VxLAN/EVPN with BGP Control Plane in a DCI Environment

Date: 17 April 2015

Table of Contents

Goal of this Document

This paper discusses the evolution of VxLAN and whether it is suitable for a DCI solution.

It covers the current implementation and emphasizes the Cisco enhanced implementation of VxLAN protocol with its control plane MP-BGP/EVPN.
It examines current shortcomings and presents the potential requirements for VxLAN to offer a solid DCI environment.

Audience

This document is written for the data center solution architects, network and system administrators, IT organizations, and consultants who are responsible for designing and deploying the network, compute, and storage devices that comprise a cloud computing solution.

Reading

This document is written in two main parts: a short summary and a detailed packet walk with all the requirements for DCI.

Some readers may want to stop after the first section (pages 1 to 7), because it contains enough to understand the evolution of VxLAN and capture a high-level overview of VxLAN deployment options for a DCI solution. Other readers may want to understand the technical details of the workflows and learn about the DCI requirements and functional improvements planned for the next release of VxLAN, which together comprise the second part.

VxLAN Evolution in the Context of DCI Requirements

The standard VxLAN protocol (RFC 7348) is aimed at carrying Layer 2 network traffic across a virtual tunnel established over an IP network; hence from a network overlay point of view there is no restriction to transport a Layer 2 frame over an IP network, because that's what network overlays are for.

Recently, the DCI market has become a buzz of activity around the evolution of VxLAN based on the introduction of a Control Plane (CP). In this network overlay context, the Control Plane objective is to leverage Ingress replication for Unicast transport while processing VTEP and host discovery and distribution processes. This method significantly reduces flooding for Unknown Unicast traffic within and across the fabrics.

Consequently, this noise requires a clarification on how reliable a DCI solution can be when based on VxLAN Unicast transport using a Control Plane.

What Does VxLAN Evolution to Control Plane Mean?

We can consider four different stages in VxLAN evolution.
These stages differ in the transport and learning processes as well as the components and services required to address the DCI requirements.

- The first stage of VxLAN relies on a Multicast transport mode with no Control Plane. Learning for host reachability is treated using Flood&Learn. This is the original transport method and behavior that became available with the first release of VxLAN. Let's call this first stage, VxLAN phase v1.0*; however, it is out of the scope of this post as it relies on Flood&Learn and has been already broadly elaborated in different forums (e.g. "Is VxLAN a DCI solution for LAN extension ?") clarifying why Multicast-based VxLAN was not suitable to offer a viable DCI solution.

- The second stage of VxLAN relies on Unicast-only transport mode. This mode leverage a control plane to announce to each VTEP the list of VTEP IP addresses and the associated VNIs. It does NOT discover nor populate any host information. It uses a Head-End Replication mechanism to be able to "flood" (in the data-plane) BUM traffic for each VNI. Nonetheless, the added value with this mode is that, an IP multicast network for the learning process across multiple sites is not mandatory anymore. However, although this mode relies on Flood&Learn discovery process, I think it's interesting to elaborate this Unicast-only mode transport, essentially due to some implementations of this basic HER stage claiming to offer a DCI solution. Let's call this stage VxLAN phase v1.5*. This phase is elaborated in the next sections.

- The third stage of VxLAN provides dynamic host discovery and distribution across all VTEP using a Control Plane MP-BGP EVPN Address Family. The host information consists of the IP and MAC identifiers of the end-point concerned by the Layer 2 adjacency with all its peers. This

mode of VxLAN is interesting as it helps with reducing drastically the flooding within and between fabrics; thus, in a nutshell, it might be considered "good enough" when diverted as a DCI model. However it is important to note that some of the functions required for a solid DCI solution are not all there. To avoid confusion with other implementation of VxLAN and for references throughout this article, let's call this stage VxLAN phase v2.0*. This mode is deeply elaborated in the next sections.

- The fourth stage of VxLAN offers similar transport and learning process as the current implementation of VxLAN MP-BGP EVPN AF (Phase 2.0*); however, it will introduce several functions required for a solid DCI solution. Let's call this stage VxLAN phase v3.0 *.

* Phase "n": be aware that there is nothing official nor standard in this nomenclature. This is mainly used for referencing in the following writing.

In a nutshell, which DCI functions are supported with VxLAN?

You can find more details in the Addendum

VxLAN Multicast (VxLAN Phase 1.0)

- Dual-homing

VxLAN Head-End Replication (VxLAN Phase 1.5)

- Multicast or Unicast transport (Head-end replication)
- Dual-homing

VxLAN BGP EVPN AF (VxLAN Phase 2.0)

- Control-Plane Domain Isolation using eBGP
- Independent Autonomous System
- Multicast or Unicast transport (Ingress Replication enabled by CP)
- ARP spoofing: reducing ARP flooding requests
- Hair-Pinning reduced with Local Default Gateway (L3 Anycast Gateway)
- Dual-homing using vPC and Anycast VTEP

Wishes for a Next phase of VxLAN BGP EVPN AF *

- Data-Plane Domain Isolation
- Unknown Unicast (including UU ARP) suppression
- Selective Storm Control (per VNI)
- Native L2 Loop detection and protection
- Multi-homing (Ethernet Segment ID supporting Designated Forwarder and Split-horizon)

* There is no commitment to any of the above. This list of functions is given for information on what could be implement in VxLAN to improve some existing shortcomings in regard to DCI

A slight clarification on Multi-homing: it relates to a DCI solution with embedded topology independent Multi- homing, versus a MLAG based dual- homing built with a tightly coupled pair of switches.

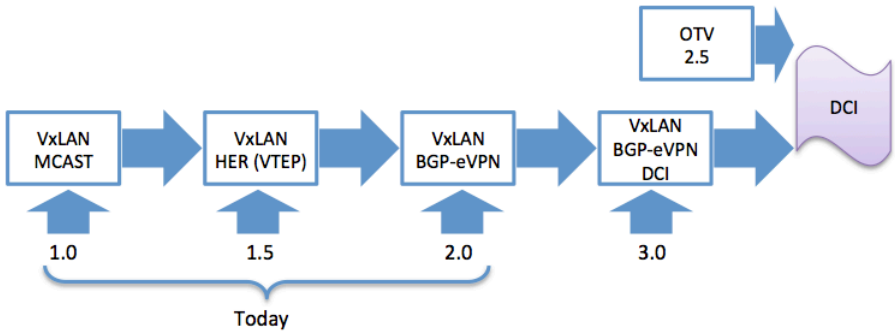


Figure 1: VxLAN evolution toward DCI functions

The figure above aims to provide a high-level overview of VxLAN evolution towards full support for DCI requirements.

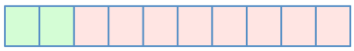
VxLAN evolution rating for DCI

Today:

Multicast transport for Flood&Learn

- Definitely excluded for any DCI solution

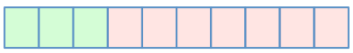
If one were to rate how efficient this implementation of VxLAN would be for DCI, we would give it a 2 out of 10 for VxLAN 1.0



Unicast-only mode with HER for VTEP discovery only

- Definitely not efficient for any DCI solution
- Relies on Flood&Learn
- Having a Unicast-only mode with HER to handle BUM doesn't mean that it should be considered a valid DCI solution, be cautious.

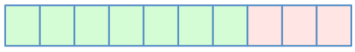
It deserves a maximum of 3 out of 10 for VxLAN 1.5



Control Plane (BGP EVPN AF)

- This can be considered "good enough" for VLAN extension but is not a validated DCI solution *per se*.
- Some DCI requirements are still missing, and we are missing the performances and scalability figures.

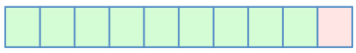
We would give it 7 out of 10 for VxLAN 2.0



Upcoming: Same Control Plane BGP EVPN AF as above but with additional services to improve and secure inter-site communication

- Should fill some important gaps for DCI requirements

If this evolution honors the promises, we may give it 9 out of 10 for VxLAN 3.0



If several models of ASIC (Custom ASIC or Merchant Silicon) support VxLAN encapsulation, it doesn't mean that all vendors have implemented a Control Plane improving the learning process for the hosts. Some vendors, such Cisco, supports Multicast transport and BGP-EVPN AF Control Plane (VxLAN Phase 2.0) for its Data Center switches (Nexus series) and WAN edge routers (ASR series). Others have implemented VxLAN 1.5, while certain vendors support only Multicast-based transport (VxLAN 1.0).

Consequently, it is important to examine the different modes of VxLAN in order to capture which implementation could be suitable for a DCI solution.

Some readers may want to stop here, because it should contain enough to understand the evolution of VxLAN and capture a high-level overview of VxLAN shortcomings in a DCI solution. Other readers may want to understand the technical details of the workflows and learn about the DCI requirements for the next release of VxLAN, which together comprise the second part. For those brave readers, read the next:

Let's Go Deeper with the Different Phases

Global Architecture for Dual-Sites

Let's assume the following DCI scenario as a base reference for the whole document. Two data center network fabrics are interconnected using a Layer 3 core network.

As described throughout this article, and often mentioned in the hybrid cloud environment, maintaining the Control Plane independent from each location is recommended as a strong DCI requirement.

CP independency is elaborated in the next pages; however, a single Control Plane stretched across two locations is also discussed for VxLAN phase 1.5.

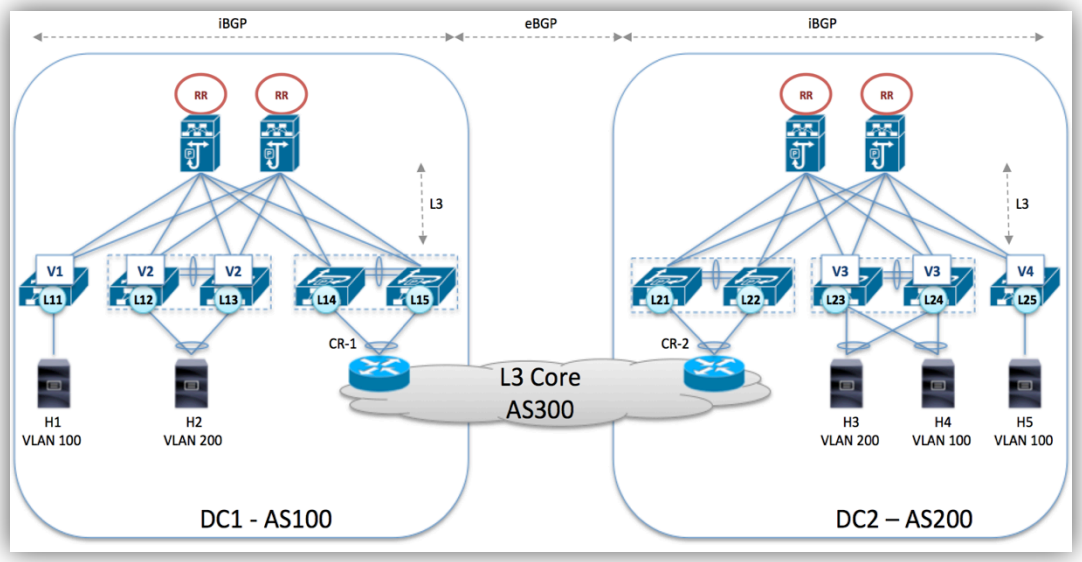


Figure 2: Global architecture for Dual-Sites

DCI and fiber-based links

This document focuses on a Layer 3 Core Network used for interconnecting sites. The Layer 3 Core network is usually acceded through a pair of Core routers enabled at the WAN Edge layer of each Data Center. This solution typically offers long to unlimited distances to interconnect two of multiple Data Centers.

However, before we go further in this document, it is important to briefly clarify another design generally deployed for Metro distances, which relies on fibers connectivity between sites.

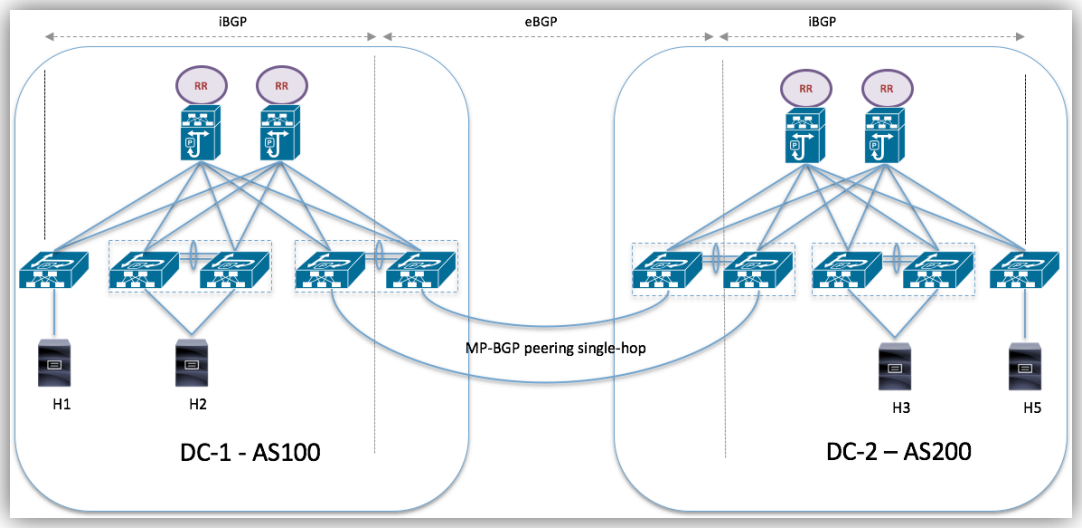


Figure 3: Design consideration for Metro distances using fibers/DWDM for DCI

From a control plane point of view, it can still support independent BGP Control Plane per Fabric network if chosen. Host information is dynamically learnt and populated according to the CP mode supported by the VxLAN implementation as described in the following pages.

However the main differences between the two designs, reside on the service that can be enabled on top of this Metro solution architecture. Indeed, conventionally, IP Multicast is not often an option with Layer 3 managed services when the Enterprise wishes to deploy a Layer 3 backbone (WAN/MAN). In a Metro distances, most of the time optical links are provided to interconnect sites. With dark fibers owned by the Enterprise itself or with DWDM managed services the network manager can easily enabled IP Multicast on his own.

That gives the flexibility to leverage IP Multicast for BUM traffic also inter-site as desired.

Note that when deploying DWDM, like with any traditional multi-hop routed network, it is recommended to enable BFD. The objective is improve the detection of remote link failure and to accelerate the convergence time. This is particularly true in case the optical switch doesn't offer this service (aka remote port-shutdown). However BFD is usually not necessarily for direct fibers (e.g. large campus deployments).

VxLAN and Reference Design Taxonomy

In the following scenarios and figures

- VxLAN Segment: This is the Layer 2 overlay network over which endpoint devices, physical or virtual machines, communicate through a direct Layer 2 adjacency.
- VTEP stands for VxLAN Tunnel EndPoint (It initiates and terminates the VxLAN tunnel)
- “V” stands for VTEP
- “VNI” stands for Virtual Network Identifier (or VxLAN Segment ID.)
- “NVE” this is the VxLAN Edge Devices that offers VTEP (also sometimes referenced as “Leaf” in this document as we assume VTEP is enabled on all Leafs in the following scenarios).
- “H” stands for Host
- “Host” may be a bare-metal server or a virtualized workload (VM). In both cases the assumption is that the “host” is always connected to a VLAN.
- “L” stands for Leaf (This is the Top Of Rack or ToR)
- “RR” stands for (BGP) Route Reflector
- “Vn” where “n” is the VTEP IP address
- “CR” stands for Core Router (redundancy of Core Routers is not shown)
- “vPC” stands for Virtual Port-Channel. This is a Multi-Chassis Ether-Channel (MEC) feature. Other forms of MEC exist such as Multi-chassis Link Aggregation (MLAG). I’m using vPC as references because it brings a keen feature known as Anycast VTEP gateway that allows sharing the same virtual VTEP address across the two tightly coupled vPC peer-switches. “

With the Dual-homing function, the same VTEP Identifier is duplicated among the 2 vPC peer-switches representing the same virtual IP address (to simplify the figures in this post, some identifiers have been simplified e.g., V2 where “2” is the

IP address). Note that the Anycast IP address is used for dual-homed vPC members and Orphan devices with a single attachment.

Dual-homing in the context of VxLAN is discussed further in this paper.

Yet this paper does not intend to deeply detail the dual-homing technology *per se*, however, it will highlight its crucial role for the DCI needs. Additional technical details on redundant vPC VTEP can be found in the following pointers:

- **Building Redundant vPC VTEPs with Cisco Nexus 9300 Platform Switches**
- **VPC Considerations for VXLAN Deployment**

In addition, for references to this document:

- VLAN 100 maps to VNI 10000
- VLAN 200 maps to VNI 20000
    - Note that VLAN are local significant to each Leaf or each port of Leaf, thus the VLAN ID mapped an a VNI can be different on remote Leaf.
- It is assumed that all VTEPs know their direct-attached hosts
    - Silent host will be discussed in a further section with VxLAN phase 2.0.
- BGP Route Reflectors are enabled to redistribute the information to all of the VTEPs that it knows. They are usually deployed for scaling purposes.
- One of the main DCI requirements is to maintain Control Plane independence on each site.
    - In the first design, a unique Control Plane with the same Autonomous System (AS) spread over east to west (E-W) locations is discussed.
    - The DCI design will evolve with 2 independent Control Planes in regard to VxLAN Phase 2.0, as recommended for DCI architectures.
- Finally, to reduce the complexity of the scenario, let's assume that the Border-Leafs used to interconnect to the outside world have no hosts attached to them, obviously the Border-Leafs may be also used as a ToR. A further section discusses this scenario.

VxLAN Phase 1.0: Multicast Transport

As a reminder, this first stage has been deeply elaborated in this article http://yves-louis.com/DCI/?p=648.

To summarize this solution, this first implementation of VxLAN imposes Multicast transport. It therefore assumes IP Multicast transport exists inside and outside the DC fabric. IP Multicast transport in the Layer 3 backbone may not be necessarily an option for all enterprises. In addition, a large amount of IP Multicast traffic between DCs could be a challenge to configure and maintain from an operational point of view (without mentioning performance and scalability).

Above all, it relies on Flood&Learn to compute the host reachability with no or limited policy tools for rate- limiting the concerned IP Multicast groups.
As a result, with excessive flooded traffic and bandwidth exhaustion, the risks of disrupting the second Data Center are very high.
The above is highlighted without counting on lack of traffic optimization due to nonexistence function of distributed Anycast gateways. As the result, "hair-pinning" or "ping-pong" effect may seriously impact the performances of the applications.

Consequently, Flood & Learn VxLAN may be used for very limited and rudimentary LAN extension across remote sites. If technically it carries Layer 2 frames over a Layer 3 multicast network, be aware that this is risky and it would necessitates additional intelligent services and tools to address the DCI requirements, such as leveraging a Control Plane, Flood suppression, independent Control Plane, Layer 3 anycast gateway, to list just few, which are not applicable with this implementation of VxLAN.

Therefore the recommendation is to NOT deploy VxLAN Multicast transport as a DCI solution.

VxLAN Phase 1.5: Unicast-only Transport Using Head-End Replication

This implementation is also known as Unicast-only mode. What it means is that it relies on Head-end replication to dynamically discover and distribute the VxLAN VTEP information necessary to build the overlay network topology. However, hosts are learnt using the traditional Flood & Learn data plane.

Packet walk

Let's analyze a packet walk step by step for Head-end replication with dynamic VTEP discovery.

In the following design, the Control Plane is spread across the two sites. An IGP protocol of choice is used to communicate between all the Leafs (OSPF, EIGRP, ISIS, etc) establishing the network overlay tunnels used for the data plane. Optionally, a redundant Control Plane is initiated on DC-1 to scalability purposes and to simplify the learning process.

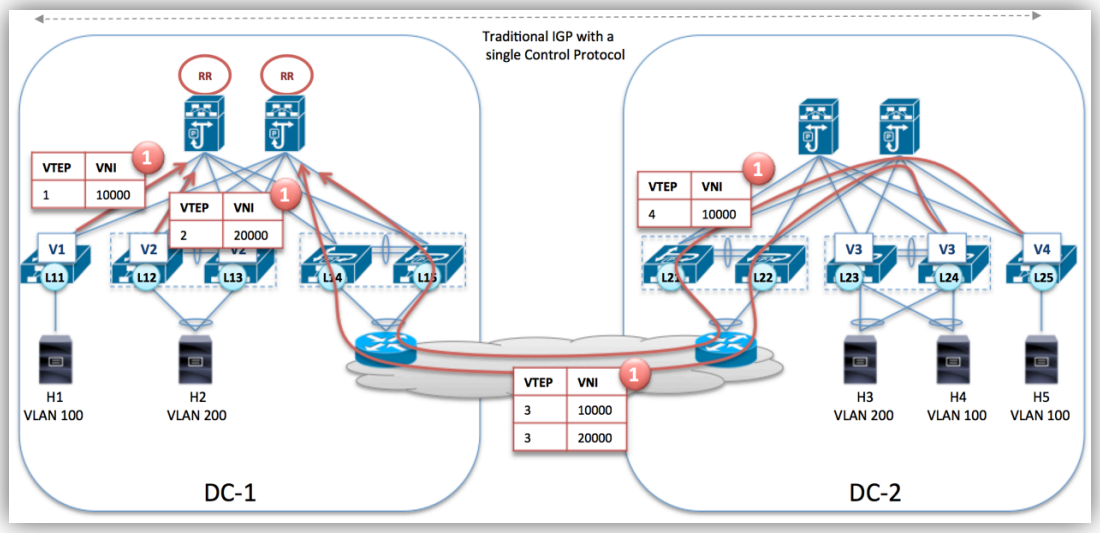


Figure 4: VTEPs advertise their VNI membership

1. All VTEPs advertise their VNI membership to the Control Plane. The target is the Route Reflector located in DC-1.

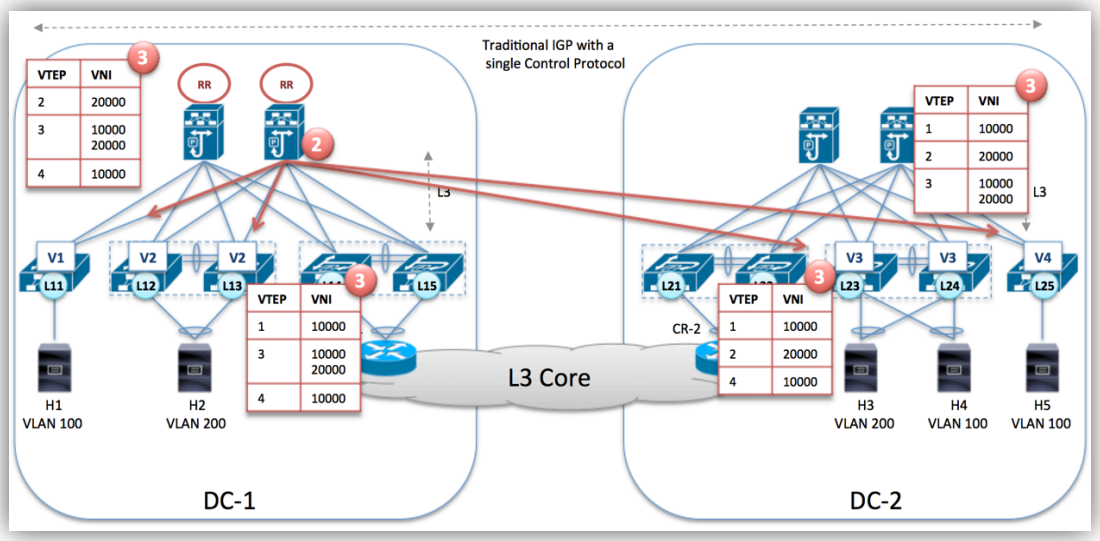


Figure 5: BGP propagates the VTEP information

2. The Control Plane consolidates the VTEP information and the Route Reflector propagates the VTEP list with their respective VNI to all VTEPs it knows.
3. Each VTEP obtains a list of its VTEP neighbors for each VNI.

Now that each VTEP has exhaustive knowledge of all existing VTEP neighbors and their relevant VNI, in the VxLAN domain H1 establishes communication with H4.

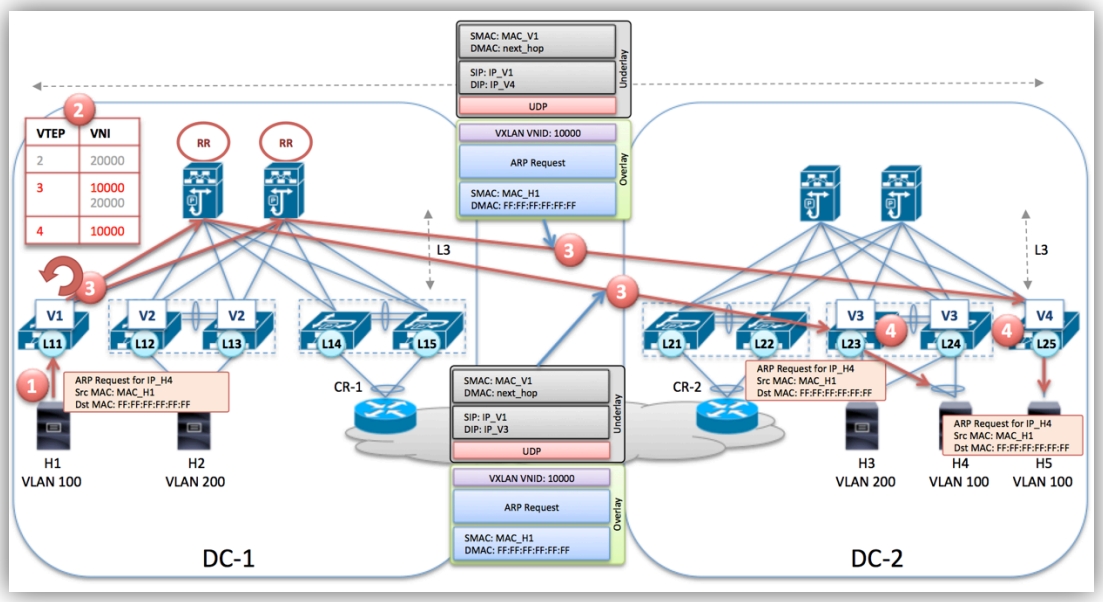


Figure 6: ARP request flooded across the DCI

1. H1 ARP's for H4. The source MAC is therefore H1, and the destination MAC is FF:FF:FF:FF:FF:FF.
2. H1 is attached to VLAN 100, which is mapped to VNI 10000. The local VTEP V1 does a lookup in its VTEP table and localizes all the VTEP

neighbors responsible for the VNI 10000. In this example, VTEP 3 and VTEP 4 are both binding VNI 10000.

3. As a result, VTEP V1 encapsulates the original broadcast ARP request from H1 with the VxLAN header using VNI 10000 and replicates this broadcast packet toward VTEP V3 and V4. VTEP 1 is the source IP address used to construct the overlay packet with VTEP 3 and VTEP 4 as the destination IP address.
4. VTEP 3 and VTEP 4 both receive the VxLAN packet identified with VNI 10000, remove the VxLAN header and notify the Dot1Q tag (VLAN 100). As a result, NVE Leafs 23 and 25 forward, respectively, the ARP request to all the respective interfaces binding VLAN 100. H4 and H5 receive the ARP request. H5 ignores the request and H4 learns and caches the MAC address for H1.

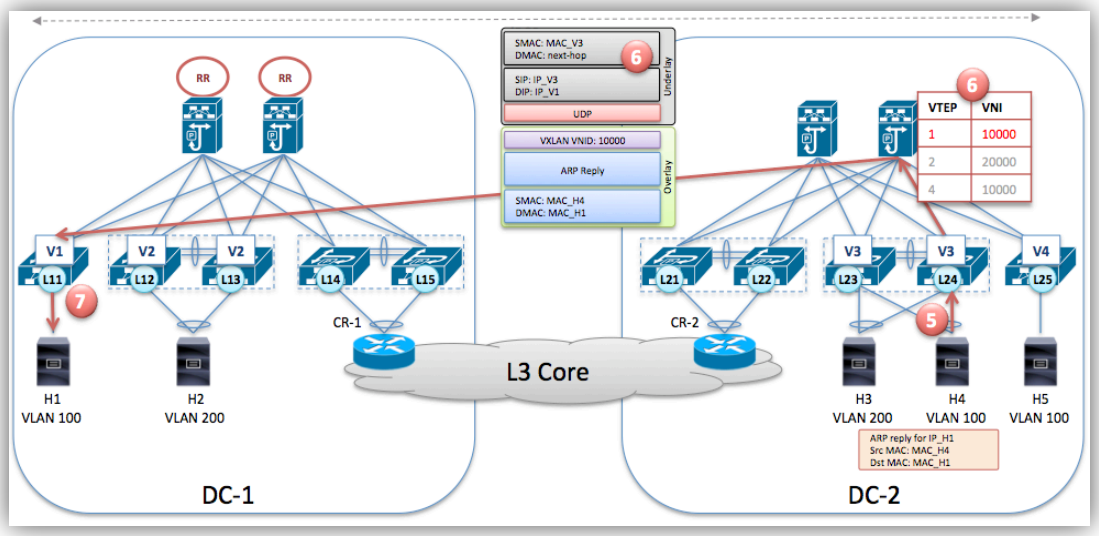


Figure 7: ARP reply Unicast

5. H4 can reply Unicast to H1 toward NVE Leaf 24. A LACP hashing algorithm is implemented across the vPC attachment and the reply will hit one of the two vPC peer-NVE. It doesn't matter which Leaf receives the frame as both VTEP vPC peer-switches share the same VTEP table entries.
6. VTEP 3 (NVE Leaf 24) notifies the response frame from H4 for a binding to VLAN 100 and consequently encapsulates the ARP reply with the identifier VNI 10000. Therefore, it sends it Unicast toward the VTEP 1 IP address as it knows that the destination MAC of the ARP reply is locally connected to VTEP 1.
7. VTEP 1 receives the VxLAN packet with VNI 10000 from VTEP 3, strips off the VxLAN header and forwards the original ARP reply to H1.

Key takeaways

As a result, we can note that even though the transport is Unicast-only mode, (Head-end replication), the learning process for host information is not optimised. The source VTEP knows the remote VTEPs for each VNI of interest. But it still needs to learn the destination MAC which relies on Data Plane-learning using Flood & Learn. Consequently, Multicast and Unknown Unicast, including ARP requests (BUM), are flooded using Head-end replication to the

remote VTEPs. ARP suppression cannot be enabled as the host reachability process relies on Flood&Learn. Risks of disrupting both data centers due to Broadcast Storm is high.

In addition, it is important to note that this implementation in general doesn't used a standard Control Plane approach such as MP-BGP. Consequently it may not allow for an independent Control Plane per site, hence the reason I used a single CP stretched across the two sites.

From a standard point of view, VxLAN does not offer any native L2 loop protection. Consequently, a manual configuration of the BPDU Guard on all network interfaces must be set.

Last but not least, Anycast L3 Gateway for network overlay is usually not supported with this mode. As the result, "hair-pinning" or "ping-pong" effect may seriously impact the performances of the applications.
In short, be aware that having a Unicast-only mode with Head End Replication doesn't mean that it is a valid DCI solution.

VxLAN Phase 2.0: VxLAN/EVPN MP-BGP Control Plane

This is an important enhanced method of VxLAN with Unicast transport leveraged for the learning process. In addition to the VTEP information discussed above, all hosts are now dynamically discovered and distributed among all VTEPs using a BGP control Plane pushing the end-point information toward all VTEPs, reducing massively the amount of flooding for the learning method.

This is achieved using MP-BGP based on EVPN NLRI (Network Layer Reachability Information), which carries both Layer 2 and Layer 3 reachability reports. This is an IETF draft (https://tools.ietf.org/html/draft-sd-l2vpn-evpn-overlay-03).

> *VxLAN Bridging and routing services in the VxLAN overlay may exist on each Leaf. Bridging is a native function that comes with the hardware-based VTEP platforms. It is required to map a VLAN ID to a VxLAN ID (VNI). Distributed Anycast routing service is a key element needed to reduce the hair-pinning effect in a distributed virtual fabric design. This is treated in a separate section.*

In the previous scenarios, a single Control Plane has been responsible for both fabrics. This implies two consequences:

- First of all, the same autonomous system (AS) is spread over the two sites, the internal BGP must connect and maintain the session with all other BGP peers in a full mesh fashion (where everyone speaks to everyone directly). In DCI infrastructure, when establishing a full-mesh between all VTEP spread across the multiple sites, this number of sessions may degrade performance of routers, due to either a lack of memory, or too much CPU process requirements.
- Secondly, for resiliency purposes, it is preferable to offer an independent Control Plane (Figure 2: Global architecture for Dual-Sites), thus if one CP fails in one location, the other site is not impacted.

Therefore, in the following part, the VxLAN Data Plane stretched across the two locations is built with two independent Control Planes aligned with the DCI requirements. Let's re-use the same scenario elaborated previously, this time with an independent Control Plane inside each fabric and eBGP EVPN peering establishment on the WAN side. The next section provides an overview of how both Control Planes will be updated by each other. That is followed by communication between two end-nodes.

Host learning process with independent control planes

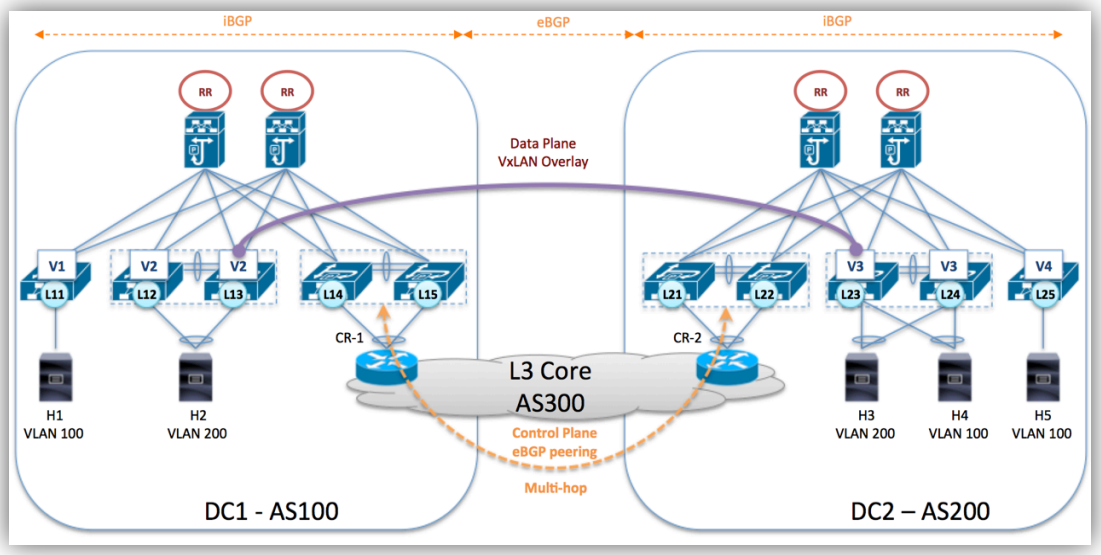


Figure 8: VxLAN eBGP Control Plane

iBGP EVPN AF is used for the control plane to populate the host information inside each fabric. MP-BGP (eBGP) peering is established between each pair of Border-Leafs, allowing for different AS per location.

The multi-hop function allows eBGP peering to be achieved over intermediate routers (e.g., ISP) without the need to treat EVPN *per se*.

The same VxLAN overlay network, established from end to end across the two Data Centers, represents the Data Plane.

Packet Walk

If we replay the same scenario as before with Phase 1.5, we can see the efficiency gained with the BGP EVPN AF Control Plane.

In the following design, an independent Control Plane is initiated for each site using iBGP. Subsequently, independent BGP Route Reflectors are initiated on both fabrics.

All VTEPs advertise Host Routes (IP and MAC identifiers) for their hosts that exist in their Network interfaces to their respective Control Plane BGP. Any VTEP establishes an iBGP peering with each BGP Route Reflector.

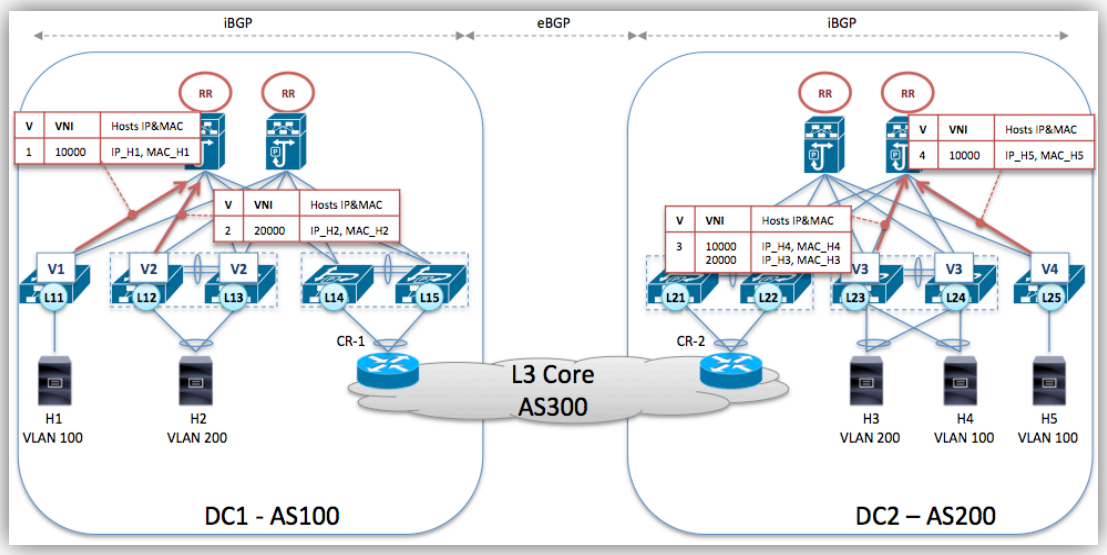


Figure 9: each NVE updates the CP with their hosts

Subsequently, each BGP Route Reflector knows all local VTEP and host information that exists in its fabric. The Route Reflector therefore can distribute to all its local VTEPs the required host information that it is aware of; however, it needs first to update its BGP peer in the remote location, and it expects as well to be updated with the remote VTEP/host information.

In order for the BGP Route Reflector to populate its local host information table with their respective VTEP, VNI, IP and MAC addresses to the remote fabric, two main options are possible:

- Route Reflector peering
- Border-Leaf peering

BGP Route Reflector Peering

The BGP neighbor is specified for the spine and eBGP will be initiated, as two different AS exist.

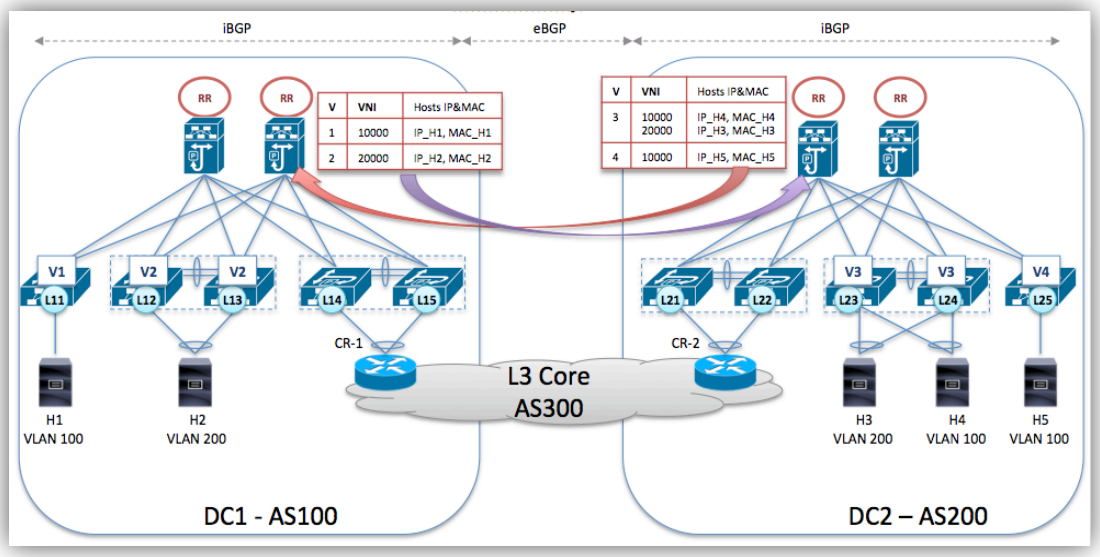


Figure 10: Route Reflector Peering

Each BGP Route Reflector can now populate its local host table toward the remote Control Plane.

BGP Border-Leaf peering:

The BGP next hop for each spine (RR) is the local pair of Border-Leafs connecting the Core layer. eBGP peering is established between each Border-Leaf pair toward the core routers. All transit routers in the path (Layer 3-managed services via SP) do not require any knowledge of EVPN. To achieve this transparency, eBGP peering is configured to support multi-hop.

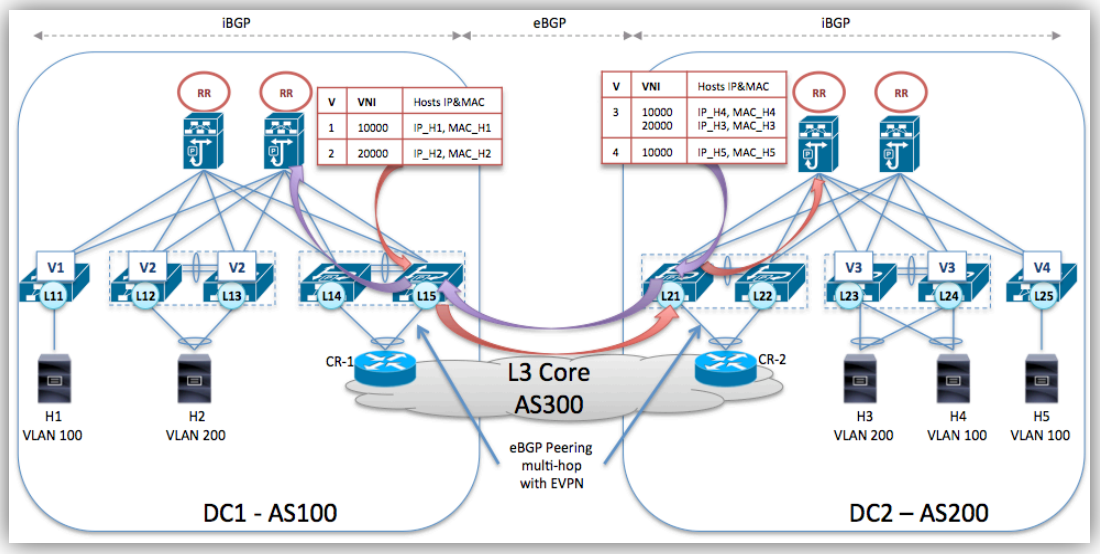


Figure 11: eBGP peering from the Border-Leaf

The BGP Route Reflector will relay the host information distribution toward the Border-Leaf, which will forward to its peer eBGP using EVPN AF.

Global Host information distribution

Each Control Plane has now collected all host information from its local fabric as well as from the remote fabric.
Each control plane needs straightaway to populate its local VTEPs with all host identifiers (IP & MAC addresses) that exist across the two fabrics.

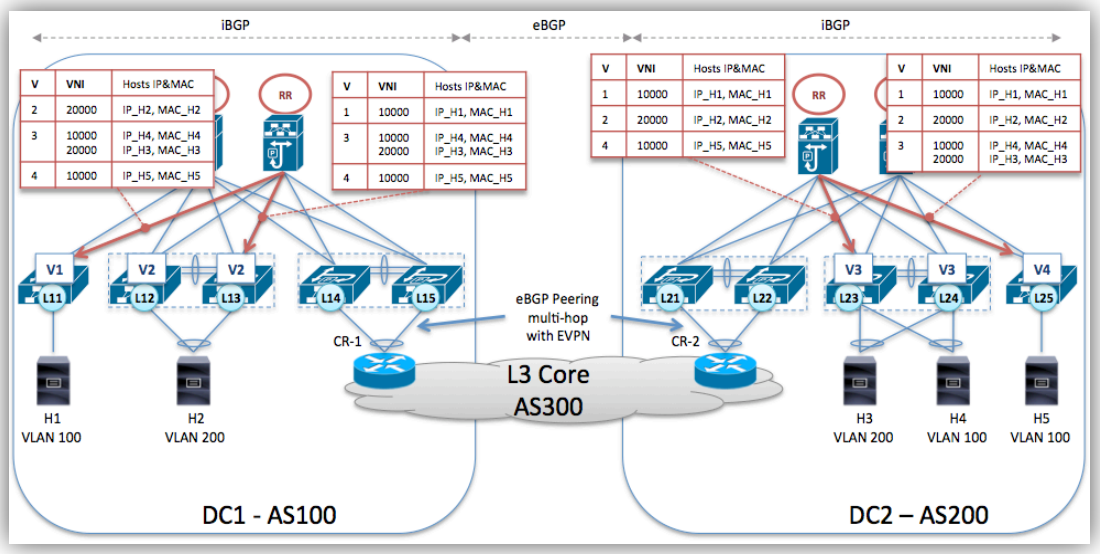


Figure 12: Global host information population

ARP request

Each VTEP has an exhaustive knowledge of all existing local and remote hosts associated to locally active VNIs, including all other VTEPs, with their relevant host information, VNI, IP and MAC addresses.

The VTEP neighbor information is distributed accordingly, as described in the previous phase.

H1 wants to establish communication with H4.

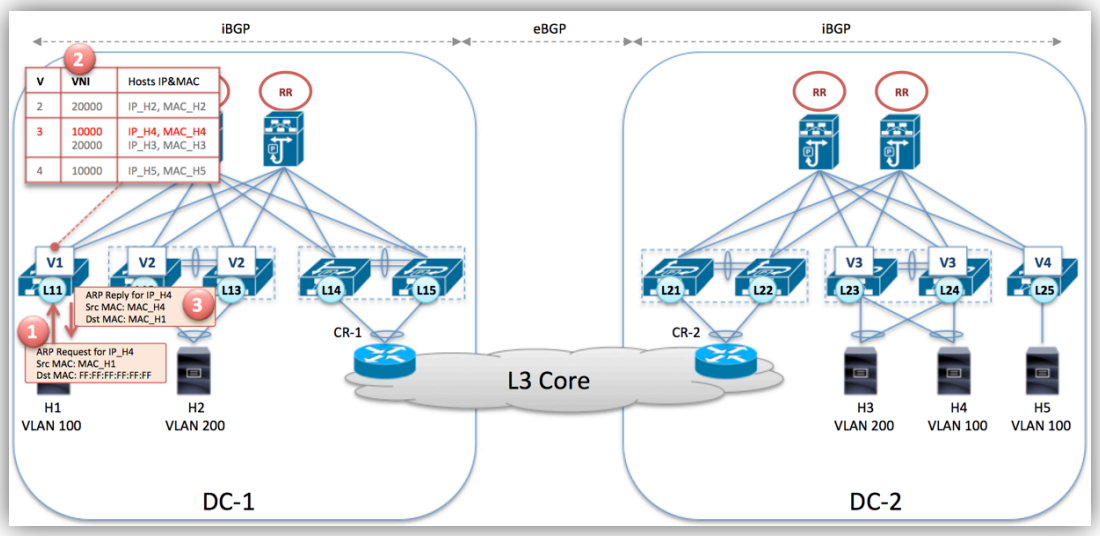


Figure 13: ARP Suppression

1. H1 ARP's for H4. The source MAC is therefore H1, and the destination MAC is FF:FF:FF:FF:FF:FF.
2. Leaf 11 gets the ARP request from H1 and notes that this frame must be mapped to the VNI 10000. VTEP 1 checks against its ARP suppression cache table and notices that the information for H4 is already cached.
3. VTEP responds Unicast to H1 with the MAC address for H4.

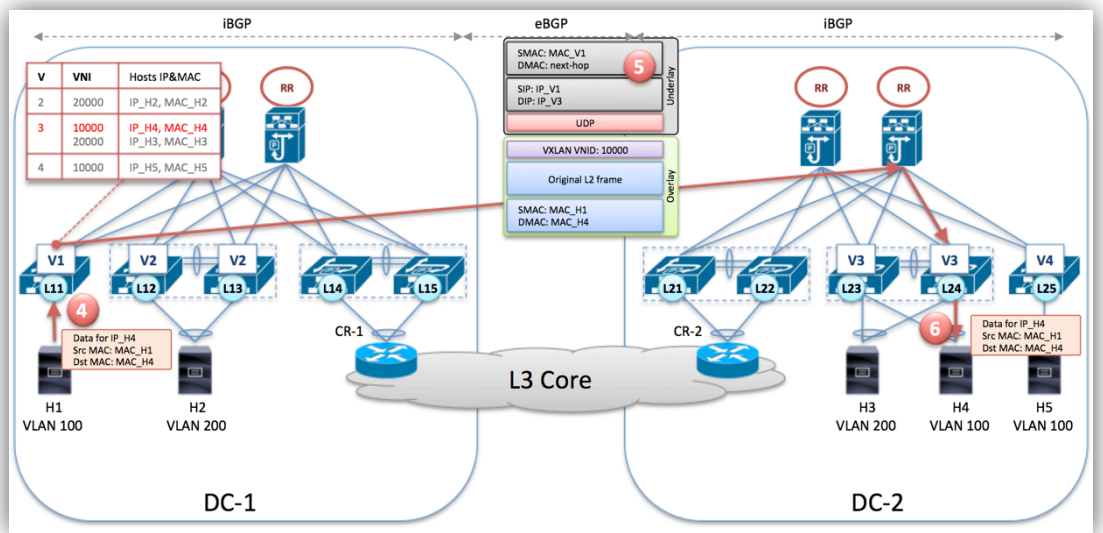


Figure 14: BGP EVPN AF and Unicast traffic

As a result, the ARP request is suppressed over and outside the fabric.

4. Host H1 can send a Unicast frame to host H4.
5. NVE Leaf 11 receives the Unicast Layer 2 frame destined for H4 over VLAN 100 and observes this flow requires it to be encapsulated with VNI 10000 with the IP destination VTEP 3, which it knows.

- The Unicast packet is load balanced across the fabric and ECMP determined the path to hit the Anycast VTEP address (V3) available on both vPC peer-switches.

6. VTEP 3 on Leaf 24 receives the frame, strips off the VxLAN header and performs a MAC lookup for H4 in order to retrieve the interface attaching H4 (e.g., E1/2). It then forwards the original frame to H4.

Thus, after each VTEP has been notified and updated with the information about all the existing hosts (VTEP, VNI, Host IP & MAC), when a Leaf sees an ARP request from one of its network interfaces (VLAN), it performs a ARP snooping on behalf of the destination, responding directly to the ARP request with the destination MAC address cached in its local table, for the local host originating the request. Accordingly, the amount of flooding for Unknown Unicast is highly reduced as ARP is not flooded when the destination of the ARP request is known.

This is assuming that all hosts have been dynamically discovered and distributed to all VTEPs. However, for silent hosts and non-IP protocol, UU are still flooded.

Silent host

Now, let's pretend we have a silent host (H6) as a target destination for a new communication. Consequently, its directly attached ToR (L25) is unaware of its existence and therefore the NVE function of the Leaf 25 cannot notify the BGP Control Plane about the existence of host H6.

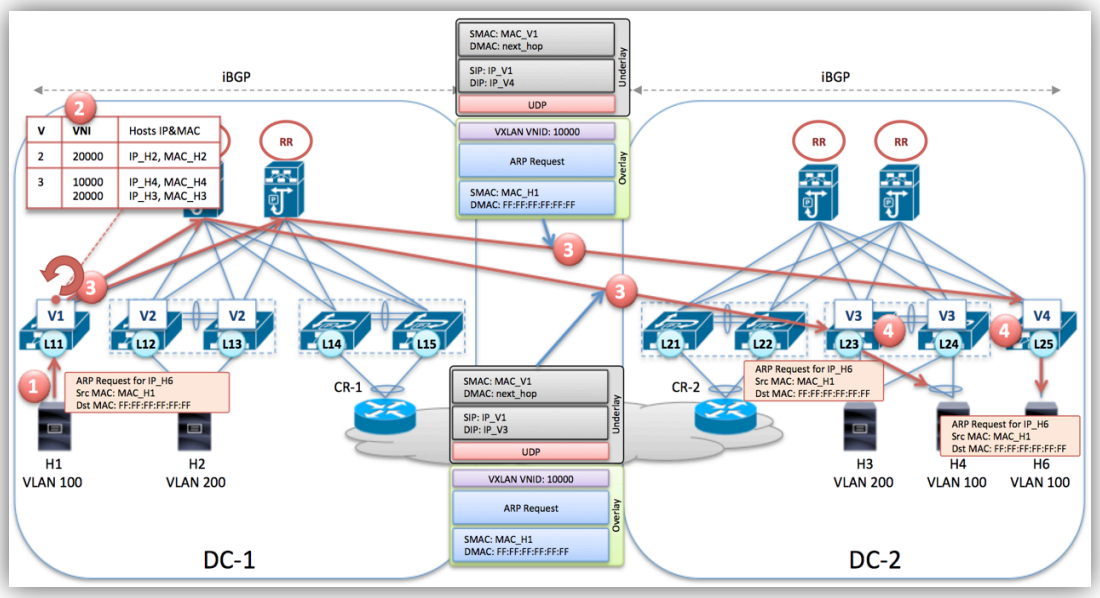


Figure 15: BGP EVPN and silent host (H6)

1. From the above scenario, Host H1 wants to communicate with H6 but has no ARP entry for this host. As a result, it "ARP requests" for H6.
2. The NVE Leaf 11 does a lookup on its VTEP/Host table but gets a "miss" on information for host H6.

3. Subsequently, VTEP 1 encapsulates the ARP request with VNI 10000 and replicates the VxLAN packet to all VTEPs that bind the VNI 10000 (HER).
   Note that if IP multicast is enabled, it can also forward the ARP request using a multicast group.
4. VTEP 3 and VTEP 4 receive this ARP request packet flow, respectively strip off the VxLAN header and flood their interfaces associated to VLAN100. H4 receives but ignores the L2 frame; however, H6 takes it for further treatment. Depending on its OS, H6 caches the host H1 information in its ARP table.

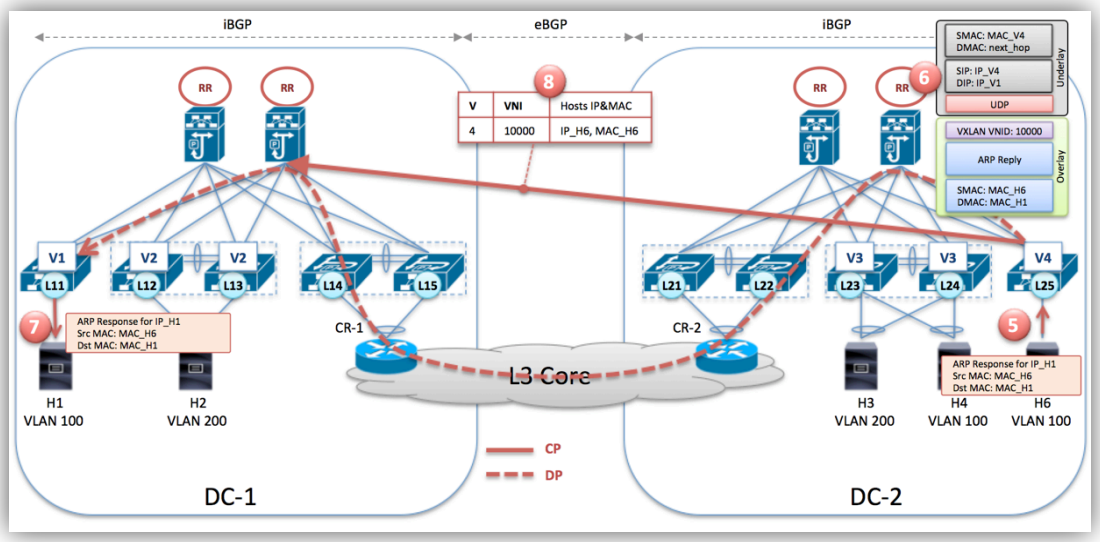


Figure 16: BGP EVPN AF Silent host update to BGP

5. Host H6 replies with its ARP response toward Host H1.

Two actions are completed in parallel:

Action 1: Data Plane transport

6. VTEP 4 encapsulates the ARP reply with the VNI 10000 and sees on its VTEP host table that H1 belongs to VTEP 1. VTEP 4 forwards Unicast the VxLAN response to VTEP 1.
7. VTEP 1 receives the packet from VTEP 4, strips off the VxLAN header, does a MAC lookup for Host 1 and forwards the response Unicast to H1 through the interface of interest (e.g., E1/12). H1 populates its local ARP cache.

Action 2: Control Plane host learning and population

8. In the meantime, VTEP 4 populates its BGP Control Plane with the new host route for Host 6.

As soon as the local Route Reflector receives a new host entry, it updates its BGP neighbor. Both BGP Control Planes are now up-to-date with the new entry for Host H6.

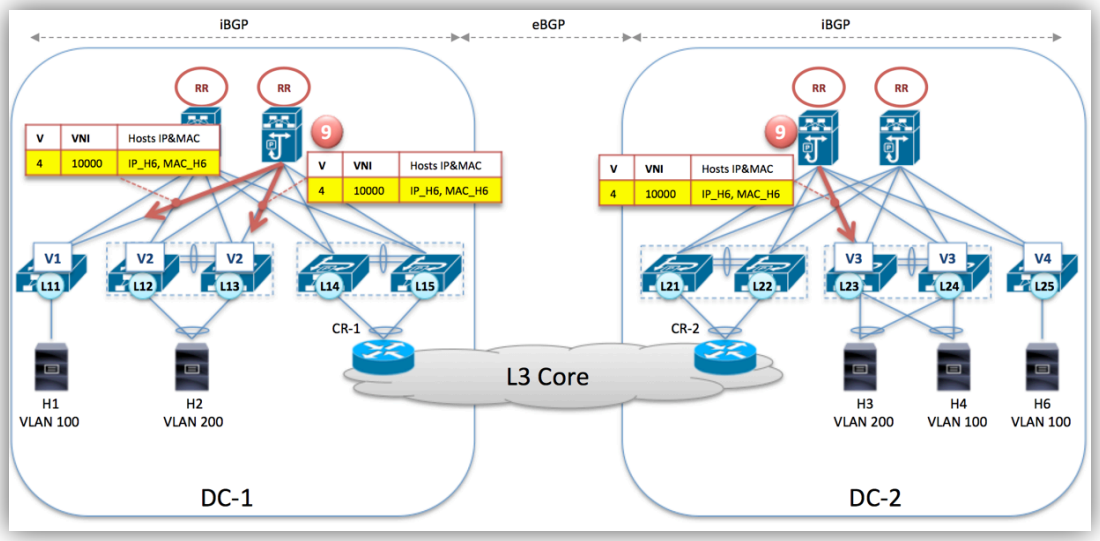


Figure 17: BGP EVPN AF Silent Host Population

9. The BGP route reflector updates all VTEPs with the new host route H6.

In a nutshell, if there is a Unicast RIB "miss" on the VTEP, the ARP request will be forwarded to all ports except the original sending port (ARP snooping).
The ARP response from the silent host will be punted to the Supervisor of its VTEP with the new host information, and subsequently populates the BGP Control Plane with its Unicast RIB (learning). The ARP response is therefore forwarded in the Data Plane.

Note: Inside the same Fabric, with MP-BGP EVPN control plane and Host discovery and distribution, it is certainly more appropriate to enabled Multicast to handle the BUM traffic, thus reducing the Ingress replication (IR) for a better scalability. In the above DCI scenario, Ingress replication is used to carry the BUM inside each site and across the two locations. The operational reason is that IP Multicast for Layer 3 managed services (L3 WAN) is not always an option for the Enterprise and we should consider the scenario that fits in all circumstances. However, if IP Multicast is available, the ARP request or any BUM can be transported as well inside a Multicast within the core network.

Host mobility

Host mobility should respond to a couple of requirements:
1 - Detect immediately the new location of the machine after its move and notify the concerned network platforms about the latest position.
2 - Maintain the session stateful while the machine moves from one physical host to another.

Traditionally when a station is powered up, it automatically generates a ARP sending its MAC address to the entire network. The CAM table of all the switches

belonging to the same layer 2 broadcast domain is updated accordingly with the known outbound interface to be used to reach the end-node. This allows any switches to learn dynamically the unique location of the physical devices.

For virtual machine mobility purposes, there are different ways to trigger an notification after a live migration, informing the network about the presence of a new VM.
When running Microsoft HyperV, Windows will send out a Gratuitous ARP (GARP) after the migration of the guest virtual machine has been successfully achieved. With VMware vCenter (at least up to ESX 5.5) a Reverse ARP (RARP) is initiated by the hypervisor sent toward the upstream switch (Leaf). Note that RARP has no layer 3 information (actually, it asked for it, like DHCP does).

With the original VxLAN Flood & Learn (Multicast-only transport and Unicast-only Transport with HER), the GARP or RARP are flooded in the entire overlay network and the learning process is achieved like within a traditional Layer 2 broadcast domain.

However, with VxLAN/EVPN MP-BGP Control Plane, the new parent switch detects the machine after the move and this information is populated consequently to all VTEP of interest.

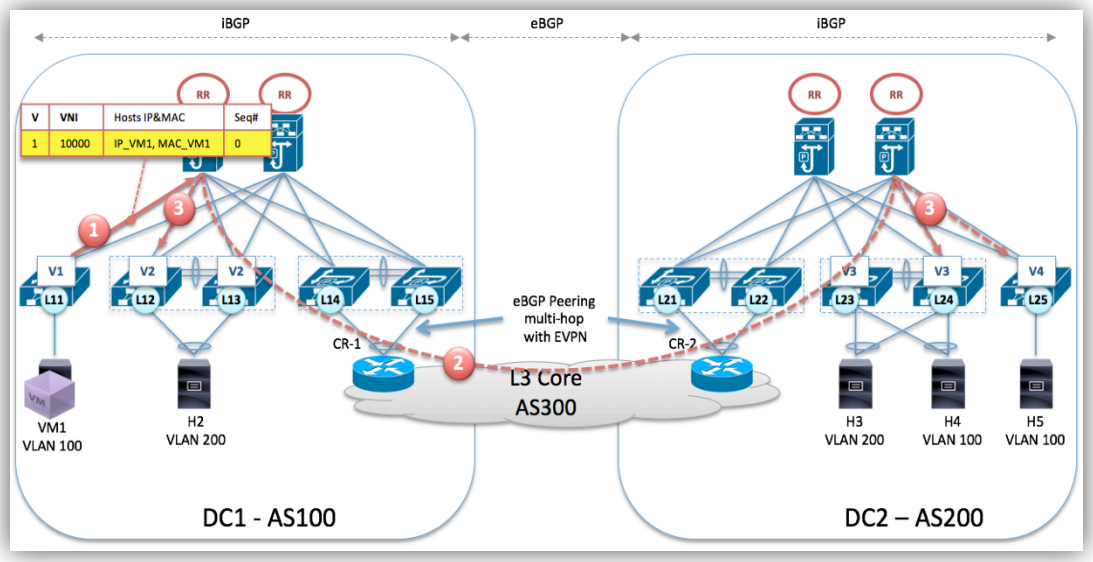


Figure 18: Host Advertisement: Sequence number 0

In the example above, VM1 is enabled and "virtually" attached to its layer 2 segment. As the result, VTEP 1 detects its presence and advertises its BGP Route Reflector about its location.

1. VTEP 1 notifies its local BGP Route Reflector with a sequence number equal to "0" informing that it owns VM1.
2. The BGP RR, updates its BGP RR peer about this new host. As the result, both BGP Control plane are aware of the host VM1 and its parent VTEP.
3. Both RR in each site, update their local VTEPs about VM1.

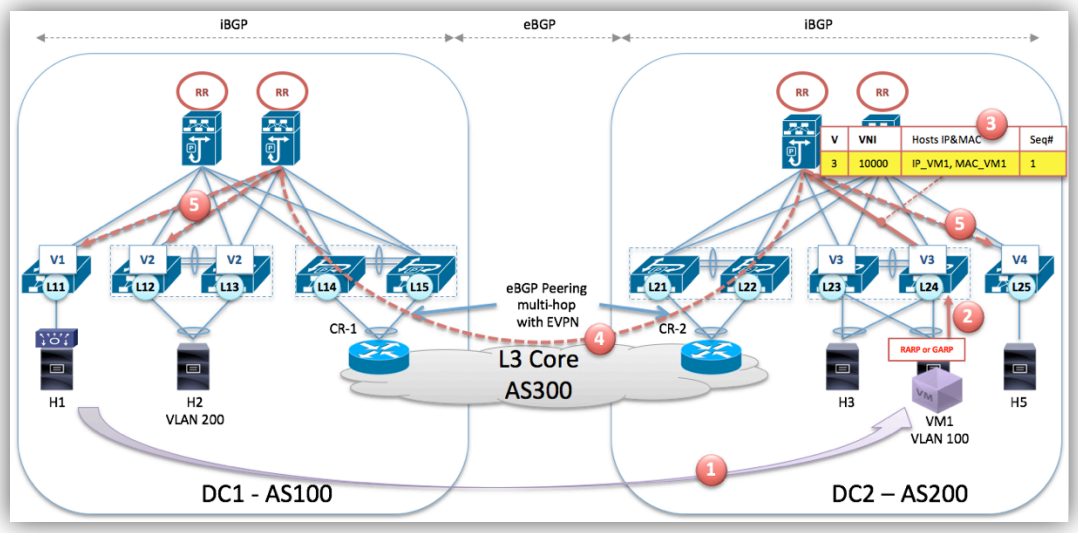


Figure 19: Host move, new advertisement using a Sequence number 1

VM1 migrates from its physical Host 1 (H1) in DC1, across the site interconnection up to the physical Host 4 (H4) located in DC2. We assume the port group in the virtual switch is configured accordingly.

1. After the move of VM1, the hypervisor sends a RARP or GARP.
2. The Leaf 23 or 24 (depending on the LACP hashing result) detects the new presence of VM1.
3. VTEP 3, which knew VM1 from a previous notification as being attached behind VTEP 1, sends an update of new host VM1 using a sequence number equal to "1" in order to overwrite the previous host information for VM1.

   If it's a RARP (vMotion), the host IP address is unknown at this period of time, consequently the association IP ⇔ MAC is deducted from the local VTEP host table.
4. The BGP RR in DC2 updates its RR peer in DC1
5. Subsequently both BGP RR update their respective local VTEPs about VM1 with its new location behind VTEP 3

VTEP 1 sees a more recent route for VM1 and will update his routing table.

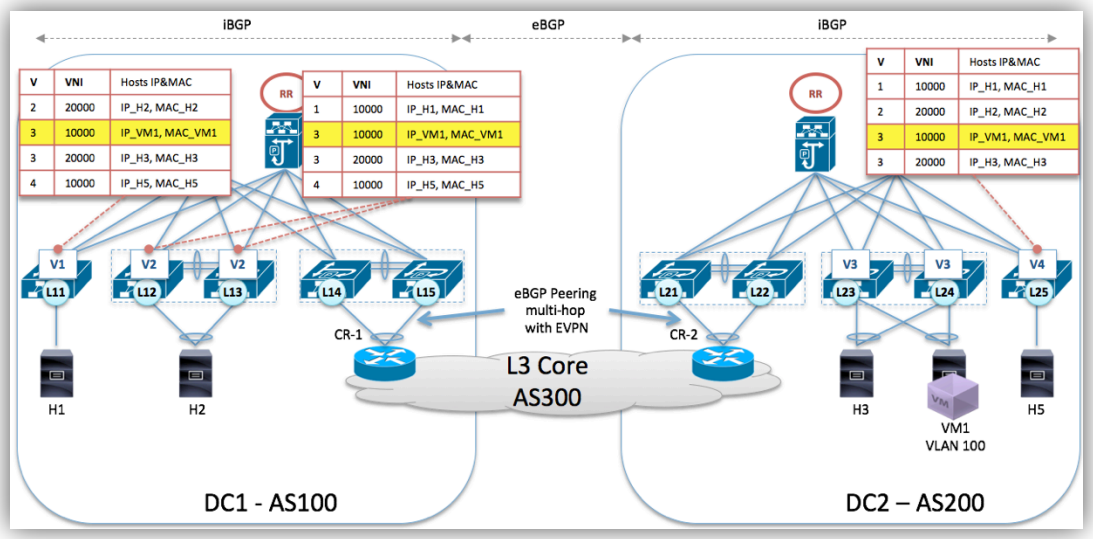


Figure 20: All VTEPs are updated with the new location for VM1

All VTEPs are updated with the new location VTEP 3 for VM1

Distributed Default gateway

In a DCI environment it is important to take into consideration the bandwidth consumption as well as the increased latency for E-W traffic across the multiple sites that may have a critical impact in term of application performances. To facilitate optimal east-west routing reducing the hair-pinning effect, while supporting transparent virtual machine mobility without interruption, the same Default Gateways must exist on each site, ensuring that the redundant first-hop routers exist within each data center.

There are two options to achieve this Layer 3 Anycast function of the default gateway on both sites:

- The 1st option is to enable a pair of FHRP A/S gateways on each site and manually filter the FHRP multicast and virtual MAC addresses to not be propagated outside the site. As the result each pair of FHRP routers believe they are alone and become active on each location with the same virtual IP and MAC addresses. When a VM migrates to the remote location, it continues to use the same default gateway identifiers (L2 & L3) without interrupting the current active sessions. The drawbacks of this option are that it requires a manual configuration at the DCI layer (e.g. FHRP filtering) and it limits the number of active default gateway per site to one single device (per VLAN). This option should be applicable to any VxLAN mode.

- The second option is to assign the same gateway IP and MAC addresses for each locally defined subnet in each Leaf. As the result, the first hop router for the servers exists across all Leaf switches for each VLAN of interest.

Note that this feature is compatible with traditional DCI solutions such as OTV or VPLS as it occurs on each ToR.

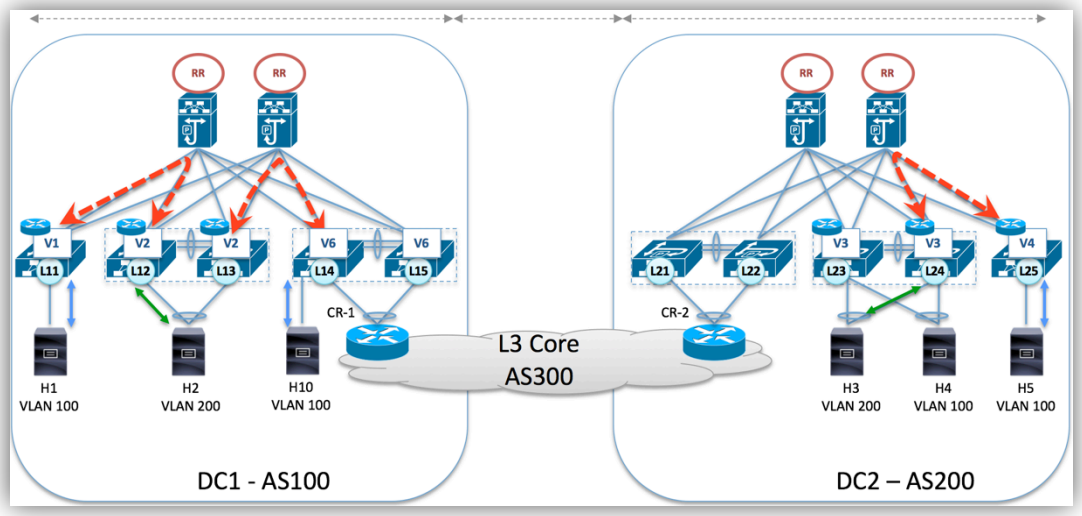


Figure 21: Anycast Layer 3 Gateway

In the above example, in DC1, H1 from VLAN 100 wants to communicate with H2 on VLAN 200. H10 from VLAN 100 wants also to speak with H2. In the meantime in DC2, H5 on VLAN 100 wants to connect with H3 on VLAN 200. They all use the same Default gateways for their respective VLAN 100. In a traditional deployment, all the routed traffic would hit a single centralised layer 3 gateway. With distributed L3 Anycast gateway, the traffic is directly routed by the first hop router initiated in each ToR.

- H1 sends the packet destined to its DF (VLAN 100), routed and encapsulated with the concerned VNI by L11 toward VTEP 2. L12 and L13 are de first hop routers for H2 in regard to VLAN 200.

- H10 sends the packet destined to the same DF (VLAN 100) with the same identifiers used by H1 toward H2 that sits behind VTEP 2.
- H5 sends the packet destined to the same DF with the same identifiers used by H1 toward H3 on VTEP 3.

As the result, although in this example all source use the same default gateway, the routing function is distributed among all ToR increasing VxLAN to VxLAN L3 routing performances, reducing the hair-pinning and therefore the latency improving the performances on multi-tier applications. Having the same gateway IP and MAC address helps ensure a default gateway presence across all leaf switches removing the suboptimal routing inefficiencies associated with separate centralised gateways.

Key takeaways for VxLAN MP-BGP EVPN AF and DCI purposes

Although the VxLAN Control Plane MP-BGP EVPN has not been originally thought out for DCI purposes, it could be diverted to provide a “good enough” DCI solution, specifically with VTEP and Host information dynamically learnt and

populated, in conjunction with ARP suppression reducing UU across the two sites. On the resilient form of the architecture, redundancy tunnel end-points are addressed with a Cisco vPC peer switches. However, we must be cautious about some important shortcomings and risks for the DCI environment that need to be clarified and highlighted to Network Managers.

Some tools can be added manually, such as L2 loop detection and protection, while some others are not available yet, such as selective BUM or Multi-Homing discussed afterward.

Dual-homing for resilient edge devices

Different solutions exist offering Dual-homing from a Layer 2 switching point of view, such as vPC or MLAG or VSS or nV. The tunnel created for the network overlay requires to be originated from a VTEP. The VTEP resides inside the Leaf mapping VLAN IDs to a VxLAN IDs. However it is crucial that the VTEP is always Up & Running for business continuity. The Host-based VTEP gateway supports an Active/Standby HA model. When the active fails, it may take some precious times before the standby VTEP takes over. To improve the convergence, transparency and efficiency, Cisco Nexus platforms support Active/Active VTEP redundancy by allowing a pair of virtual Port-Channel (vPC) switches to function as a logical VTEP device sharing an Anycast VTEP address. The vPC peer switch is leveraged for redundant host connectivity, while each device is individually running Layer 3 protocol with the upstream router in the underlay networks: Fabric layer and Core layer.

The intra-fabric traffic is load-balanced using ECMP (L3) among the 2 VTEPs configured with the Anycast VTEP address, while the vPC members in the network side, load distribute the traffic using an Ether-Channel load-balancing protocol such as LACP (L2).

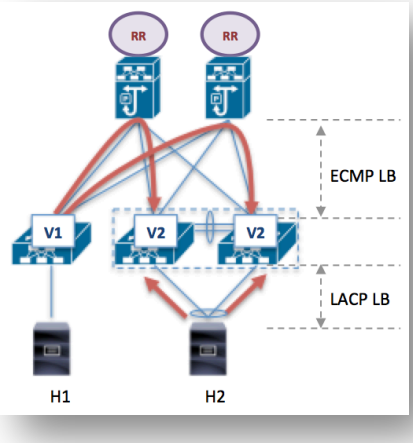


Figure 22: vPC peer-switch & Anycast VTEP address

For the VxLAN overlay network, both switches use the same Anycast VTEP address as the source to send the encapsulated packets. For the devices dual-attached in the underlay network, the two vPC VTEP switches appear to be one logical VTEP entity.

For VxLAN Data Plane as well as BGP Control Plane traffic, the Layer 3 is not directly concerned with vPC peering, as it is pure Layer 3 communication (Layer 3 intra-fabric to Layer 3 Core communication). Consequently, the Layer 3 Data Plane (VxLAN) is routed across L3 Border-Leaf devices. As a traditional standard routing protocol, it natively supports a L3 load-balancing algorithm such as ECMP or PBR via the two Layer 3 Border Leafs. If one DCI link or DCI device fails, traffic will be re-routed automatically to the remaining device.

Having said that, and contrary to the scenarios described previously, the Border-Leaf is certainly not being just dedicated to route the network fabric to the outside world but is also going to be used as a ToR offering device dual-homing connectivity (host, switches, network services, etc.). That gives the flexibility to use a pair of Border-Leafs either for DCI connectivity (Data Plane and Control Plane) only, or use the Border-Leaf for both DCI connectivity as well as vPC VTEP for dual-homed devices with regard to VLAN to VNI mapping. The last use-case is certainly the most deployed.

The following figure shows each VxLAN overlay network (VNI) routed toward the core layer. A pair of Border-Leafs is used as redundant Layer 3 gateways for DCI purposes only. The VxLAN tunnel is initiated and terminated in the remote NVE of interest.

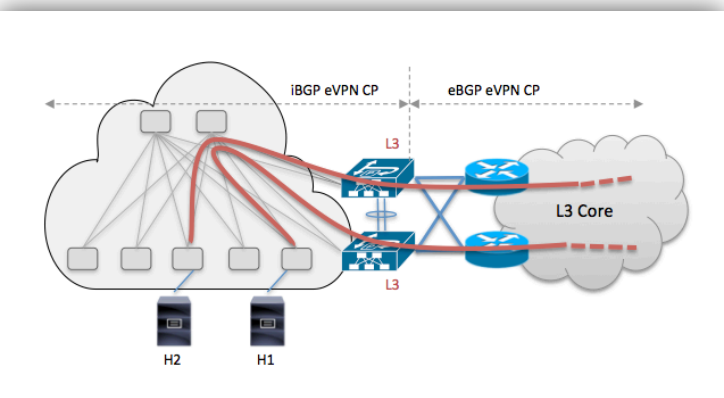


Figure 23: Border-Leaf L3 DCI only

In the following figure, the Border-Leaf vPC peer-switches are used for two functions. The first role is used to offer redundant Layer 3 gateways for DCI purposes. The second role is to initiate the VxLAN tunnel through its local VTEP, independent of the DCI functions.

Anycast VTEP gateway is leveraged for redundant VTEP, improving the resiliency of the whole VxLAN-based fabric as the function of redundant VTEP gateway is distributed. If one fails for any reason, it will impact only the physical ToR concerned by the failure, but traffic is just immediately sent via the remaining VTEP on the data-plane level. In addition, it allows improving performance (the same VNI packets can be distributed between the two VTEPs).

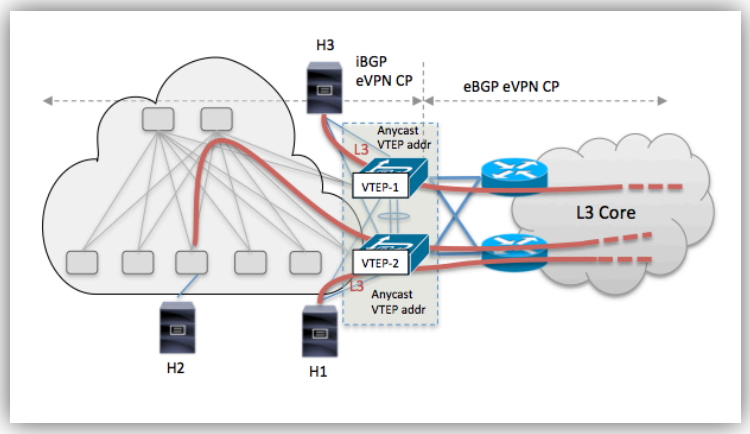


Figure 24: Border-Leaf vPC VTEP and L3 DCI

VxLAN deployment for DCI only

Another case study exists when VxLAN protocol is diverted to only run a DCI function while traditional Layer 2 protocols such as STP, vPC, FabricPath or others are deployed within each Data Center.

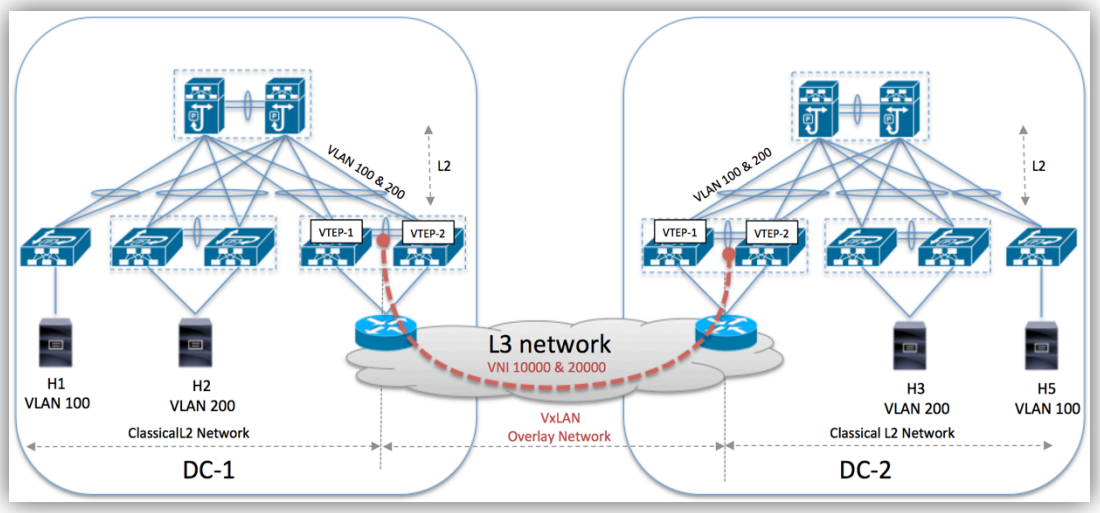


Figure 25: VxLAN for DCI only

This design above looks pretty much like a traditional DCI solution in the sense that tunnels for the overlay networks are initiated and terminated at the DCI layer for DCI purposes only.

Each Fabric uses a traditional Layer 2 transport. There is no fabric dependency, one fabric can run vPC/RSTP, and the other can be FabricPath. All VLANs that require to be stretched across multiple sites are extended from each ToR concerned by those VLAN, toward the pair of Border-Leafs used for the DCI function. These VLANs are mapped to their respective VNI at the Border Leaf (DCI layer) supporting the VTEP: the VxLAN configuration, including the secondary VTEP address, is similar to both VTEP vPC peer-switches. The VTEPs

on each site initiate the overlay network toward the remote VTEPs in a redundant configuration.

If one VTEP device fails, the other takes over immediately (vPC Layer 2 and Anycast VTEP address convergence).

VLAN and VNI selection for DCI purposes

It is usually recommended to extend outside a fabric, only a set of required VLAN, not all of them. It is not rare to deploy 1,000 or more VLANs within a Data Center, but maybe only 100 of them are needed in the remote sites.

In a conventional DCI solution, as was noted previously, the VLANs are selectively encapsulated into an overlay network (e.g. OTV, VPLS) at the DCI edge layer. Hence, it's easy to select and configure the set of VLANs that are required to be extended.

On the other hand, with VxLAN, it is possible to control the number of VLANs to be stretched across the fabrics by using a mismatch between VLAN ID ⇔ VNI mapping. Only VLAN mapped to the same VNI will be extended through the Overlay network.

For example, in DC-1, VLAN 100 is mapped to VNI 10000, and in DC-2, VLAN 100 is mapped to VNI 10001. As a result, VLAN 100 will never be stretched across the two fabrics.
Instead, with ToR-based VxLAN VTEP gateway, the VLAN is local significant, subsequently, this allows VLAN translation natively, by mapping different VLAN IDs to the same VNI.

BFD and fast convergence

To offer fast convergence, one function required is fast failure detection (e.g., BFD, route tracking). However, another crucial function is a host/MAC reachability update (flush, discover, learning, distribution, etc.).

The function of host reachability on BGP EVPN relies on a centralized BGP Route Reflector that contains all the information required to forward the VxLAN packet to the desired destination. Both RR are synchronized and contain the same list of VTEP/Host entries. If one RR fails, the other takes over the entire job. One main difference with the traditional DCI overlays such as OTV, VPLS or PBB-EVPN, is that the overlay tunnel is initiated at the Leaf device (ToR) and terminates on the remote Leaf. The DCI layer is just a Layer 3 gateway and doesn't handle any VTEP/Host information (for the purpose of a DCI edge role). With the VxLAN overlay network, the encapsulation of the VLAN frame is distributed and performed in each Leaf/VTEP. Each Leaf supports its local redundancy design based on vPC, for dual-homed Ethernet devices.

It is important to note that BFD improves the detection of a remote failure in hundred of ms, allowing ECMP to select the remaining VTEP very quickly, in case of a issue (e.g. CPU) with one Leaf.

It is important that in any instances of a failure with a dual-fabric based on VxLAN MP-BGP EVPN, all host entries are maintained and up-to-date in the VTEP tables.

What is Needed for a Solid DCI Solution

The second section discussed previously, described the weaknesses of VxLAN Unicast-only mode with Head-end replication for VTEP. The third section described VxLAN with MP-BGP EVPN AF and has demonstrated a more efficient way to discover and distribute host information among all the VTEPs.

However, I mentioned that all of the requirements for a solid DCI deployment have not been met. Hence, let's reviews the shortcomings that should be addressed in the next release of VxLAN to be considered as a solid DCI solution.

Multi-homing

Sometimes mistaken for Dual-homing, Multi-homing goes beyond the two tightly coupled DCI edge devices. For the purpose of DCI, it aims to address any type of backdoor Layer 2 extension.

In the following design, the DC interconnection is established from the spine layer (this is mainly to keep the following scenario clear to understand; otherwise, it doesn't matter where the DCI is initiated).

A backdoor Layer 2 link is created between the two leafs, L15 and L21, meaning all the VLANs are trunked.

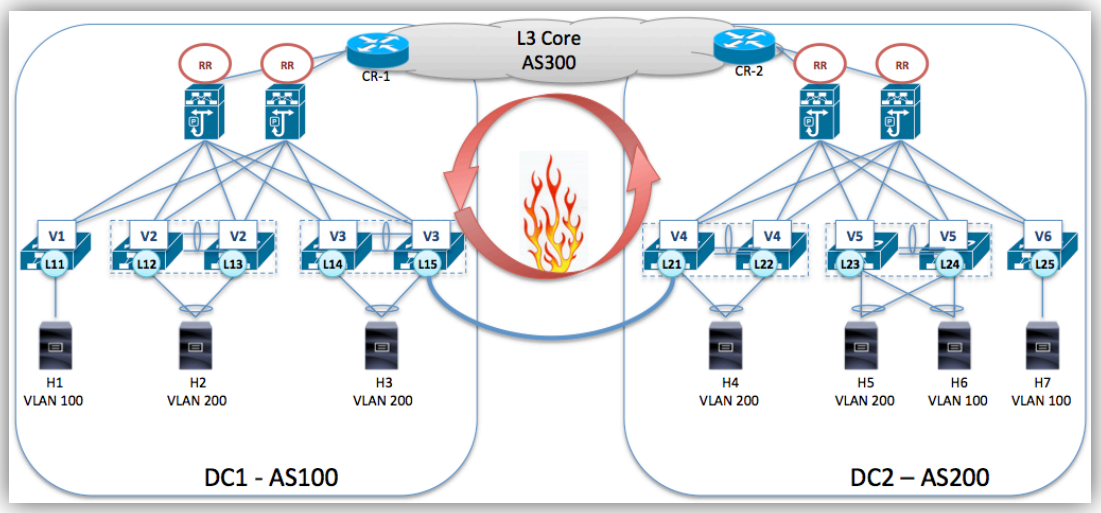


Figure 26: Backdoor Layer 2 link

As a result, on the Leaf directly concerned by the backdoor connection, the VLANs mapped to a VNI (e.g., VLAN 200), will face a large loop and a broadcast storm will be flooded through the VNIs of interest (Ingress replication) and from site to site. Consequently, the VNI will immediately saturate the original DCI service, taking all the available DCI bandwidth over the other VNI.

With the hardware-based VxLAN VTEP gateway, VLANs are local Leaf significant. Theoretically, only Leafs that bind the VLAN affected by the storm of BUM will be impacted. In the above design (Figure 26: Backdoor Layer 2 link), assuming only VLAN tag 100 exists in L11 and L25, they should not be directly impacted by the

storm created by the backdoor link, which here, concerns only VLAN 200. The reason is that each VLAN maps a different VNI and BUM will be flooded through the VNI of interest. However, L23 and L24, which support both VLAN 100 and VLAN 200, will be disrupted.

A behavior to consider is, if BUM is transported over a common Multicast group, the risk is that all VLANs may suffer from the storm.

From a standard point of view, VxLAN has neither an embedded multi-homing service nor L2 loop detection and protection tools. To avoid this situation, it is critical to protect against such L2 loop. This is also true inside the network fabric itself, for example, a wrap cable extended across two ToR within the same DC.

A workaround for Layer 2 loop protection

Consequently, the first necessity is to protect against potential Layer 2 loop. The most known feature for that purpose is certainly BPDU Guard. This feature is available and can already be enabled in the current implementation of VxLAN. BPDU Guard is a Layer 2 security tool that prevents a port from receiving BPDUs. It can be configured at the global or interface level. When you configure BPDU Guard globally to the Leaf, it is effective only on operational Edge ports. In a valid configuration (Host attachment), Layer 2 Edge interfaces do not receive BPDUs. If a BPDU is seen on one of the Edge Port interfaces, this interface signals an invalid configuration and will be shutdown, preventing a L2 loop to happen.

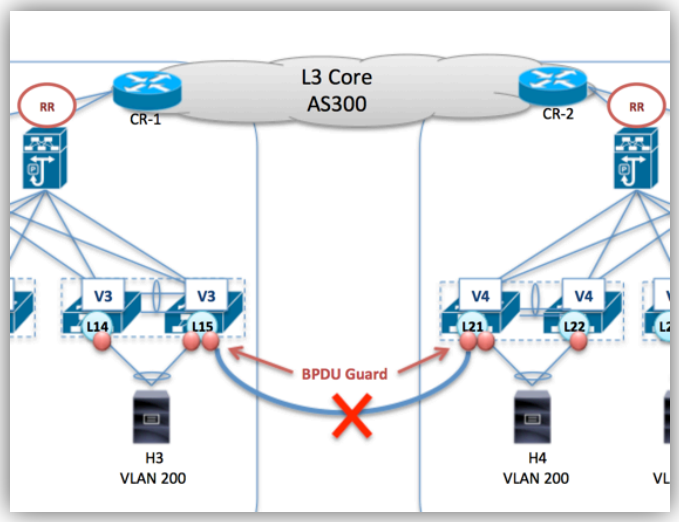


Figure 27: BPDU Guard

The drawback of BPDU Guard is that when cascading classical Ethernet switches running STP, it requires a special attention. BPDU Guard is disabled for trunk interfaces. Which is fine when connecting the traditional network, but it may add some complexity for operational deployments, as it implies some manual configuration.
The recommendation is to force all interfaces of the full ToR-based VxLAN to be configured in Edge Port and enable globally BDPU guard. BPDU Guard can be

disabled afterward per interface connecting traditional Layer 2 network.

Note that if the interface is trunked for attaching a hypervisor, then BPDU Guard is disabled; hence risks exist for a back-door link to be established.

Multi-homing and layer 2 loop detection improvement is required.

What is missing with the current implementation of VxLAN, is an embedded security feature that allows multi-homing while detecting and protecting end-to-end layer 2 loop. This doesn't necessarily supplant the BPDU Guard discussed previously, but brings an additional native service to protect against such a design mistake. This would also be used in site merging scenarios, where a back-door L2 exists between two fabrics, not due to a human mistake, but for example because of the migration from a traditional Data Center to Fabric with a VxLAN stretched end-to-end.

In Overlay network technologies such as VxLAN, managing BUM traffic is a key requirement. In the case where interconnection across multi-sites is multi-homed as described above, it is necessary that one, and only one edge device forwards the BUM traffic into the core or towards the remote Leaf when appropriate.

If by mistake two L2 segments are welded together (locally or across the two fabrics), an embedded function automatically detects the presence of another active device (e.g. sending and listening Hello's) for the same Ethernet segment and the election of a unique forwarder will be triggered immediately to choose one device responsible for BUM traffic.

OTV uses from ground-up a concept called Dual-Site Adjacency with a "Site VLAN" used to detect and establish adjacencies with other OTV edge devices, determining the Authoritative Edge devices for the VLANS being extended from the site. This is achieved inside each site using multiple Layer 2 paths for the Site VLAN (site adjacency) and outside the data center using the overlay network (overlay adjacency), offering natively a solid and resilient DCI solution. As the result only one device is responsible to forward broadcast traffic, preventing the creation of end-to-end layer 2 loop.

Rate Limiters

If a Layer 2 issue such as a broadcast storm is propagated across the remote locations, there are risks that the DCI links as well as the remote site are disrupted. In such a situation, none of the resources from the different locations will be available due to the total capacity of the bandwidth immediately overloaded. Consequently, it is required that the remote site is protected by identifying and controlling the rate allocated to the broadcast storm.

In a traditional DCI deployment, where tunnels are initiated at the DCI layer, it is possible to configure the Storm-control to rate-limit the exposure to BUM storms across different DCI blocks.

This required configuration is simplified in OTV because of the native suppression of UU frames and for broadcast containment capability of the protocol (ARP caching). Hence, just the storm of broadcast traffic can be rate-limited for the interface facing the WAN with a few command lines. However this implies treating the original VLAN frame. In the case of a VxLAN fabric interconnected using OTV, this forces all VxLAN overlays to be de-encapsulated at the DCI edge layer.

On the other hand, with VxLAN, it is challenging to rate-limit a VNI Unicast packet transporting BUM traffic with Ingress Replication. With Unicast mode IR, nothing differentiates a Unicast frame from a BUM. In addition, BUM will be also replicated, as many times as there are VTEP affected by this VNI.

In the case the fabric and L3 core is Multicast enabled, then, BUM can be transported within a Multicast group. Therefore, it will be possible to differentiate the flooded traffic from the Unicast workflow and rate-limit that BUM traffic over the Multicast groups at the Core layer. However, it is important to get that not all switches support Multicast rate limiter.

Configuration/complexity

Network Managers are used to configuring OTV in a few command lines without the need to care about the Control Plane.

VxLAN deployment imposes long and complex configurations (BGP, EVPN, VLAN to VNI mapping) with risks of human error. A smart tool to simplify the configuration is a programmatic approach using Python or Perl or any other modern language to ease the deployment of VxLAN. Even though this helps to accelerate the deployment while reducing the risk of mistakes, it requires some time to compose the final scripts to be used in production.

VxLAN and DCI solution: Conclusion

DCI is not just a layer 2 extension between two or multiple sites. DCI/LAN extension is aiming to offer business continuity and elasticity for hybrid cloud in all its forms (Private to Private, Private to Public, Public to Public). It offers disaster recovery and disaster avoidances services. As it concerns on Layer 2 broadcast domain, it is really important to understand the requirement for a solid DCI/LAN extension and the weaknesses that rely on the current implementation of VxLAN.

On the other hand, VxLAN with MP-BGP EVPN AF has some great functionalities that can be leveraged for DCI solution, if we understand its shortcomings. Flood and Learn VxLAN may be used for rudimentary DCI. VxLAN MB-BGP EVPN is taking a big step forward with its Control Plane and can be used for extending the Layer 2 across multiple sites. However it is crucial that we keep in mind some weaknesses related to DCI purposes. Current solution such as OTV provides a mature and proven DCI solution with all embedded DCI features from ground-up. Fortunately, the open standard VxLAN and the necessary intelligence and services are being implemented into the protocol to provide in the future a viable and solid DCI solution.

Addendum

DCI LAN Extension Requirements

- Failure domain must be contained within a single physical DC
  - Leverage protocol Control Plane learning to suppress unknown Unicast flooding.
  - Flooding of ARP requests must be suppressed or reduced and controlled using rate-limiting across the extended LAN.
  - Generally speaking, rate-limiters for the Control Plane and Data Plane must be available to control the broadcast frame rate sent outside the physical DC.
  - The threshold must be set carefully with regard to the existing broadcast, unknown Unicast and Multicast traffic (BUM).

- Dual-Homing
  - Redundant path distributed across multiple edge devices are required between sites, with all paths active without creating any Layer 2 loop.
  - Any form of Multi-EtherChannel split between two distinct devices needs to be activated (e.g., EEM, MC-LAG, ICCP, vPC, vPC+, VSS, nV Clustering, etc.).
  - VLAN-based Load balancing is good; however, Flow-based Load-balancing is the preferred solution.

- Multi-homing
  - Multi-homing going beyond the two tightly coupled DCI edge devices.
  - Any backdoor Layer 2 extension must be detected.

- Layer 2 Loop detection and protection
  - Built-in loop prevention is a must-have.

- Independent Control Plane on each physical site
  - A unique active Control Plane managing the whole VxLAN domain may lose access to the other switches located on the remote site.
  - Reduced STP domain confined inside each physical DC.
  - STP topology change notifications should not impact the state of any remote link.
  - Primary and secondary root bridges or Multicast destination tree (e.g., FabricPath) must be contained within the same physical DC.

- Layer 2 protocol independence inside the network fabric
  - The DCI LAN extension solution should not have any impact on the choice of Layer 2 protocol deployed on each site.
  - Any Layer 2 protocols must be supported inside the DC (STP, MST, RSTP, FabricPath, TRILL, VxLAN, NVGRE, etc.) regardless of the protocols used for the DCI and within the remote sites.
  - Hybrid Fabric should be supported on both sides.

- Remove or reduce the hair-pinning workflow for long distances

- Layer 3 default gateway isolation such as FHRP isolation, Anycast gateway or proxy gateway.
- Intelligent ingress path redirection such as LISP, IGP Assist, Route Health Injection, or GSLB.

- Fast convergence
  - Sub-second convergence for any common failure (link, interface, line card, supervisor, core).

- Multi-sites
  - The DCI solution should allow connecting two or more sites.

- Any Transport
  - The DCI protocol must be transport-agnostic, meaning that the DC interconnection can be initiated on any type of link (dark fiber, xWDM, Sonet/SDH, Layer 3, MPLS, etc.).
  - Usage of IP Multicast may be an option for multi-sites but should not be mandatory. The DCI solution must be able to support a non-IP Multicast Layer 3 backbone because an enterprise would not necessarily obtain the IP Multicast service from their provider (especially for a small number of sites to be Layer 2 interconnected).
  - Enterprises and service providers are not required to change their existing backbone.

- Path Diversity
  - Path diversity and traffic engineering to offer multiple routes or paths based on selected criteria (e.g., Control Plane versus Data Plane).
  - This is the ability to select specific L2 segments to use an alternate path (e.g., Data VLAN via path#1 and Control VLAN via Path#2).

A clarification on the taxonomy:

Entropy/Depolarization: Variability in the header fields to ensure that different flows being tunneled between the same pair of TEPs, do not all follow the same path as they traverse a multi-pathed (usually ECMP) underlay network.

Load Balancing: The ability of an overlay to send different flows to different TEPs/EDs when a destination site is multi-homed (ideally also dual-homed, although dual-homing the way we've implemented it wouldn't lend itself to this, we are reducing the destination to a single any cast address, trumping our ability to prescriptively load balance)

Path Diversity: The ability of an encapsulating TEP to steer different encapsulated flows over different uplinks. Different uplinks should either belong to different underlay networks or be mapped to underlay Traffic Engineering tunnels (MPLS-TE, Multi-topology routing, Segment Routing or other) that guarantee the end-to-end divergence of the paths taken by traffic that enters the TE tunnels.