Agenda

• VXLAN Recap
• Real World VXLAN Deployments
• Migration Strategy
Data Center “Fabric” Journey
Modern DC Fabric

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
- Scale – Reduce core state
  - Distribute and partition state to network edge
- Flexibility/Programmability
  - Reduced number of touch points
VXLAN Benefits

• Flexible placement of any workload in any rack throughout and between data centers
• Decoupling between physical and virtual networks
• Large Layer 2 network to provide work load mobility
• Centralized Management, provisioning, and automation, from a controller
• Scale, performance, agility and stream lined operations
• Better utilization of available network paths in the underlying infrastructure
# Cisco VXLAN Portfolio

## Cisco VXLAN Solutions

<table>
<thead>
<tr>
<th>Scale</th>
<th>Secure Multi-tenancy</th>
<th>Workload Mobility</th>
<th>Workload Anywhere</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR1000 CSR1000</td>
<td>Nexus 1000</td>
<td>Nexus 3100</td>
<td>Nexus 5600</td>
</tr>
<tr>
<td>L2 Gateway</td>
<td>L3 Gateway</td>
<td>BGP EVPN Control Plane</td>
<td>Anycast Gateway</td>
</tr>
</tbody>
</table>

### Cisco VXLAN Solutions

- **ASR1000 CSR1000**
- **Nexus 1000**
- **Nexus 2000**
- **Nexus 3100**
- **Nexus 5600**
- **Nexus 7000**
- **Nexus 9000**
- **ASR9000**

### Features

- **Scale**
- **Secure Multi-tenancy**
- **Workload Mobility**
- **Workload Anywhere**

**BRKDCN-2020 © 2016 Cisco and/or its affiliates. All rights reserved. Cisco Public**
VXLAN Frame Format

MAC-in-IP Encapsulation

Original Layer-2 Frame

VXLAN Header
- VXLAN Flags
- VNI (24 Bytes)
- Reserved

UDP Header
- Source Port
- UDP Length
- Checksum 0x0000

Outer IP Header
- IP Header Misc. Data
- Protocol 0x11 (UDP)
- Header Checksum
- Source IP
- Dest. IP

Outer MAC Header
- VLAN Type 0x8100
- VLAN ID Tag
- Ether Type 0x0800

Src VTEP MAC Address

Dest. MAC Address

Src. MAC Address

14 Bytes (4 Bytes Optional)

8 Bytes

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

Src and Dst addresses of the VTEPs

VNI

Reserved

VXLAN Flags

RRRRIRRR

8 Bytes

Allows for 16M possible Segments

UDP 4789

VNI

Reserved

Reserved

8 Bytes

50 (54) Bytes of Overhead

50 (54) Bytes of Overhead

Underlay

Overlay
VXLAN Overview (1)
VXLAN Overview (2)

VTEP – VXLAN Tunnel End-Point
VNI/VNID – VXLAN Network Identifier
Centralized Gateway (FHRP)

**VXLAN Routing**

- Centralized Routing in a Layer-2 VXLAN Network
  - Routing between VNI (Different Subnet)
  - Bridging within VNI (Same Subnet)
- Inter-VXLAN Routing at Core/Aggregation Layer
- vPC provides MAC state synchronization and HSRP peering
  - Redundant VTEPs share Anycast VTEP IP address in the Underlay
- Bottleneck for throughput
“Traditional” VXLAN deployments

**Head End Replication**
- Overlay Control-Plane provides dynamic VTEP discovery
- Head-End Replication enables Unicast-only mode (aka ingress Replication)
- Only scales for VERY small deployments, any scaled deployment will quickly encounter ingress replication related limitations even with minimal amount of BUM traffic

**Flood & Learn**
- Uses “Flood and learn” technique
- Relies on multicast underneath for BUM traffic
- Still has traffic hair pinning issues for default gateway placements

*Multicast Independence requires the usage of the Overlay Control-Plane or static configuration*
Key Takeaway

• "Traditional" VXLAN deployments suffer from performance bottlenecks due to flood and learn and/or head end replication for B.U.M traffic

• Default gateway placement is still a challenge in large fabrics

• Salt/Pepper load-balancing of gateways still creates bandwidth hotspots

• In ‘real world’ deployments traditional VXLAN flood/learn with or without head-end replication has limited applicability simply does not scale for 100’s of switches
If only…..

• There was a protocol that could scale to millions of MAC addresses
• Multiple VRF’s
• Provide a robust communication mechanism
• Provide a framework for authenticated peer-discovery
• Granular filtering
VXLAN Control Plane – Available Today!

- Multi-Protocol BGP (MP-BGP) based Control-Plane using EVPN NLRI (Network Layer Reachability Information)
- Make Forwarding decisions at VTEPs for Layer-2 (MAC) and Layer-3 (IP); Integrated Route/Bridge (IRB)
- Reduce Flooding
- Reduce impact of ARP on the Network
- Standards Based

Protocol Learning

- Workload MAC and IP Addresses learnt by VXLAN Edge Devices (NVEs)
- Advertises Layer-2 and Layer-3 Address-to-VTEP Association (Overlay Control-Plane)
- Flood Prevention
- Optimized ARP forwarding
Host and Subnet Route Distribution

VXLAN/EVPN

- Host Route Distribution decoupled from the Underlay protocol
- Use MultiProtocol-BGP (MP-BGP) on the Leaf nodes to distribute internal Host/Subnet Routes and external reachability information
- Route-Reflectors deployed for scaling purposes
EVPN – Ethernet VPN

VXLAN Evolution

- EVPN over NVO Tunnels (VXLAN, NVGRE, MPLSoE) for Data Center Fabric encapsulations
- Provides Layer-2 and Layer-3 Overlays over simple IP Networks
Host and Subnet Route Distribution

VXLAN/EVPN

- Host Route Distribution decoupled from the Underlay protocol
- Use MultiProtocol-BGP (MP-BGP) on the Leaf nodes to distribute internal Host/Subnet Routes and external reachability information
- Route-Reflectors deployed for scaling purposes

BGP Route-Reflector

----- iBGP Adjacency
Gateway Functions in VXLAN

VXLAN Routing

Distributed Gateway
- Route or Bridge at Leaf
- Distributed Gateway (Anycast) for Routing
- Disaggregate state by scale out
- Optimal Scalability
- Requires VXLAN/EVPN!
Distributed IP Anycast Gateway*

VXLAN/EVPN

- Distributed Routing with IP Anycast Gateway (Integrated Route/Bridge IRB)
  - Routing between VNI (Different Subnet)
  - Bridging within VNI (Same Subnet)
- Inter-VXLAN Routing Leaf/Access Layer
  - All Leafs share gateway IP and MAC for a Subnet (No HSRP)
  - A Host will always find its Gateway directly attached anywhere it moves

*Requires EVPN Control-Plane.
## Symmetric vs Asymmetric – A quick word

The RFC defines two modes of routing between VXLAN overlays:

<table>
<thead>
<tr>
<th>Asymmetric</th>
<th>Symmetric</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Ingress VTEP performs route + bridge</td>
<td>• Both ingress and egress VTEP can forward at L2 or L3</td>
</tr>
<tr>
<td>• Egress VTEP only performs bridge</td>
<td>• Only VNI’s of locally attached VTEP’s need to be configured on the leaf.</td>
</tr>
<tr>
<td>• Complex configuration</td>
<td>• Simplifies configuration</td>
</tr>
<tr>
<td>• Need to learn “Everything” ”everywhere”</td>
<td>• Better scale in terms of total VNI’s</td>
</tr>
<tr>
<td></td>
<td>• Cisco supports ONLY Symmetric IRB</td>
</tr>
</tbody>
</table>
Key Takeaway
MP-BGP EVPN is what you should use

• Control plane learning for end host Layer 2 and Layer 3 reachability information.
• Ability to build a more robust and scalable VXLAN overlay network
• Supports multi-tenancy, provides integrated routing and bridging
• Minimizes network flooding through protocol-driven host MAC/IP route distribution
• ARP suppression to minimize unnecessary flooding
• Peer discovery and authentication to improve security, optimal east-west and north-south traffic forwarding
• Cisco supports only symmetric IRB mode
• When buying a next-gen switching platform ensure that it can perform VXLAN routing in hardware (Note: Merchant only Trident2 or Tomahawk based ToR switches cannot “natively” route VXLAN, ONLY Trident2+ can)
• When buying a next-gen switching platform ensure that it can support MP-BGP w EVPN
Agenda

- VXLAN Recap and WHY MP-BGP is a requirement
- Real World VXLAN Deployments
- Migration Strategy
Host Overlays
Host overlay design considerations

- Host overlays run between VTEPs and empower server teams to delivery virtual network services without needing to involve the network team.

- Typically results in a sub-optimal networking solution.

- Results in CPU impact as most off the shelf NIC’s and hypervisors do not provide HW offload for VXLAN.

- Extra effort due to lack of correlation between IP underlay and overlay, network team has no idea where workloads live.

- Typically needs a gateway, either software or hardware which must be integrated with the physical network.

- Please consider, licensing costs, compute overhead, performance penalties gateway infrastructure requirements and impact to Opex.
Solutions

- Use VMTracker with auto-config to enable automated provisioning of network resources in a vSphere environment.

- VMTracker will communicate with vCenter to:
  - ESX host to VM port-group mappings
  - Physical port attachments
  - VDS assignment to VM

- VMTracker will provision the following
  - VLAN provisioning
  - VNI allocation
  - L3 anycast GW provisioning
  - VRF provisioning

- You get HW based forwarding, full visibility, and network automation without any of the drawbacks in the previous slide
That being said ..

- There may be scenarios where host based overlays may solve some challenges such as cases where
  - Support for a mix of software and hardware VTEPs
  - Support multi-vendor Fabrics
  - Overlay and underlay are operated by different teams (OpenStack)

- Cisco VTS (Virtual Topology System) is one possible solution. The VTC (Virtual Topology Controller) is a single point of management for configuration/management and operation of VXLAN EVPN.

- The management layer supports orchestration with hypervisors like vSphere or OpenStack

- Control-plane using IOS-XR w MP-BGP EVPN and SW VTEP leveraging the VTF (Virtual Topology Forwarder)
Key Takeaways – Host Overlays

- Document scale requirements in granular detail
- Verify supported scale using the verified scalability guides
- Verify you are running the latest recommended version of code
- Careful with scaling host overlays. Consider not just network automation/virtualization, but performance SLA’s, cost of licensing, edge gateway hardware (more servers).
- Cisco VM-Tracker and/or Cisco VTS provide a comprehensive solution for host based overlay designs
- Specific host overlay use cases such as distributed firewall at hypervisor level, micro segmentation can be delivered with much lower TCO and orders of magnitude improvement in ROI via Cisco ACI
Key considerations for successful VXLAN EVPN design

Underlay
- Importance of Underlay
- MTU considerations
- Overlapping vs non overlapping underlay IP addressing
- Inband vs Out-Of-Band management
- IGP Selection and fast failover criteria
- Multicast protocol selection
- Multicast RP placement and HA strategy
- Multicast OIF Scaling on Spine

Overlay
- Head End replication
- RD and RT numbering strategy
- VRF for L3 VNI assignment strategy
- Route reflector form factor
- Route reflector placement
- BUM Traffic handling
- Host connectivity strategy
- QoS requirements
- Layer 3 External connectivity to legacy
- Layer 2 Extension including Loop Prevention
- Multicast in the overlay
- Integration with Firewalls and Load-Balancers
Before you begin

• IMPORTANT: Always refer to the verified scalability guide and the recommended software release on CCO for the latest numbers. This is a number given at a point in time based on Cisco Engineering’s QA efforts

• Different for different platforms (Nexus 9000 vs Nexus 7000 vs Nexus 5600) and needs careful consideration and planning during design phase to ensure you don’t run into issues
General Underlay Considerations

- Similar to MPLS VPN / in Provider networks, the Underlay is your IGP and on top the BGP process with the VRFs are running

- The **Underlay** network is the **heart of your network**!
  - Your Overlay can’t be more stable or have faster convergence then your Underlay
  - A failure in the Underlay, affect all VRFs of your Overlay

- Consider to **isolate** your **Underlay from outside**
  - Authentication / network management can run via mgmt0 or separate VRF
  - Also automation can run via mgmt0 or inside of own VRF
  - Which other reason you would have then to extend Underlay?
General Underlay Considerations

• Similar to MPLS VPN / in Provider networks, the Underlay is your IGP and on top the BGP process with the VRFs are running

• The **Underlay** network is the **heart of your network!**
  • Your Overlay can’t be more stable or have faster convergence then your Underlay
  • A failure in the Underlay, affect all VRFs of your Overlay

• Consider to **isolate** your **Underlay from outside**
  • Authentication / network management can run via mgmt0 or separate VRF
  • Also automation can run via mgmt0 or inside of own VRF
  • Which other reason you would have then to extend Underlay?
Underlay Routing Protocol

- **BGP will work but ...**
  - It’s a Distance Vector Routing Protocol
    - Only knows about Autonomous Systems
    - No idea about Link Speed or Path Cost (only about AS)
  - A lot of Peering's – **Use Subnet based peering**
    - Loopback to Loopback will only work if Reachability is there
    - Adding a 2nd Routing Protocol for Loopback Reachability?

- When using BGP in the Underlay and BGP in the Overlay (EVPN), do we still have Underlay / Overlay separation?

- Can I use eBGP in the Underlay and iBGP in the Overlay?
IGP and Fast-Failover Criteria

• Most operations teams are not ready for this as the line between overlay and underlay is blurred

• IGP’s also simplify multi-pathing and offer separation of the underlay and overlay control plane

• Most deployments rely on IGP for underlay and MP-BGP EVPN for overlay.

• Typical IGP best practices apply to VXLAN designs
  • Use point-to-point as the network-type (OSPF)
  • Use BFD (ideally offloaded to hardware) for fast failover and not aggressive timers
  • Tie IGP + BGP + multicast into the same BFD instance for best convergence
Routing Mode on Nexus 9000

• Considerations for Server Leaf
  • The assumption is the Server Leaf will have mainly host routes in his FIB
  • Recommendation is keep the default max-mode host
  • Routes from outside of the Fabric (LPM routes) probably been minimal

• Considerations for Spine
  • The Spine will learn in Control Plane all routes, but not adding to FIB / RIB
  • The Spine will learn in his FIB / RIB only the Underlay routes, expectation is this will be relatively small, similar to a P-node in MPLS
  • Recommendation is keep also here the default max-mode host
  • There might be corner cases like Underlay is exposed to outside with existing large RIB
MTU

- VXLAN adds a header (54 byte) to every packet
- Ensure Jumbo MTU is enabled on all links (system QoS enabled globally)
- If enabling jumbo MTU is not possible ensure you have > 1500bytes
- For SVI’s and “VXLAN payload” do not exceed 9100 bytes to be on the safe side
MTU

• A change of MTU in Underlay could be disruptive and could impact also all Overlay VRFs / tenants. Get it right the first time.

• Underlay should use max MTU possible from start
  • Nexus 9000 and Nexus 7000 can use MTU 9216 Byte for Underlay
  • Nexus 5600 can use max MTU 9192 Byte for Underlay (due to internal header)

• Overlay MTU require lower MTU due to VXLAN overhead
  • VXLAN adding header of up to 54 bytes
  • For Overlay, consider to use fixed MTU over all SVI and uplinks to 9100 or 9000 Byte to have common value and took into considerations to space for additional overhead
  • Server NICs usually can’t go higher than 9000 Byte
Loopback Interface

- **Usage of Router ID Loopback IP**
  - Used to build Underlay IGP neighborship
  - Used to build MP-BGP neighborship inside Fabric
  - Seen in VRF BGP entries e.g. as “Imported from”

- **Usage of VTEP Loopback IP**
  - If vPC is used, only secondary IP is used
  - If no vPC is used, only primary IP is used
  - Used as next hop in BGP updates for redistributed networks (type 5)
  - Used as host and L2route table next-hop as well in MAC table (type 2)

- Exception is enhancement “advertise-pip” explained later
IP addressing and In-band vs Out-of-Band

• Typical VXLAN design needs
  • A loopback per Leaf/Spine for MP-BGP/IGP router ID
  • A loopback per Leaf for VXLAN decapsulation
  • A /31 per link from Leaf to Spine
  • An IP address per VPC leaf pair

• Pre-Reserve IP address range with growth in mind

• Use ‘ip unnumbered loopback0’ if supported in software to further optimize Leaf<>Spine IP addressing needs

• If using inband for management, this range should not overlap with the existing network. Hard to re-IP this!

• If using inband for management, Spine needs to support VTEP (VXLAN encap/decap) functionality
Loopback Interface

• For **Underlay Loopbacks**, the **IP address design is most important!**
  • This you will see in each “show ip route …."
  • Is the seen IP used for VTEP or for Router-ID
  • Schema should be able to differentiate by Loopback IP address between Spine, Server Leafs and Border Leafs
  • You could even consider to put the “POD” identity into the Loopback IP

• Beside your “Fabric Uplink” and Underlay Loopback, probably you need one SVI local and between vPC peers
  • More on this later
  • Will be also in VRF default
Loopback Interface

- **Recommendation** is to use one Loopback for Router ID, and **different Loopback for VTEP especially for vPC Leafs**
  - Idea is that **different Loopback for Router ID** will enable **learning** of the **prefixes in the Fabric** even VTEP Loopback is not up
  - VXLAN hold down the VTEP Loopback at boot up for certain time
  - In certain scenarios with vPC the VTEP Loopback will be brought down
  - The VTEP IP will always used as next hop for advertisements
  - Consider this schema for all leaves to have a consistent environment
Store and Forward

- The **Nexus 5600** need to run in **store-and-forward**
  - Default mode is cut-through and need to change as the header must be rewritten, as well as the ingress frame encapsulated
  - This change require a reboot on both platforms

- The **Nexus 9300** does not need store and forward
  - According feedback from engineering for VXLAN the default cut-through should been used

- How do this on **Nexus 5600**
  - Command is “hardware ethernet store-and-fwd-switching”
  - No verification command today (being tracked)
  - With PoaP the reload will be triggered automatically
Multicast

• Enable PIM on the underlay interfaces (unless using head-end replication)

• Do NOT tune PIM timers, rely on BFD for fast failover assuming platform supports PIM tying into BFD

• RP’s can be placed on the Spine

• PIM ASM or PIM SSM can be used. ASM provides higher validated scale in most cases.

• Either MSDP or PIM anycast can be used for RP high availability.
Multicast Routing in Underlay

- BUM traffic needs to be forwarded inside of the Fabric via this path
  - BUM = Broadcast, Unknown Unicast and Multicast

- You can run Fabric Underlay without multicast routing, called ingress replication and will do “head end replication”
  - Recommendation is to use this for small environments only
  - E.g. for static ingress replication recommended not more then 16 VTEPs

- For using multicast replication for BUM, the multicast routing in the Underlay Fabric will be only local
  - The intention is NOT to expand this to outside
  - It will be independent from any multicast routing outside in the Core
Multicast Routing in Underlay

- Natural choice for the Multicast Rendezvous Point (MC RP) for ASM as well BiDir are the Spines
  - The MC RP should be redundant … it didn’t hurt if there are more, e.g. you having 4 Spines and with this 4 MC RP

- Interoperability of Nexus 9300, Nexus 7000 and Nexus 5600 as Leaf for multicast routing in Underlay
  - Nexus 7000 to Nexus 9300 – PIM ASM
  - Nexus 7000 to Nexus 5600 – PIM BiDir
  - Nexus 9300 to Nexus 5600 – Not Today
  - Next slides more details
# Multicast Routing in Underlay Interoperability

<table>
<thead>
<tr>
<th></th>
<th>Nexus 9000</th>
<th>Nexus 7x00</th>
<th>Nexus 5600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer-2 (Ingress Repl.)</td>
<td>✔</td>
<td>✖</td>
<td>✖</td>
</tr>
<tr>
<td>Layer-2 (PIM ASM)</td>
<td>✔</td>
<td>✔</td>
<td>✖</td>
</tr>
<tr>
<td>Layer-2 (PIM BiDir)</td>
<td>✖</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Layer-3</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

What does this mean? .. Hold on..

---

1 Hardware capable
PIM ASM General

- PIM ASM (any source multicast routing aka PIM sparse)
  - Stream initialize via the MC RP … the “shared tree”
  - After first packet the shortest path between sender and receiver is used … the “source tree”
  - For Fabric the “shared tree” and “source tree” path is the same, because all traffic will go via the Spines (except directly between vPC peers)

- For PIM ASM there is a clear “direction” of the traffic, from the sender to the MC RP, and from MC RP towards the receiver
  - The sender and receiver for the MC streams, the BUM traffic of the end devices, will be the Server Leafs no end devices

- Default multicast load balancing for ASM is per group via the Spines
PIM ASM Scale

- Scalability PIM ASM
  - For each Bridge Domain with unique MC group address will be one *,G entry in the mroute table
  - For each sender for this Bridge Domain, potential all Leafs, will be one S,G entry in the mroute table
  - If you using 50 groups for your Bridge Domains / VLANs, and having 50 Server Leafs, this could result in 50 *.G and 2500 S,G entries so total 2550 mroute entries
  - Please be aware of the max mroute entries per platform (VXLAN table)
    - Nexus 5600 total 800 VXLAN MC groups (BiDir)
    - Nexus 7000 with SUP 2E, 1000 VXLAN MC groups (unclear BiDir and/or ASM)
    - Nexus 9300 128 VXLAN MC groups, and 256 VTEPS, max 4000 ASM mroute
PIM ASM Scale

• For MC RP and PIM ASM you should sync the S,G entries, usually done via anycast RP (no need for MSDP)
  • S,G are the Source of the traffic, here the Leafs, and the Group, bind to the Bridge Domain or VLAN

• Because of potential high amount of S,G entries the multicast group IP address design is important
  • Best load balancing over the Spines for BUM traffic for one MC group per Bridge Domain / VLAN … but maybe too many S,G entries
  • Best scalability all Bridge Domains / VLANs share the same MC group … but no load balancing as well all Leafs will receive all BUM traffic
  • Compromise is to use one MC group per VRF over all Bridge Domains … limit the amount of S,G entries, by allowing the BUM traffic load balancing, as well BUM traffic from away from Leafs not having this VRF configured
PIM BiDir General

• PIM BiDir (Bi-Directional Multicast routing)
  • The streams always goes from source to RP, and from RP to the receiver
  • Always the “shared tree” is used, never the “source tree”
  • In the Fabric always the Spine is in forwarding path, so there is no shorter path then via the Spine, aka no benefit of “source tree”

• Because only “shared tree” is used, there are only *,G entries in the mroute table and no S,G … scale much higher

• For PIM BiDir there is no clear “direction” of the traffic, it flows to and from the RP for a given group
  • The sender and receiver for the MC streams, the BUM traffic of the end devices, will be the Server Leafs no end devices
PIM BiDir General

- **PIM BiDir redundancy** is via called “Phantom RP”
  - Only one Spine is the active RP for the MC groups assigned to him, the standby RP didn’t get any BUM traffic as long the active RP is available
  - With this no sync protocol between active and standby RPs is required

- **IP addressing for the PIM BiDir Loopbacks**
  - The IP for PIM BiDir RP should not be configured, usually one higher
    - Best Practice from Service Provider setups
  - The “primary RP” have for the MC RP a /30 subnetmask, due to be able having “IP +1”
  - The “Phantom RP” have a /29 mask with same IP then primary RP
PIM BiDir General

• With only one PIM BiDir RP there will be no load balancing for BUM traffic over the Spines

• To achieve load balancing for BUM, multiple PIM BiDir RP required with different MC groups per RP
  • For single MC group all traffic goes via one RP / single Spine
  • Load balancing over both RP is achieved by different groups
  • With 2 Spines, each has is active RP for one MC group, as well Phantom RP for the other MC group
PIM BiDir General

- **PIM BiDir redundancy** is via called “Phantom RP”
  - Only one Spine is the active RP for the MC groups assigned to him, the standby RP didn’t get any BUM traffic as long the active RP is available
  - With this no sync protocol between active and standby RPs is required

- **IP addressing for the PIM BiDir Loopbacks**
  - The IP for PIM BiDir RP should not be configured, usually one higher
    - Best Practice from Service Provider setups
  - The “primary RP” have for the MC RP a /30 subnetmask, due to be able having “IP +1”
  - The “Phantom RP” have a /29 mask with same IP then primary RP
PIM BiDir General

- With only one PIM BiDir RP there will be no load balancing for BUM traffic over the Spines
- To achieve load balancing for BUM, multiple PIM BiDir RP required with different MC groups per RP
  - For single MC group all traffic goes via one RP / single Spine
  - Load balancing over both RP is achieved by different groups
  - With 2 Spines, each has is active RP for one MC group, as well Phantom RP for the other MC group
Conclusion

• Keep the Fabric Underlay multicast routing separated from outside

• Place the multicast rendezvous points at the Spine level
  • Make the rendezvous points redundant, 4 are good too

• Consider the platform capabilities when choosing the multicast routing protocol, if you have both multicast routing options:
  • Chose PIM ASM when Fabric contain Nexus 9000
  • Chose PIM BiDir with Fabric mix of Nexus 5600 and 7000 and if you need granular control which BUM traffic been received on the Leafs

• Balance the Bridge Domain to MC group mapping between the scalability of the Leafs and granularity of BUM traffic forwarding
DHCP Relay in VXLAN EVPN

- VXLAN EVPN add to **DHCP relay** adds a **challenge with anycast IP**
  - All Server Leafs using for one SVI the same IP address, the distributed gateway IP or anycast IP
  - With this the approach to add as GIADDR (option 80) the server SVI IP will not work, because path back is not distinct
  - In VXLAN EVPN a **unique IP** for the **requested Server Leaf** will be added to GIADDR which is in examples below either Loopback or mgmt0
  - According the standard, the DHCP server need to use the GIADDR to send back the DHCP offer. This path is deterministic with use of Loopback or mgmt0
- Due to challenge above, for **DHCP relay for VXLAN EVPN** need to use **DHCP option 82** and the sub-types on the DHCP server
DHCP Relay in VXLAN EVPN

• Some DHCP server have currently problems with these sub-options of DHCP option 82, current investigation shows this is mainly Windows 2012 servers, documented in CSCux88796

• The DHCP relay commands are used to enable different options of 82 of DHCP protocol required for setup in multi tenancy environment

• The “server ID override sub-option” is used for the client send DHCP renew and release from the client to the Server Leaf
  • By this modification on the Server Leaf the client will send renew and release to the Server Leaf, which will then forward the messages to the DHCP server
  • This will be the server SVI IP address which is the anycast IP on Server Leaf
DHCP Relay in VXLAN EVPN

• The “link selection sub-option” provides a mechanism to separate the subnet / link on which the DHCP client reside, and gateway IP (giaddr).
  • The Server Leaf add as GIADDR a unique IP, either a Loopback or his mgmt to which the DHCP server responds
    • In other environments the GIADDR is enough to define a distinct path back to the DHCP relay agent, but not with distributed gateway
  • The Server Leaf will add the server SVI subnetwork as link selection
  • From this range the DHCP server will provide the IP address to the client and bind it the MAC of the client
General Considerations – PXE boot

• DHCP is not that often used in data center, one of the exceptions is PXE boot for staging servers which adding new challenges
  • Most of the times the DHCP request will come untagged
  • As “PXE boot” bridge domain we need to have a L2 VNI which is configured to be used as untagged on HIF
  • Options described in “Layer 2” presentations:
    • Either “switchport mode access” if servers can be re-patched
    • Or “system nve allow-vxlan-native-vlan” for a trunk HIF to use the native VLAN

• The following examples are how to address this without use of auto-config trigger LLDP
  • With LLDP trigger we can set base on MAC database on DCNM the correct VLAN / L2 VNI, so DHCP request will be send it the desired VLAN without extra config per port
General Considerations – PXE boot

- Because in VXLAN EVPN we assume multi-tenancy, **how many “PXE boot domains” we need?**
  - One PXE boot server for all VRFs … unique PXE boot domain per VRF?

- Usually **different operating system require different PXE boot** processes and by this different PXE boot server
  - One PXE boot server per operating system and different for VRFs?

- On **which HIF is which “PXE boot domain” as native VLAN required?**
General Considerations – PXE boot

- Which path to DHCP server should been used?
  - Same then the client traffic via Border Leaf and Core?
  - Using VRF management / mgmt 0 the switches?
  - Using complete different VRF to forward this traffic?

- For automation profiles as well manual config we can use per SVI different DHCP server settings, network itself is flexible
  - Sometimes many options makes things more complicate

- Planning and get in sync with server teams might be the bigger challenge than to configure DHCP relay
PXE boot options without DHCP Relay

- There are few cases where DHCP Relay did not work
  - Due to automation no individual Loopback can be added on Server Leaf
  - DHCP server is not reachable via mgmt 0
  - DHCP options 82 not implemented on DHCP server
  - Due to bug ID CSCux88796 in some scenarios the DHCP options 82 on DHCP server does not work

- For such cases consider other options:
  - Run PXE boot VLANs as L2 bridge domains, and add DHCP interface directly into this bridge domain (no SVI in Fabric required for this)
  - Use external DHCP relay device which has a SVI in each PXE boot VLAN, and can use GIADDR like in traditional environment
  - Use separate infrastructure for staging of servers, then re-cable server after the installation process
EVPN VXLAN Fabric External Routing

VXLAN Overlay EVPN MP-BGP

Border Leaf

Leaf

Global Default VRF Or User Space VRFs

Routing Protocol of Choice

IP Routing

VXLAN Overlay EVPN VRF/VRFs Space
EVPN VXLAN Fabric External Routing (Cont’ed)

For Layer 3 interfaces, use one per overlay VRF instance. The routing protocol neighbor is in the EVPN VRF address family.

Layer 3 interfaces on the external devices can be in either tenant VRF instances or the global default VRF instance.

Interface-Type Options:
- Physical Routed Ports
- Subinterfaces
- VLAN SVIs over Trunk Ports (static routing possible via vPC)
**EVPN VXLAN External Routing with BGP**

**Sample Configuration – On the Border Leaf**

### On the VXLAN Border Leaf

```
router bgp 64910
  router-id 12.12.12.111
  address-family ipv4 unicast
  address-family 12vpn evpn
  neighbor 12.12.12.1 remote-as 64910
    update-source loopback100
    address-family ipv4 unicast
    send-community both
    address-family 12vpn evpn
    send-community both
  neighbor 12.12.12.2 remote-as 64910
    update-source loopback100
    address-family ipv4 unicast
    send-community both
    address-family 12vpn evpn
    send-community both

vrf BLUE
  address-family ipv4 unicast
  advertise 12vpn evpn
  redistribute direct route-map RM-ALL-FABRIC-RMAP-REDIST-SUBNET
  maximum-paths ibgp 2
  neighbor 32.32.32.2 remote-as 100
    description CORE NEIGHBOR 32.32.32.2 BGP AS 100
    address-family ipv4 unicast
    send-community both
    route-map RM-ALL-BGP-TO-CORE-FILTER out
```

### iBGP multipath required in order to benefit from ECMP

**IMPORTANT:** also required on the server-leafs in order to benefit from ECMP.

**IMPORTANT:** Issues seen if configured as eiBGP (seen in Lab)

### The eBGP neighbor is on the outside.

It’s in address-family ipv4 unicast of the tenant VRF routing instance.
EVPN VXLAN External Routing with BGP
Sample Configuration – On the Border Leaf

On the VXLAN Border Leaf

```bash
---snip---
vrf BLUE
  address-family ipv4 unicast
  advertise 12vpn evpn
  redistribute direct route-map RM-ALL-FABRIC-RMAP-REDIST-SUBNET
  maximum-paths ibgp 2
  neighbor 32.32.32.2 remote-as 100
  description CORE NEIGHBOR 32.32.32.2 BGP AS 100
  address-family ipv4 unicast
  send-community both
  route-map RM-ALL-BGP-TO-CORE-FILTER out

route-map RM-ALL-FABRIC-RMAP-REDIST-SUBNET permit 100 match tag 8300

route-map RM-ALL-BGP-TO-CORE-FILTER deny 20
match ip address prefix-list PL-ALL-DENY-L3-ALL-HOSTS

route-map RM-ALL-BGP-TO-CORE-FILTER permit 100 match ip address HOSTS

ip access-list HOSTS
  10 permit ip any any

ip prefix-list PL-ALL-ALLOW-LOOPBACK seq 10 permit 12.12.12.0/24 le 32
ip prefix-list PL-ALL-DENY-L3-ALL-HOSTS seq 10 permit 0.0.0.0/0 ge 28 le 32
```

For better scalability, apply prefix-list to filter out /32 IP host routes. Advertise prefix routes only to the external eBGP neighbor.

Used to redistribute specific local subnets
EVPN VXLAN External Routing with BGP

Sample Configuration

```
router bgp 64910
  vrf BLUE
    address-family ipv4 unicast
    network 20.0.0.0/24
    neighbor 30.10.1.2 remote-as 200
    address-family ipv4 unicast
      prefix-list outbound-no-hosts out

interface Ethernet2/9.10
  mtu 9216
  encapsulation dot1q 10
  vrf member BLUE
  ip address 30.10.1.1/30

interface Ethernet1/50.10
  mtu 9216
  encapsulation dot1q 10
  vrf member BLUE
  ip address 30.10.1.2/30

router bgp 100
  address-family ipv4 unicast
    network 100.0.0.0/24
    network 100.0.1.0/24
    neighbor 30.10.1.1 remote-as 100
    address-family ipv4 unicast
```
EVPN VXLAN External Routing with BGP

You could expect we receive in VRF Green also routes from VRF Blue, because Core have one VRF … why is this not the case?

From Core perspective all tenant/VRF prefixes come from same AS
  • Even though the Core has all prefixes from all tenant/VRFs he will NOT send them back into the same AS
  • Without default route from Core, no communication possible between tenant/VRFs inside of the same Fabric

Recommendation is to announce default route from Core > BL

On the VXLAN Server Leaf

This is the iBGP route. The next hop is the VTEP address of the border leaf.

This is the external route.

n9396-vtep-1# sh ip route vrf BLUE 100.0.0.0/24
IP Route Table for VRF "BLUE"
'*' denotes best ucast next-hop
'***' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%'<string>' in via output denotes VRF <string>

100.0.0.0/24, ubest/mbest: 1/0
  *via 10.1.1.16*default, [200/0], 01:01:14, bgp-100, internal, tag 100 (evpn)segid: 0x9858 tunnelid: 0xa010110 encap: 1
Default route implications and options
Originate 0.0.0.0/0 local on Border Leaf

• How to implement:
  • Add static 0.0.0.0/0 on Border Leaf pointing to next hop IP
    • Either virtual firewall IP, external learned network like MC RP, or to Null 0
    • Either redistribute static or use network statement in VRF BGP IPv4 address family

• Typical Use case:
  • The “exit” from the VRF is a Firewall
  • No default route is available in the Core
  • Complete separation from Fabric and external routing has high priority
Default route implications and options
Originate 0.0.0.0/0 local on Border Leaf

• **Pro:**
  • Separation Fabric and external routing
  • For devices which don’t support wide range of routing protocols
  • Static routing preferred and less complex setup on most firewalls
  • Possible with vPC via SVIs on Border Leaf

• **Con:**
  • It is static
  • Requires also static routing external to the Fabric
  • Need to add FHRP protocol on both sides for failover
  • Higher operation effort, if adding/removing networks is dynamic
Default route implications and options
Receive 0.0.0.0/0 via IGP from Core

• How to implement:
  • Receive 0.0.0.0/0 as route via IGP like OSPF from Core
  • Add this to VRF BGP IPv4 address family via network statement
  • Possible also to redistribute only the 0.0.0.0/0 route into BGP
  • The Server Leaf don’t need to receive all more specific routes from Core
  • The Fabric prefixes need to be redistributed into IGP or summaries need to be build on the Border Leaf

• Typical Use case:
  • Current Core runs IGP like OSPF or EIGRP and BGP should be avoided
  • Dynamic routing is desired, but network services only support e.g. OSPF
Default route implications and options
Receive 0.0.0.0/0 via IGP from Core

• **Pro:**
  • For Core running IGP, no change required
  • Dynamic advertisement of moving / adding networks to the Fabric
  • Easier to use for network services as BGP

• **Con:**
  • Partial redistribution between IGP and BGP required
    • Adds complexity especially if control function like ACLs are used
  • Lose capabilities and options of BGP like marking prefixes via standard BGP communities as suggested to signal the tenant / VRF
  • Not possible via SVI towards vPC member connected device
Default route implications and options
Receive 0.0.0.0/0 prefix via eBGP from Core

• How to implement:
  • BGP to Core or network service from Fabric Border Leaf’s
  • 1:1 transport of Fabric prefixes, or use summarization on Border Leaf
  • Suggestion is to add BGP standard communities on Border Leaf to be able to identify the VRF / tenant in the Fabric

• Typical Use case:
  • Core or network service supports BGP with own AS
Default route implications and options
Receive 0.0.0.0/0 prefix via eBGP from Core

• Pro:
  • Easiest to implement and use (if operations has BGP skills)
  • Full set of BGP capabilities can be used to influence routing external to the Fabric
    • E.g. adding BGP standard community to identify VRF / tenant in Fabric

• Con:
  • Requires BGP on the Core or network services
  • Not possible via SVI towards vPC member connected device

• The following output shows an example for 0.0.0.0/0 received via eBGP on Border Leaf and forwarded to Server Leafs
Default route implications and options
Conditionally or unconditionally

• **Conditionally:**
  • The default route is only advertised into the fabric if there is a path available on the Border Leaf to the Core

• **Unconditionally:**
  • No matter if there is a path to the Core, the default route is always advertised into the Fabric
Conditionally vs. Unconditionally

Core advertises default route to Border Leaf

Border Leaf learns default route via eBGP from Core AND via iBGP from neighbor Border Leaf

Both Border Leaf advertise default route via iBGP to RR/Leafs (ECMP)

Border Leaf

Spine

VXLAN Overlay
EVPN MP-BGP

Routing Protocol of Choice

Core advertises default route to Border Leaf

IP Routing
Conditionally vs. Unconditionally

Both uplinks from Border Leaf 1 fail

Border Leaf 2 learns default route via eBGP from Core

Conditionally:
Only Border Leaf 2 advertise default route via iBGP to RR/Leaves (ECMP)

Border Leaf 1 doesn't learns default route via eBGP from Core.
But learns via iBGP from neighbor Border Leaf

Result:
Leaf's receive only default route from Border Leaf 2 as this is the only direct way to the Core

Core advertises default route to Border Leaf
Conditionally vs. Unconditionally

Leaf learns default route only from Border Leaf with direct path to Core

N56-1# sh ip bgp vrf cola:blue
BGP routing table information for VRF Cola:Blue, address family IPv4 Unicast
BGP table version is 299, local router ID is 172.20.51.193
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

<table>
<thead>
<tr>
<th>Network</th>
<th>Next Hop</th>
<th>Metric</th>
<th>LocPrf</th>
<th>Weight</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>*&gt;0.0.0.0/0</td>
<td>9.0.0.22</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>64999 i</td>
</tr>
</tbody>
</table>

Only the default route from BORDER-2 is advertised

BORDER-1 doesn't advertise the default route anymore, as there is no direct path out of the fabric from BORDER-1.

CONDITIONAL advertisement is recommended
External Connectivity at Border (Layer-3)

- 2-Box Solution
- Routing-Session for each VRF
- Dedicated Interface (Physical or Logical) per VRF
- Configuration intensive
- Can be Automated
Seamless MPLS Integration – BorderPE

- 1-Box Solution
- Single Routing-Session
- Single Interface for all VRF
- Lean Configuration
- Can be Automated

Use-Case:
- External Connectivity
- Layer-3 DCI

Nexus 7x00 with F3 – 7.3 (shipping)
ASR9k TH/TYP – 5.3.2 (shipping)
ASR1k – (roadmap)
Seamless MPLS Integration – BorderPE

```
router bgp 65500
neighbor to-MPLS remote-as 65599
  update-source loopback0
  address-family vpnv4 unicast
    import l2vpn evpn reoriginate
  address-family vpnv6 unicast
    import l2vpn evpn reoriginate
neighbor to-VXLAN remote-as 65500
  address-family l2vpn evpn
    import vpn unicast reoriginate
```

Nexus 7x00 with F3 – 7.3 (shipping)
ASR9k TH/TYP – 5.3.2 (shipping)
ASR1k – (roadmap)
If that seems complicated..

- **DCNM10** includes full automation for both day0 provisioning as well as day1/day2 tasks
- **Ignite** is an open source tool on github that provides an automation framework for Day0 provisioning
- **Nexus Fabric Manager** (Nexus9000 only) allows turn key provisioning from an appliance
- **Cisco VTS** (Virtual Topology System) to provision overlay
- Industry’s first **puppet** plugin for MP-BGP EVPN
- Bring your own controller (**BYOC**)
- OR ……. 
ACI Fabric supports discovery, boot, inventory, and systems maintenance processes through the APIC

- Fabric discovery and addressing
- Image management
- Topology validation through wiring diagram and systems checks
- Fully automated BGP RR and VXLAN VTEP discovery / configuration and removal
Agenda

• VXLAN Recap and WHY MP-BGP is a requirement
• Real World VXLAN Deployments
• Migration Strategy
General Migration Consideration
Strategy of Migration

• We have lot of experience of Migration, moving from STP environment to use VSS / vPC as well from STP / VPC to FabricPath

• More important than the technology is for migration clear definition of responsibility and coordination of the different steps … outside of the scope of this session

• Where we can help is to define how the interconnect from new to old world should look like, as well with technical migration strategy … this we want to get right
General Migration Consideration
Migration Steps Overview

• Don’t do “flag day” approaches!
  • From past experience, no good idea to migrate on the fly running switches with servers to VXLAN EVPN
  • Would cause longer down time which clients don’t have, and higher risk, which clients don’t want

• Follow the approach of small steps
  • Build up new environment beside exiting … make this stable and use it for testing as well building experience for the client
  • Then create the connection from existing to new environment
  • Probably faster than “flag day” … depending on how often things go wrong
General Migration Consideration
Migration Steps Overview

• **Bring up new Fabric … without servers**
  • Obviously first the Spines to be able interconnect the Leafs
  • Then the Border Leafs, which later create the connection to existing devices as well the routing to the Core
  • Last one or couple of Server Leaf pair(s)

• **Establish the connection between new and existing**
  • First do Layer 2 connection … probably with move of STP root for VLANs
  • Second move Layer 3 with anycast IP into the Fabric
  • Lasts step, cut the existing layer 2 connection for migrated VLANs
General Migration Consideration
Migration and Automation

• Control of Migration steps is most important
  • Preferred is manual configuration
  • Consider to use the Border Leaf

• Do NOT mix auto-config and manual config
  • With auto-config assumption is the same config is used in the whole Fabric
  • When you mix this you break this rule with unpredictable effects
  • It is not supported to mix auto-config and doing afterwards in this part manual changes
  • One reason that general Server Leaf should not used to connect existing environment to the Fabric
General Migration Consideration
Layer 2 / Spanning Tree – In General

• The **STP root should be in the Fabric**
  • Would mean **STP root change in existing environment for VLANs start migration**, the other VLAN STP root will be not affected
  • Probably STP root change is the biggest impact … duration depending if RPVST is used as well how many VLAN moved per step
  • Possible with VXLAN to run STP root outside of Fabric, but risk is higher due to Fabric itself can’t break a Layer 2 loop
  • Avoid BPDU filter, because wouldn’t allow STP to break a Layer 2 loop
General Migration Consideration
Layer 2 / Spanning Tree – In General

• For Layer 2 use of Leaf running as vPC to existing environment
  • Consider dedicated Leaf for L2 connection … later been Server Leaf
  • The STP BLK ports will be inside of the existing environment, full bandwidth to new Fabric will be utilized
    • Drawback is that communication between existing environment might go via Fabric
    • With vPC the same MAC will be used to create bridge ID (via peer-switch)
  • Separate Border Leaf who does only routing to the Core without vPC
  • Don’t even think about using the Fabric Spines
  • In existing environment consider Distribution as connection, because most time central point as well vPC / VSS is there in use
  • Recommendation is to use vPC for connection to existing environment
    • Main reason is higher resiliency and lower convergence in case of failure, also better use of bandwidth to the new Fabric
General Migration Consideration
Layer 2 / Spanning Tree – In General

- vPC leafs for L2 connection to existing environment
- non-vPC border leafs for L3 connection to existing Core
- vPC leafs for L2 connection to existing environment

Diagram:
- Existing Core
- vPC PL
- Leaf
- Distribution
- Border Leaf
- VXLAN EVPN Fabric
- CE Link with STP
- Routed Fabric Link
- Routed Core Link
General Migration Consideration
Layer 2 / Spanning Tree – In General

• If isolated Leafs have been used to connect existing environment
  • Assumption is STP root is in the Fabric, and STP normal with root guard is used on Fabric ports to existing environment
  • The STP BLK will be a port on the Fabric switches, because of root guard.
    • There can be only one STP root (The Leaf with lower system MAC will win)
  • Less optimal use of bandwidth from existing environment to Fabric
    • Benefit, traffic inside of the existing environment will stay inside existing environment and not cross the Fabric
  • To use isolated Leafs to connect existing environment should be avoided
General Migration Consideration
Layer 2 / Spanning Tree – In General

For clients which refuses vPC

Drawback: STP BLK ports, also between existing Distribution
General Migration Consideration
Layer 2 / Spanning Tree - Questions

• Do we have single environment or multiple for consolidation?
  • Should the new Fabric consolidate multiple separated environments?
  • Is there one big STP environment to be migrated?
  • This will tell: How many L2 connection points we need on “L2 Leaf” … can we live with single connection, or need multiple “L2 Leafs”?

• How stable is STP in the current environment?
  • How often do we see TCN BPDU’s in the environment?
  • Many STP root changes?
  • This will tell: How many VLANs we can move STP root per step, how many changes are required in total
General Migration Consideration
Layer 2 / Spanning Tree - Questions

• Is any VSS or vPC in use?
  • Did the current distribution run kind of VSS or vPC?
  • This will tell: Makes the connection between existing environment and new Fabric much more resilient, reduce impact of single box reload

• Do we have overlapping VLAN IDs?
  • Are there multiple different environments with overlapping VLAN IDs need to be consolidated?
  • This will tell: Do we need to think about VLAN translation options on the “Migration Leaf” and which kind option might fit
General Migration Consideration
Layer 2 / Spanning Tree - Questions

• Is the STP root today on single box, or distributed?
  • Did one switch, or 2 switches hosting the STP roots for all VLANs?
  • This will tell: How many steps we need to migrate the STP root into the new Fabric

• Did we have “daisy chain” STP environment?
  • Is there clear structure of distribution with one access switch tier?
  • Due to infrastructure or history, multiple tiers of access switches?
  • This will tell: How complex is the STP environment, how much connection points we need for the Border Leafs
General Migration Consideration
Layer 3 / Routing – In General

• For Layer 3 in the Fabric use anycast IP
  • HSRP is working only without VNI to the SVI, like for network services
    • Because to add hosts to control plane EVPN config is required, will not work

• As external routing protocol you have lot of options
  • Recommended is eBGP, makes design less complex
  • For the Fabric a 0.0.0.0/0 to outside is required … and enough
    • If other IGP is required, don’t do mutual redistribution
    • In case redistribution is required, do this only from Fabric prefixes towards the IGP of the Core
  • Better way is to build summaries on Border Leaf, and advertise only them
General Migration Consideration
Layer 3 / Routing – In General

- **Use of summaries on Border Leaf during migration**
  - During migration using summaries might lead to traffic going to “wrong” side
    - E.g. in existing data center still announce /24 with higher metric and the new Fabric start to announce with /20 … traffic will not send to new Fabric for migrated L3
  - Consider during migration not to use summaries, until a full block is migrated to the new Fabric
General Migration Consideration
Layer 3 / Routing - In General

• Start with filtering host routes from Border Leaf to Core
  • Filter the host prefixes to Core is recommended, will help also in case no summaries been build during migration
  • Keep in mind the Loopback in the VRFs on the Border Leaf or maybe also Server Leaf, allow these host mask to the Core
  • For migration, to filter host prefixes also minimize the impact for back out the change / moving back the Layer 3 to the existing environment
General Migration Consideration
Layer 3 / Routing - Questions

• Who should announce networks in process of migration?
  • Should this be done by the Fabric, or by existing devices?
  • Usually the environment have the “active DGW” is doing this
  • This will tell: When to announce from Fabric the server networks, maybe as interim with bigger metric as fallback

• What is the routing protocol between Data Center and Core?
  • Many options in the field, from static routing over EIGRP / OSPF to BGP
  • This will tell: How smooth will be the migration, as well the redistribute options you can use on the Border Leaf
General Migration Consideration
Layer 3 / Routing – Questions

• Default Gateway of servers
  • Who is in current environment the DGW of the servers?
  • Who should be the DGW of the servers in new Fabric?
  • This will tell: For which bridge domain is the Fabric “just Layer 2”, and for which the Fabric will route and hosting the DGW

• Are there any Firewall terminating Layer 3 for servers?
  • Is this also the plan for new Fabric?
  • Should the Fabric provide the VRF for existing security zone, with exit point to Firewall?
  • This will tell: How many VRFs did we need, as well again about Fabric “just layer 2” or more routing in the Fabric
General Migration Consideration
Layer 3 / Routing – Questions

• Overlapping IP addresses are a topic?
  • In case different environments been consolidated could happen
  • Mainly a topic for Service Provider environments
  • This will tell: If we need a kind of NAT device between VRFs / tenants as well about how the routing from Border Leaf to Core must be setup
Option 1 … Direct Connections
Option 2 … Interim L2 Phalanx
Requirements might lead to this option

- Multiple connection points between new Fabric and existing STP environments required
  - Physical disperse connection points
  - Infrastructure is not capable of providing all links to one location
- Many different independent STP domains
- Current STP environment is instable, first need to isolate the existing STP failure domains
- Overlapping VLAN IDs, and VLAN translation should happen outside of the new Fabric
- No vPC or VSS in existing environment
Option 2 … Interim L2 Phalanx
Overlapping VLAN IDs
General Considerations

• When existing environments need to merge into one Fabric, overlapping VLAN IDs is often an issue due to historical reasons

• There are different solutions for this, which fit best depending on:
  • Is there a fix block of overlapping VLAN IDs?
  • How many or sum of overlapping VLAN IDs we need to deal with?
  • Amount of different environments need to be consolidated?
Overlapping VLAN IDs
Solution Overview

• Port-Local VXLAN VLAN mapping
  • Based on Nexus 9300
  • One Leaf is connected to multiple existing environments
  • On a per port basis the mapping is done
  • Fit best for small block of overlapping VLANs
  • For use case overlapping VLANs is the exception
Overlapping VLAN IDs
Solution Overview

• VXLAN VLAN mapping on dedicated Leaf pairs
  • Supported on all platforms
  • One Leaf pair required per existing environment
  • Mapping happens individual per area on this Leaf
  • For diverse VLAN overlapping and most scalable solution
  • For use case overlapping VLANs is usually the norm
Overlapping VLAN IDs
Port-Local VXLAN VLAN Mapping

- In this example:
  - VLAN 3000 and 3001 overlap in different environments
    - The overlapping VLAN IDs don’t need mapping to Segment IDs!
  - They will map to different vn-segment IDs on a per port basis

```
interface Ethernet1/7
  switchport mode trunk
  switchport vlan mapping enable
  switchport vlan mapping 3000 3100
  switchport vlan mapping 3001 3101
  switchport trunk allowed vlan 3100,3101

interface Ethernet1/8
  switchport mode trunk
  switchport vlan mapping enable
  switchport vlan mapping 3000 3102
  switchport vlan mapping 3001 3103
  switchport trunk allowed vlan 3102-3103
```
Overlapping VLAN IDs
Port-Local VXLAN VLAN Mapping

• Available since 7.0(2)I2(2) on Nexus 9000
  • Current limit is 100 PV mappings per interface, total 1K L2 VNIs per Leaf

• Allows to map the same 802.1Q VLAN tag to different VNIs on different interfaces of the same Leaf
  • The VLAN ID to Segment ID mapping happens per switch and not per port
  • The Overlapping VLAN IDs will NOT get a Segment ID!

• No extra config for overlapping VLAN range
  • The overlapping VLANs will be mapped per port basis
  • The overlapping VLAN range can be used also inside of the Fabric
Overlapping VLAN IDs
Port-Local VXLAN VLAN Mapping

• Clear structure required which VLAN ID overlaps
  • Less scalable but very flexible
  • The VLAN range which overlaps should be known and best been in one block like VLAN 1-500 overlap in all environments

• “Free block” of VLAN IDs must be available for “translation VLANs”
  • If VLAN ID 1-10 overlaps, and coming from 4 different environments, we need as VLAN block for translation 40 free VLAN IDs
  • For historical grown environments maybe above ID 1000 is free
Migration Example

Layer 3 Routing
Static, OSPF, BGP

Layer 2 vPC Trunk

- STP compatibility with Classic Network
- VLAN 10 maps to VNI Blue_1
- VLAN 11 maps to VNI Blue_2
- Classic Devices are still the Default Gateway
- Equally applicable to L4-7 services (FW/LB) in the Classic Network
- No ARP suppression enabled
- Once migration completed, insert needed services and move Default Gateway ACI BDs

Tag could be VLAN ID or VNID.
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 through the Cisco Live mobile app or from the Session Catalog on CiscoLive.com/us.

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
- Lunch & Learn
- Meet the Engineer 1:1 meetings
- Related sessions