp suppression-cache detail

Flags: + - Adjacencies synced via CFSoE/vPC peer
  R – Remote Adjacency
  L2 – Leant over I2 interface

Total number of entries: 2
Address         Age         MAC Address   Vlan  Physical Interface Flags
100.1.1.2       00:01:02  0026.980c.1ec2  100  Ethernet2/6
100.1.1.3       00:01:03  0026.980c.1ec3  100  
```

interface nve 1
source-interface loopback 0
member vni 100 mcast-group 239.0.0.1
member vni 1000
ten-host-reachability control protocol bgp
  suppress-arp
  ingress-replication control protocol bgp
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Same Subnet aka Bridging (Data Packet)

7. VM1 generates a data packet with PM1_MAC as destination MAC
8. TOR receives the packet and performs Layer-2 lookup for the destination
9. TOR adds VXLAN-Header information (Destination VTEP, VNI, etc) and forwards the packet across the Layer-3 fabric, picking one of the equal cost paths available via the multiple Spines
10. The destination TOR receives the packet, strips off the VXLAN header and performs lookup and forwarding toward PM1

In case of VM1 is not known to PM1, PM1 would ARP for VM1. Destination TOR would Proxy for VM1. No Silent-Host discovery problem.
A common VNI (VNI X) is provisioned amongst the different VTEPs to carry routed traffic.

- Routed traffic between VTEPs will be encapsulated in VNI X.
- Standard longest prefix match routing takes place:
  - Host routes for all known remote hosts are installed at every VTEP ➔ Forward over VNI X.
  - Local hosts are covered by directly connected prefix, a host route will not be present.
Distributed IP Anycast Gateway

Packet Flows – unknown Host (H4) in remote Subnet

- H1 → H4
  - Routed via SVI B (VTEP1) to VLAN A (VTEP1) → Bridged over VNI A (as unknown unicast flood)

- H4 → H1
  - Routed via SVI A (VTEP2) → VNI X → SVI B (VTEP1) → VLAN B’ (as H1 is known based on response)

- Standard longest prefix match routing: As long as H4 is not learnt by VTEP1, the only path to H4 is the locally connected subnet (VLAN A’)

VTEP1

VLAN A'
SVI A'

VTEP2

VLAN A
SVI A

VNI X

VLAN A
SVI A

VLAN B'
SVI B

H2
H1
H4
Distributed IP Anycast Gateway

Packet Flows – unknown Host (H1) in remote Subnet

- H4 → H1
  - Routed via SVI A (VTEP2) to VLAN B (VTEP1) → Routed over VNI X (as destination Subnet known)

- H1 → H4
  - Routed via SVI B (VTEP1) → VNI X → SVI A (VTEP2) → VLAN A’ (as H1 is known based on response)
VXLAN CP Configuration – Advertisement of prefix routes

```
router bgp 65000
  address-family l2vpn evpn
  neighbor x.x.x.x
    remote-as 65000
    address-family l2vpn evpn

vrf customer1
  address-family ipv4 unicast
  redistribute hmm route-map foo
  redistribute direct route-map bar
```

- **Inject host IP addresses for host prefixes in BGP-EVPN**
- **Use a route-map to restrict the redistribution to host prefixes**
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Different Subnet aka Routing (ARP)

1. VM1 sends ARP request for Default Gateway – 10.10.10.1

2. The ARP request will be received at TOR and punted to the Supervisor, where MAC and IP is learned and distributed.

3. TOR acts as regular Default Gateway and sends ARP response with GW_MAC to VM1.

---

**VM1 ARP Cache**

<table>
<thead>
<tr>
<th>IP</th>
<th>MAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.20.20.20</td>
<td>0000.dead.beef</td>
</tr>
</tbody>
</table>

**SVI IP Address (VRF Blue)**

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000.dead.beef</td>
<td>20.20.20.1</td>
</tr>
</tbody>
</table>

**VNI**

<table>
<thead>
<tr>
<th>L2 VNI</th>
<th>L3 VNI</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>50000</td>
</tr>
</tbody>
</table>
Distributed IP Anycast Gateway

Packet-Walk – IP Forwarding within the Different Subnet aka Routing (Data Packet)

4. VM1 generates a data packet destined to PM2 IP (20.20.20.20) with GW_MAC as destination MAC
5. TOR receives the packet and performs Layer-3 lookup for the destination (known)
6. TOR adds VXLAN-Header information (Destination VTEP, VNI, etc) and forwards the packet across the Layer-3 fabric, picking one of the equal cost paths available via the multiple Spines
7. The destination TOR receives the packet, strips off the VXLAN header and performs lookup and forwarding toward PM2
SVI Resiliency with anycast MAC

An active GWY at every (redundant) Leaf

- Anycast MAC forwarding on local BD
- Requires a MC-Port Channel facing south to avoid MAC Flapping
- No use of FHRP
- Only available with an Overlay Control Plane
SVI/VNI/VLAN Scoping and Provisioning

Orchestration leads to scale optimization

All VNIs/SVIs everywhere
- Umbrella catch-all provisioning
- Full ARP state on all Leaf Nodes
- Can be manually provisioned up-front
- Open to L2 Flooding everywhere

VNIs/SVIs scoped as hosts attach
- Provision on host attach/policy
- ARP state only for local subnets
- Requires orchestration
- L2 Flooding is scoped

Cisco Live!
Integration with Orchestrators and Host Attachment

• Compute Controller (e.g. vCloud Director) integrated overlay provisioning
  • Integrates physical and virtual end-points
  • Topology discovery with triangulation
Automating VNIs/SVIs as hosts attach

1. Virtual or Physical Machine comes online

2. New Trigger Event on TOR
   - New MAC Learn with VLAN
   - VDP* (N1kv and OVS) with VNI
   - VMTracker with VLAN ID + Port-Group
   - LLDP¹ (Bare Metal) with NIC MAC
   - CLI with VNI or VLAN

3. TOR initiates LDAP query for respective identifier (e.g. VLAN, VNI or MAC)

4. Respective Configuration will be Downloaded (Pull) and instantiated

¹Roadmap
*802.1Qbg - VSI Discovery and Configuration Protocol (VDP)
Multi-Data Center Connectivity

LAN Extensions and IP mobility

Ethernet extensions between independent fabrics
IP traffic is forwarded via the optimal path (no hair-pinning)
Multi-Data Center Connectivity

LAN Extensions

Domain Boundary:
Failure and Event Containment
Clear Administrative Delineation
Scoped Control Plane

<table>
<thead>
<tr>
<th>MAC</th>
<th>IP</th>
<th>VTEP</th>
<th>L2 VNI</th>
<th>L3 VNI</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM1_MAC</td>
<td>10.10.10.10</td>
<td>STOR_LO</td>
<td>10000</td>
<td>20000</td>
</tr>
<tr>
<td>PM2_MAC</td>
<td>20.20.20.20</td>
<td>DTOR_LO</td>
<td>30000</td>
<td>20000</td>
</tr>
</tbody>
</table>
Multi-Data Center Connectivity

L2 Handoff

Terminate VXLAN segments at DC-edge

- Per VNI load balancing
- Optimized failure containment

OTV/VPLS/EVPN

CRS-1

Core

VXLAN X

Access

VXLAN Y
OTV/VPLS & VXLAN

L2 Handoff

Join-interface

IP/MPLS Core

OTV/VPLS

VLAN

VXLAN

OTV/VPLS

VLAN

L2 internal interface

Peer link

L2 internal interface

VXLAN L2 Gateway

L3 Fabric

VXLAN L2 Gateway

VXLAN

OTV

OTV
Multi-Data Center Connectivity

IP Mobility for optimized routing

LISP/BGP Signalling:
Relay mobility state between sites

Direct Path Forwarding

Host routes

DC1

DC2

Moving Hosts

Fabric Mobile Host Detection

Host routes
MPLS IP-VPN/LISP & VXLAN

L3 Handoff

MPLS/IP Core

N7K

L3 Fabric

N3K, N5600, N6K-X, N7K, N9K

VXLAN L3 Gateway

VXLAN L3 Gateway

VXLAN L3 Gateway

VXLAN L3 Gateway

VXLAN L3 Gateway

VXLAN L3 Gateway

VXLAN L3 Gateway
End-to-end IP mobility with LISP & BGP

L3-VXLAN & LISP IP Mobility

LISP Mobility:
- LISP encapsulation from client sites
- No host routing in the IP core
- Direct Path Forwarding

BGP event = Move signal to LISP

Data Plane

Control Plane

MAC | IP | VNI | Next Hop | Enca p | Seq
---|----|-----|----------|--------|-----
1   | 1  | 5000| XTRs     | VXLAN  | 1   

Data Center 1

Data Center 2

BGP host route

BGP host route

LISP Host Mobility

Redistribution BGP → LISP

Fabric Mobile Host Detection

Host 1 VLAN 10

VNI 5000

Access

Aggregation

Core

Core
Underlay Deployment Considerations
Multicast Enabled Underlay

Host Overlay to Hybrid Overlay

- Host Overlay VTEPs join multicast groups as hosts using IGMP reports

- Host overlays will work over an L2 underlay, ensure IGMP snooping is in place to scope the reach of multicast

- A multicast enabled L3 underlay is the better option as it enables a hybrid overlay (host and network VTEPs)
  - Ensure that the first hop router for the host in the underlay is configured to service the IGMP reports from the host VTEP
Multicast Enabled Underlay

Network Overlay

- May use PIM-ASM or PIM-BiDir (Different hardware has different capabilities)

<table>
<thead>
<tr>
<th>Mcast mode</th>
<th>N1KV</th>
<th>Nexus 7K with F3 LC</th>
<th>Nexus 3K</th>
<th>Nexus 5K/6K</th>
<th>Nexus 9K Standalone</th>
<th>CSR 1000V ASR1K</th>
<th>ASR9K</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGMP v2/v3</td>
<td>PIM-ASM &amp; Bidir-PIM</td>
<td>PIM-ASM (Bidir – Future)</td>
<td>Bidir-PIM</td>
<td>PIM-ASM (Bidir – Future)</td>
<td>Bidir-PIM (ASM – Future)</td>
<td>PIM-ASM &amp; Bidir-PIM</td>
<td></td>
</tr>
</tbody>
</table>

- Spine and Aggregation switches make good RP locations in clos and traditional topologies respectively

- Reserve a range of multicast channels to service the overlay and optimize for diverse VNIs

- In clos topologies with lean spine, using multiple RPs across the multiple spines and mapping different VNIs to different RPs will provide a simple load balancing measure

- Design a multicast underlay for a network overlay, host VTEPs will simply leverage this network.
Multi-Pathing and Entropy

- Symmetric Underlay Network topologies facilitate ECMP routing:
  - Multi-path load balancing
  - Fast Re-convergence on link Failures
- Polarization: Encapsulated flows appear as a single flow which hashes to a single path
- Entropy in the encapsulation header to depolarize tunnels
  - Variable UDP source port in VXLAN outer header
  - Underlay must support ECMP hashing on L4 port numbers
Unicast in the Underlay – Interfaces (1)

How should my Underlay look like

- Know your IP addressing and IP scale requirements
  - Use 1 prefix for all Underlay Links and Loopbacks

- Routed ports/interfaces
  - Interfaces between Spine and Leaf are in routed mode (no switchport)
  - For each Leaf / Spine connection, at least a /31 is required

- Local to Remote VTEP (Loopback) adjacency requires routed interface in-between
  - Exception: connection from SW VTEP

2 Spine * 3 Leaf = 6 Links
6 Link * 2 (/31) + 3 VTEP
15 IP Addresses required ➔ (/27)
Unicast in the Underlay – Routing Protocol (1)

How should my Underlay look like

- Routing-Protocol of choice (many flavors available)
- OSPF – watch your type
  - p2p preferred (only LSA type-1)
    - suits well for routed interfaces/ports (optimal from a LSA database perspective)
    - Full SPF calculation on link-change
    - broadcast (LSA type-1 & 2 + BR/DR election)
      - additional election and database overhead
- IS-IS
  - independent of IP (CLNS) and well suited for routed interfaces/ports
  - not everyone is familiar with it
Unicast in the Underlay – Routing Protocol (2)

How should my Underlay look like

• **eBGP**
  • neighbor is interface IP when using routed interfaces/ports approach
  • Use of loopbacks would require additional routing

• **The Routing-Protocol Combo**
  • IGP for underlay topology & reachability (e.g. IS-IS, OSPF)
  • iBGP for VTEP (loopback) reachability
  • iBGP route-reflector for simplification and scale
Folded Clos Topology

Providing Topology Symmetry

- Fully Symmetric, BW rich topology, Optimized for East-West traffic
- Lean Spine does not do any VXLAN termination/gateway
- Access to other networks through border leaf block
• Fully Symmetric, BW rich topology, Optimized for North-South traffic
• VXLAN termination/gateway @ Access and Core (or Aggregation)
• Access to other networks through Core
Access/Aggregation/Core Wide BW Topology

WAN Handoff Resiliency Models

L2 Handoff
- L2 GWY Resiliency in pairs of VTEPs (vPC based)
- VNI Based Load Balancing

L3 Handoff (w/Distributed GWY)
- ECMP across all L3 GWYs
- Can combine with L2 VNI based balancing and Resilient VTEPs
Instrumentation and Overlay Awareness

- Infrastructure awareness of encapsulated traffic:
  - Outer/Encapsulation header
  - Overlay shim header
  - Internal/Payload header
  - Payload

- Overlay aware Switching & Routing infrastructure:
  - ACLs, QoS, Netflow

- Network Analysis Module (NAM) inspects encapsulated traffic
Over-speed, Encapsulation & Effective Throughput

- Encapsulation adds bits to the traffic being sent
- When receiving traffic at full line rate, the encapsulated traffic will exceed the line-rate BW of the egress interface
  - Packet drops
  - Diminished effective throughput
- The uplink BW should be greater than the downlink BW to avoid congestion by encapsulation
  - This is naturally done in the network

\[
\begin{align*}
1500\text{bytes/packet (10Gbps)} & \Rightarrow 1542 \text{ bytes/packet (10.1 Gbps)} \\
64\text{bytes/packet (10Gbps)} & \Rightarrow 106 \text{ bytes/packet (10.3 Gbps)}
\end{align*}
\]
Summary and Conclusion
Summary recommendations & takeaways

• Optimize the location of L2 and L3 GWYs to optimize routing and minimize failure exposure

• Leverage L3 VXLAN services enabled by control protocols as the main service and L2 extensions as the exception

• Design the underlay with the VXLAN overlay in mind

• A combination of pull protocols and push protocols may render optimal scale and resiliency

• Design the network hierarchically: both the underlay as well as the overlay

• L3 Gateways are key to a sound overlay design

• Link the provisioning of the overlay and scoping of VNIs to the host orchestration system for optimal scale
Complete Your Online Session Evaluation

• Give us your feedback to be entered into a Daily Survey Drawing. A daily winner will receive a $750 Amazon gift card.

• Complete your session surveys though the Cisco Live mobile app or your computer on Cisco Live Connect.

Don’t forget: Cisco Live sessions will be available for viewing on-demand after the event at CiscoLive.com/Online
Continue Your Education

• Demos in the Cisco campus
• Walk-in Self-Paced Labs
• Table Topics
• Meet the Engineer 1:1 meetings
• Related sessions
Thank you
TOMORROW starts here.
Design Cisco Education Offerings

<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
<th>Cisco Certification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designing Cisco Network Service Architectures (ARCH)</td>
<td>Provides learner with the ability to perform conceptual, intermediate, and detailed design of a network infrastructure that supports desired capacity, performance, availability required for converged Enterprise network services and applications.</td>
<td>CCDP® (Design Professional)</td>
</tr>
<tr>
<td>Designing for Cisco Internetwork Solutions (DESGN)</td>
<td>Instructor led training focused on fundamental design methodologies used to determine requirements for network performance, security, voice, and wireless solutions. Prepares candidates for the CCDA certification exam.</td>
<td>CCDA® (Design Associate)</td>
</tr>
</tbody>
</table>

For more details, please visit: [http://learningnetwork.cisco.com](http://learningnetwork.cisco.com)
Questions? Visit the Learning@Cisco Booth or contact ask-edu-pm-dcv@cisco.com
## Data Center / Virtualization Cisco Education Offerings

<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
<th>Cisco Certification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cisco Data Center CCIE Unified Fabric Workshop (DCXUF);</td>
<td>Prepare for your CCIE Data Center practical exam with hands on lab exercises running on a dedicated comprehensive topology</td>
<td>CCIE® Data Center</td>
</tr>
<tr>
<td>Cisco Data Center CCIE Unified Computing Workshop (DCXUC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementing Cisco Data Center Unified Fabric (DCUF1);</td>
<td>Obtain the skills to deploy complex virtualized Data Center Fabric and Computing environments with Nexus and Cisco UCS.</td>
<td>CCNP® Data Center</td>
</tr>
<tr>
<td>Implementing Cisco Data Center Unified Computing (DCUCI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introducing Cisco Data Center Networking (DCICN); Introducing Cisco</td>
<td>Learn basic data center technologies and how to build a data center infrastructure.</td>
<td>CCNA® Data Center</td>
</tr>
<tr>
<td>Data Center Technologies (DCICT)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product Training Portfolio: DCAC9k, DCINX9k, DCMS, DCUCS, DCNX1K,</td>
<td>Get a deep understanding of the Cisco data center product line including the Cisco Nexus9K in ACI and NexusOS modes</td>
<td></td>
</tr>
<tr>
<td>DCNX5K, DCNX7K</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For more details, please visit: [http://learningnetwork.cisco.com](http://learningnetwork.cisco.com)

Questions? Visit the Learning@Cisco Booth or contact ask-edu-pm-dcv@cisco.com
## Cloud Cisco Education Offerings

<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
<th>Cisco Certification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designing the FlexPod Solution (FPDESIGN); Implementing and Administering the FlexPod Solution (FPIMPADM)</td>
<td>Learn how to design, implement and administer FlexPod solutions</td>
<td>FlexPod Design Specialist; FlexPod Implementation &amp; Administration Specialist</td>
</tr>
<tr>
<td>UCS Director (UCSDF)</td>
<td>Learn how to manage physical and virtual infrastructure using orchestration and automation functions of UCS Director.</td>
<td></td>
</tr>
<tr>
<td>Cisco Prime Service Catalog</td>
<td>Learn how to deliver data center, workplace, and application services in an on-demand, automated, and repeatable method.</td>
<td></td>
</tr>
<tr>
<td>Cisco Intercloud Fabric</td>
<td>Learn how to implement end-to-end hybrid clouds with Intercloud Fabric for Business and Intercloud Fabric for Providers.</td>
<td></td>
</tr>
<tr>
<td>Cisco Intelligent Automation for Cloud</td>
<td>Learn how to implement and manage cloud deployments with Cisco Intelligent Automation for Cloud</td>
<td></td>
</tr>
</tbody>
</table>

For more details, please visit: [http://learningnetwork.cisco.com](http://learningnetwork.cisco.com)
Questions? Visit the Learning@Cisco Booth or contact ask-edu-pm-dcv@cisco.com