ARISTA DESIGN GUIDE

Data Center Interconnection with VXLAN

Version 1.0
November 2014

The requirement to operate multiple, geographically dispersed data centers is a fact of life for many businesses and organizations today. There are many reasons for distributing applications and data in more than one data center, such as increasing levels of service availability, improving application performance and driving operational efficiencies. Data center interconnection ensures that data is consistent, allows for the rapid movement of virtualized workloads, and facilitates the deployment of application high availability solutions between locations.

This design guide describes an Arista solution for providing simple, robust, and cost-effective data center interconnection (DCI) that enables layer-2 services to be bridged between multiple sites over existing layer-3 IP networks. This solution is completely standards-based and can be deployed over existing routed TCP/IP transport networks. Based on multi-chassis link aggregation (MLAG) and Virtual eXtensible Local Area Network (VXLAN), it provides a highly available DCI framework for linking two or more data centers. It requires no special hardware and can be deployed using any Arista switching hardware that supports VXLAN+MLAG.

TABLE OF CONTENTS

THE DRIVERS FOR DATA CENTER INTERCONNECTION

Operating multiple, geographically dispersed data centers has been regarded as best practice for many years, and all types of businesses and organizations use this deployment model. Today, it may even be a statutory requirement for some businesses (e.g., banking and finance), or a part of an organization's governance policy.

The original reason for building and operating multiple data centers was to ensure business continuity. The rationale was simple: it was highly unlikely that an issue affecting power or service provision in one part of the world would also impact another location (provided they are adequately far apart). A beneficial side-effect of dispersing applications and data was that they could be positioned close to the ultimate consumers of the service, which in turn meant users experienced better response times, resulting in improved productivity (or in the case of customers, greater satisfaction with the service). The consequence of locating services close to users was that there was less demand for bandwidth on expensive wide-area connections, thus reducing operational costs.

With the dispersion of applications and data came the need to ensure consistency of information between locations. This has driven the need to interconnect data center locations for the purposes of data replication and to enable shared information processing between servers forming geographically dispersed clusters.

More recently, additional drivers for data center interconnection (DCI) have emerged. The widespread use of server virtualization has delivered the potential to provide live migration of compute workloads (e.g., applications, virtual desktops, etc.) between sites. The original reason to provide the live migration of workloads was for improved service availability. However, this ability to move workloads between locations has given rise to other use cases; for example, applications can be moved to locations to be close to service users (e.g., "follow-the-sun") or to where power is cheaper (e.g., "follow-the-moon").

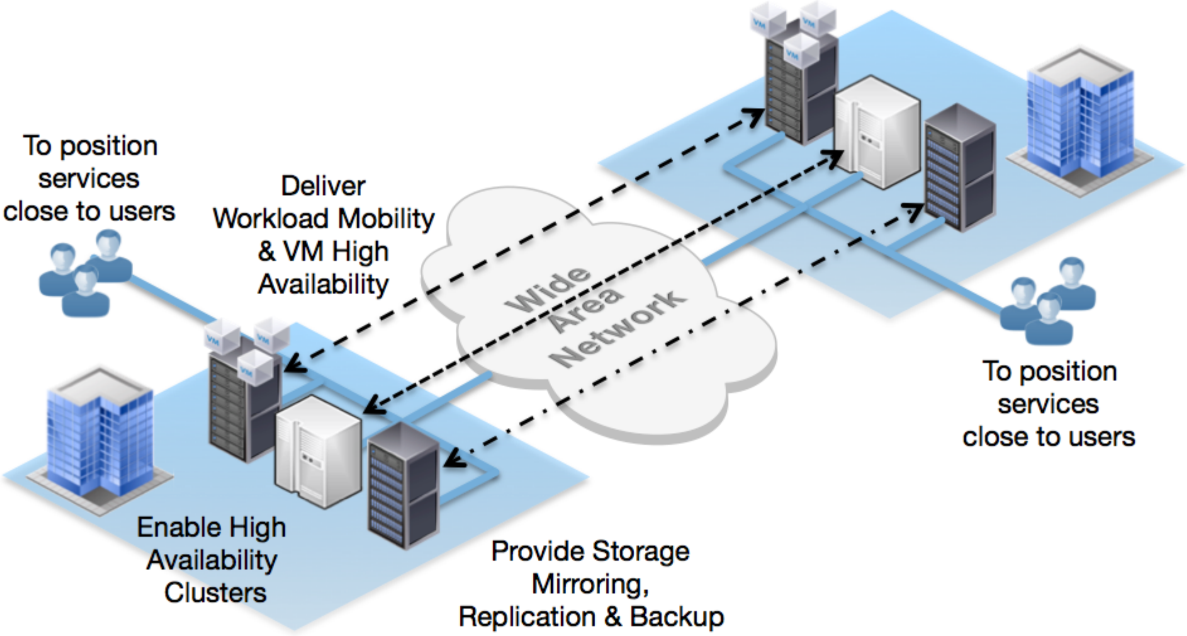
To position services close to users
Deliver Workload Mobility & VM High Availability
VM
Wide Area Network
To position services close to users
Enable High Availability Clusters
Provide Storage Mirroring, Replication & Backup

Figure 1: Some of the Typical Use Cases for Data Center Interconnection

With the advent of cloud computing models for IT service delivery, the dynamic provisioning of workloads, and the capability to instantiate applications on demand (often without consideration of location), the possibility that the constituent elements of an application could be split across different data centers becomes a reality. Likewise, cloud solutions may choose to “re-balance” a workload based on specific operational criteria. For example, virtual machines (VM) may be automatically moved from a server requiring downtime for maintenance, or workloads may be redistributed across physical hosts in order to optimize power distribution or cooling. It is quite possible that VMs could be placed on physical servers in different data centers.

Data center interconnection is often required as a temporary measure to facilitate the permanent migration of applications and data from one location to another. For example, many organizations are seeking to consolidate and rationalize their data centers in order to improve efficiency and reduce operating costs. In this case, a temporary DCI solution can be deployed to transfer data and applications from a site due for closure to their new home. Migration can also be a requirement when businesses merge or are acquired, where the need to rationalize resources and locations is deemed necessary. The reverse scenario is also true, as in the case of a business divesting part of its operation, where IT services need to relocate to the new organization's facilities.

Ultimately there are many circumstances driving need for data center interconnection and, in reality, businesses may have a combination of the requirements described above. It is likely that new requirements will emerge in the future, especially as adoption of the cloud computing paradigm continues. In the next section we will consider some of the technical requirements for delivering a DCI solution.

DATA CENTER INTERCONNECTION TECHNOLOGIES

Most network designers would prefer that data center interconnection over wide area networks (WAN) be performed at layer 3, without layer-2 connectivity spanning data centers. However, application requirements often make it necessary to provide layer-2 interconnection between data centers, with common IP subnets and VLANs stretched across sites. While it is true that some VM mobility solutions have removed the need to provide layer-2 connectivity between locations, many current solutions still require layer-2 interconnection. This is also true of many high availability clustering systems and storage virtualization and replication solutions, which mandate layer-2 adjacency between physical nodes.

However, native, long-distance, layer-2 connections such as local area network (LAN) extension services can be extremely difficult to find, especially outside of metropolitan areas or at distances that ensure data centers are adequately geographically dispersed to provide full protection from outages. Even where native layer-2 services are available, it may be undesirable to use them due potential “fate-sharing” between sites, as in the case of a layer-2 problem such as a broadcast storm on one site impacting the other data center.

PSEUDOWIRE & TUNELLING SOLUTIONS

The challenges associated with providing cost-effective, long-distance layer-2 interconnection led to the development of a set of solutions based on tunneling mechanisms. These were designed to allow layer-2 traffic to be transported over Layer-3 networks. Originally these solutions centered on the concept of a “pseudowire,” which allowed for the emulation of point-to-point Ethernet links by tunneling all traffic over wide area networks. Tunneling mechanisms such as Generic Router Encapsulation (GRE), Layer-2 Tunneling Protocol (L2TP), or Multi Protocol Label Switching (MPLS), provided the underlying framework for this approach. Unfortunately, the initial solutions that emerged in this space, such as EoGRE (Ethernet over GRE), EoMPLS (Ethernet over MPLS), AToM (Any Transport over MPLS) did not scale well. For example, to allow any-to-any communication across multiple sites, full mesh connectivity of point-to-point tunnels links was required. As a result, they proved difficult to deploy and very complex to troubleshoot. Also, due the unfiltered transmission of all traffic, these solutions did not deliver the necessary level of fault isolation between locations.

In order to overcome some of the limitations of the point-to-point tunneling approach described above, technologies such as Virtual Private LAN Service (VPLS) were developed to allow the multipoint emulation of LAN services over underlying pseudowires. While this solution addressed some of the challenges associated with earlier approaches, in that it allowed multipoint connectivity and offered improved fault isolation between sites, it added the complexity associated with the underlying MPLS infrastructure necessary to support it.

In recent years, much of the industry effort to find a simple, robust, multipoint mechanism for carrying layer-2 traffic over layer-3 networks has polarized into distinct camps, with very little cross-vendor support. These solutions, which include OTV (Overlay Transport Virtualization) from Cisco, and Juniper's EVPN (Ethernet Virtual Private Network), have aimed to address customer requirements for DCI solutions, but have lacked any widespread support from other vendors, essentially locking customers into a single supplier for their inter-data-center connectivity. Even within the individual vendors, support for their respective solutions can be very limited, often restricted to high-end and costly devices; for example, Cisco's OTV is only available on the Nexus 7000 platform with M-series modules and the ASR 1000 router.

VXLAN: A SCALABLE, STANDARDS-BASED DCI SOLUTION

Given a blank sheet of paper it would not be difficult to define the desirable characteristics of layer-2-capable data center interconnect solutions. Ideally, any modern DCI solutions should meet the following basic criteria:

- Transparent to applications
- Transport agnostic – can operate over any IP-based service
- Multiple path and multiple site support
- Capable of providing fault isolation between data centers
- Interoperable and standards based
- Simple to implement
- Platform independent – no need for specialized hardware or equipment
- Managed and controlled as part of the wider DC infrastructure – not an "alien" technology

The foundation of the Arista solution is **Virtual eXtensible Local Area Network** (VXLAN), an open IETF specification designed to standardize an overlay encapsulation protocol, capable of relaying layer-2 traffic over IP networks. VXLAN has wide industry support and was authored by Arista, Cisco and VMware with support from Broadcom, Citrix and Red Hat among others.

Arista's solution for data center interconnect meets these requirements and represents the industry's first truly open standards-based, simple to deploy and manage DCI system. It is cost-effective, running on standard switching hardware, and it provides active-active switch redundancy and can interoperate with wide range of other data center switches, including those from other vendors.

INTRODUCING THE ARISTA NETWORKS DCI WITH VXLAN SOLUTION

VXLAN was designed for the creation of logical layer-2 domains on top of an underlying IP network, initially to enable network virtualization in the data center. VXLAN identifies individual layer-2 domains using a 24-bit **virtual network identifier** (VNI), allowing for up to 16 million independent domains to be specified. layer-2 Ethernet frames are encapsulated in IP UDP datagrams and are relayed transparently over the IP network. It is the inherent ability to relay unmodified layer-2 traffic transparently over any IP network that makes VXLAN an ideal technology for data center interconnection.

Within the VXLAN architecture, **virtual tunnel end points** (VTEP) perform the encapsulation and de-encapsulation of layer-2 traffic. Each VTEP is identified by an IP address, which is assigned to a **virtual tunnel interface (VTI)**. The VTEP receives standard layer-2 Ethernet frames, selects the correct VNI and forms an IP UDP packet for transmission to one or more destination VTEPs. The source IP address is that of the sending VTI; the destination IP address is that of the receiving VTI.

The VNI is typically determined based on the IEEE 802.1Q VLAN tag of the frame received. The destination VTEP (or VTEPs in the case of multicast or broadcast traffic) is selected using a destination-to-VTEP map. This map is very similar to the MAC bridging table, except MAC addresses are associated with IP addresses rather than switch interfaces.

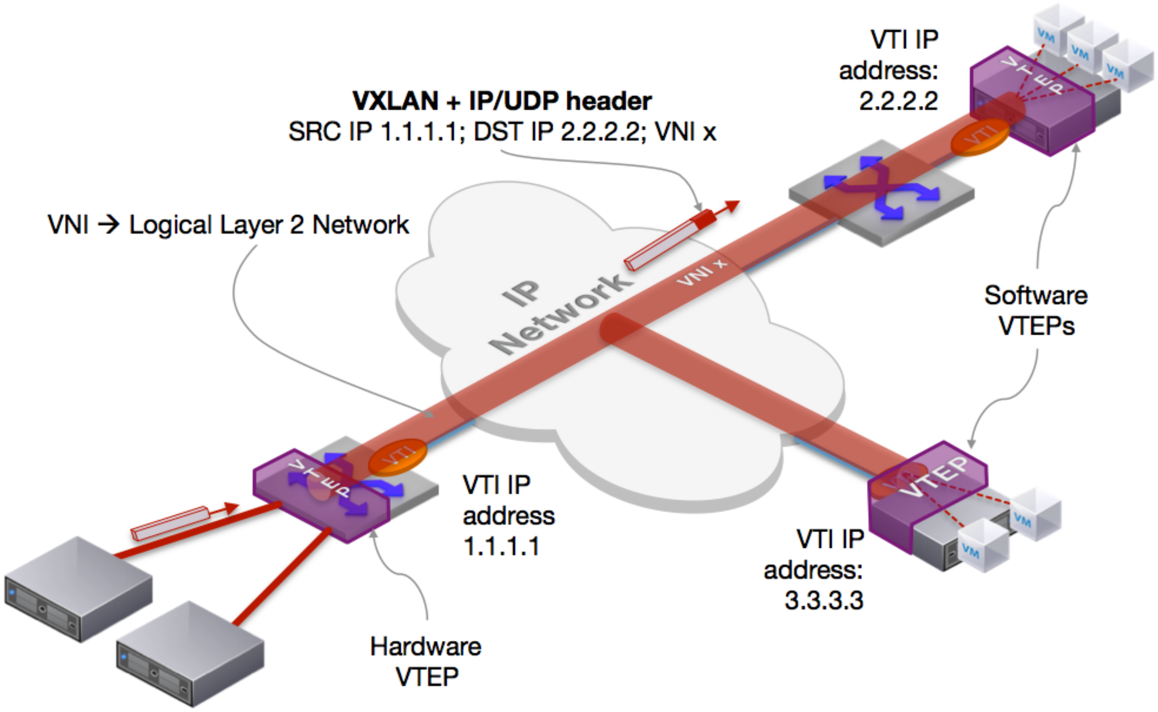

Figure 2: VXLAN Components and Operation

MAC LEARNING AND THE FORWARDING OF BROADCAST & MULTICAST TRAFFIC

In order to transparently support layer-2 traffic, each VTEP must handle the forwarding of broadcast and multicast traffic as well as ensure previously “unseen” MAC addresses and their “locations” are learned. In a layer-2 network, MAC address learning is performed by flooding frames with unknown unicast addresses on all ports within a given VLAN. For VXLAN, the approach specified in the IETF RFC is based on IP multicast. In this scenario, one or more IP multicast groups are set up to carry unknown unicast broadcast and multicast traffic to VTEPs associated with a given VNI.

This approach means that the underlying IP network used to transport VXLAN encapsulated traffic must support IP multicast. However, many wide area networks used to interconnect data centers do not support or implement IP multicast. To alleviate this issue, Arista has introduced a feature referred to as “Head End Replication” (HER), which takes incoming broadcast, multicast, and unknown unicast traffic and sends a single unicast copy to each of the VTEPs receiving traffic for a given VNI.

ARISTA DCI SOLUTION COMPONENTS

The foundation of the Arista solution for data center interconnection is the VXLAN Hardware Gateway, deployed in conjunction with multi-chassis link aggregation (MLAG).

In addition, Arista’s Virtual ARP (vARP) mechanism can be used to ensure redundant first-hop router interfaces are localized within the data center, in order to prevent traffic destined to exit the data center from consuming bandwidth on the links interconnecting data centers.

ARISTA NETWORKS VXLAN HARDWARE VTEP GATEWAY

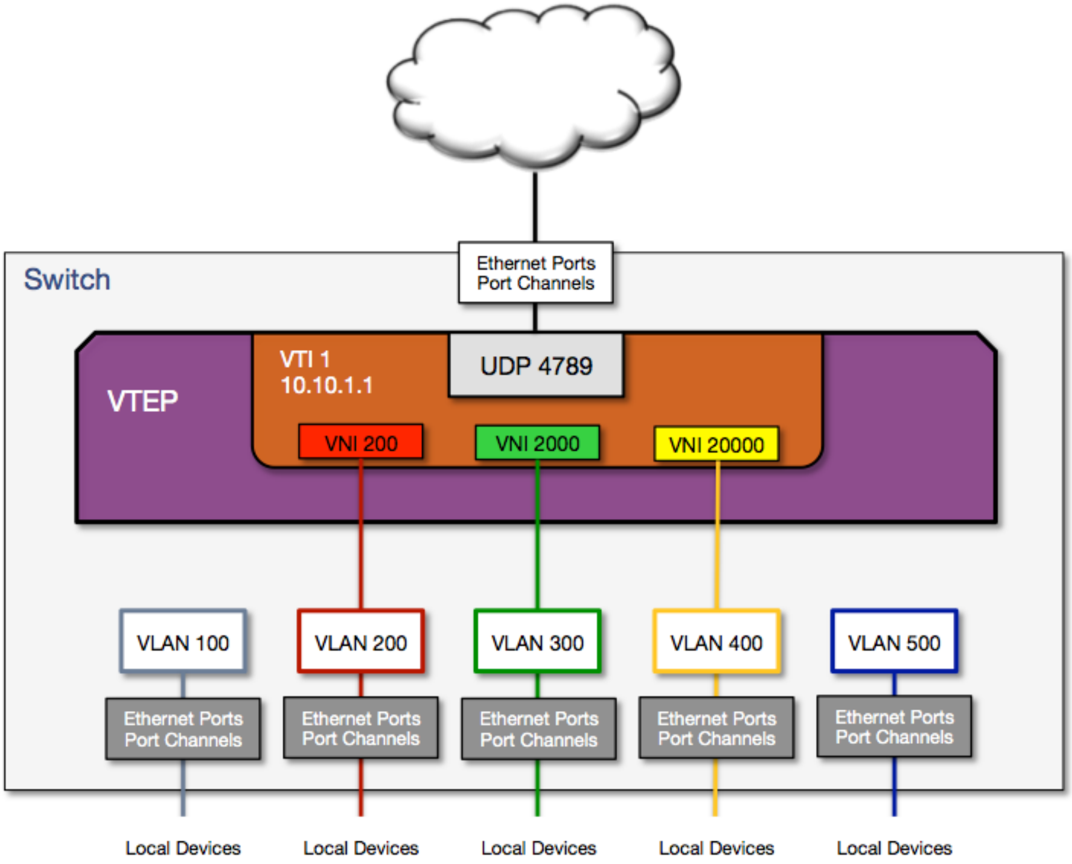

Figure 3: Arista Hardware VXLAN Gateway Architecture

Configuring the Arista VXLAN Hardware Gateway is extremely simple. An IP address for each VTEP is specified using a loopback interface, which is in turn mapped to the VTI (referred to as the “vxlan 1” interface). The default IP port for VXLAN traffic is UDP 4789, although this can be changed if required

Each VTEP performs local-VLAN-to-VXLAN-VNI mapping, enabling VLAN translation to be performed as part of the tunneling process if required (i.e., source and destination VLAN IDs can be different). The Arista VXLAN Hardware Gateway is capable of emulating layer-2 local area networks (LANs), so both point-to-point and point-to-multi-point logical layer-2 topologies are supported with the ability to deliver broadcast, multicast and unknown unicast traffic. Each VTEP filters spanning tree BPDUs ensuring that each DC is an independent spanning tree domain, helping isolate potential faults.

The initial version of Arista’s VXLAN gateway handled the flooding of broadcast, multicast and unknown unicast traffic using a single, configurable multicast group. Later versions of Arista EOS® introduced the option to use Head End Replication (HER), which allows traffic with multiple destinations (i.e., broadcast or multicast frames) or traffic requiring flooding (i.e., unknown unicast frames) to be relayed in individual IP unicast packets, thus eliminating the need to implement multicast on the layer-3 transport network.

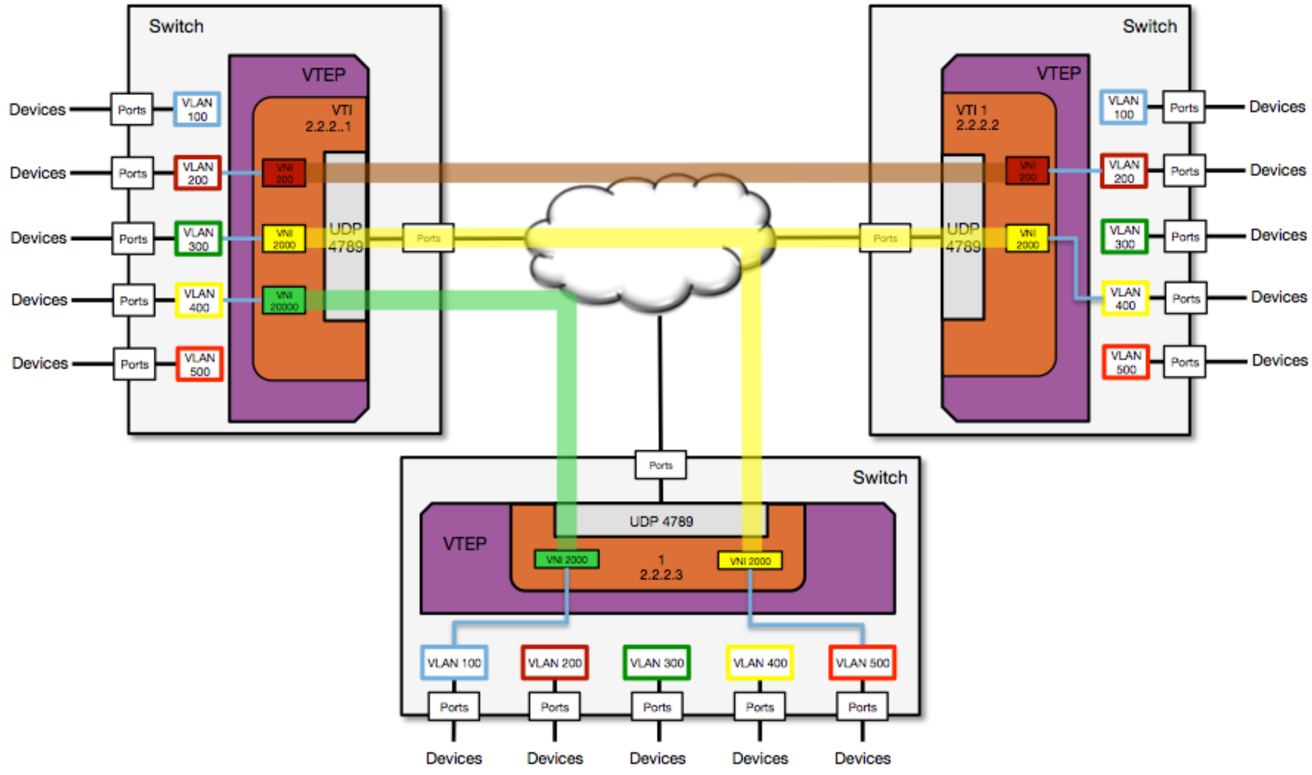

Figure 4: Arista VXLAN Point-to-Point & Multipoint VNIs with VLAN Translation

LICENSING & PLATFORM SUPPORT

VXLAN requires the Arista V- Virtualization Features License. Full details of VXLAN feature availability and platform support can be found at: http://www.arista.com/en/support/supported-features

ARISTA MULTI-CHASSIS LINK AGGREGATION (MLAG)

Arista's multi-chassis link aggregation (MLAG) enables ports on two separate switches to be combined into a single Ethernet port channel. For example, two 10-gigabit Ethernet ports, one each from two MLAG-configured switches, can connect to two 10-gigabit ports on a host, switch, or network device to create a link that appears as a single 20-gigabit port channel. MLAG-configured ports effectively provide layer-2 multi-path forwarding, thus increasing bandwidth, providing higher availability and improving efficiency by avoiding the need to have either active-standby connections or to rely on spanning tree to block alternate paths.

With MLAG, two aggregation switches create a single logical layer-2 switching instance that utilizes all connections to the switches. Interfaces on both devices participate in a distributed port channel, enabling all active paths to carry data traffic while maintaining the integrity of the Spanning Tree topology.

Arista MLAG provides these benefits:

- Provides higher bandwidth links as network traffic increases.
- Utilizes bandwidth more efficiently with fewer uplinks blocked by STP.
- Connects to other switches and servers by static LAG or IEEE 802.3AX Link Aggregation Control Protocol (LACP) without the need for proprietary protocols.
- Aggregates up to 32 10-Gbps Ethernet ports across two switches: 16 ports from each switch.
- Supports normal Spanning Tree Protocol (STP) operation to prevent loops.
- Supports active-active layer-2 redundancy.

An MLAG consists of two or more links that terminate on two cooperating switches and appear as an ordinary link aggregation group (LAG) to the connecting device (e.g., switch, host, storage system, load-balancer, firewall etc.). The two switches that form the MLAG relationship are referred to as MLAG peer switches, which communicate through an interface called a MLAG Peer Link. While the Peer Link's primary purpose is exchanging MLAG control information between peer switches, it also carries data traffic from devices that are attached to only one MLAG peer and have no alternative path. An MLAG domain consists of the peer switches and the control links that connect the switches.

A dedicated MLAG peer VLAN is configured on each of the peers to maintain the peer link and relay control information using TCP. An IP address is assigned to the peer VLAN interface on each switch.

The MLAG domain ID is a text string configured in each peer switch. MLAG switches use this string to identify their respective peers. The MLAG system ID (MSI) is the MLAG domain's MAC address. The MSI is automatically derived when the MLAG forms and does not match the bridge MAC address of either peer. Each peer uses the MSI in STP and LACP PDUs.

The topology in Figure 5 contains two MLAGs: one MLAG connects each device to the MLAG domain. Each peer switch connects to the two servers through MLAG link interfaces. In this example, the MLAG for Host A contains two links, while the MLAG for Host B has 4 links. Switch A and Switch B are peer switches in the MLAG domain "MLAGDomain01" and connect to each other through the peer link.

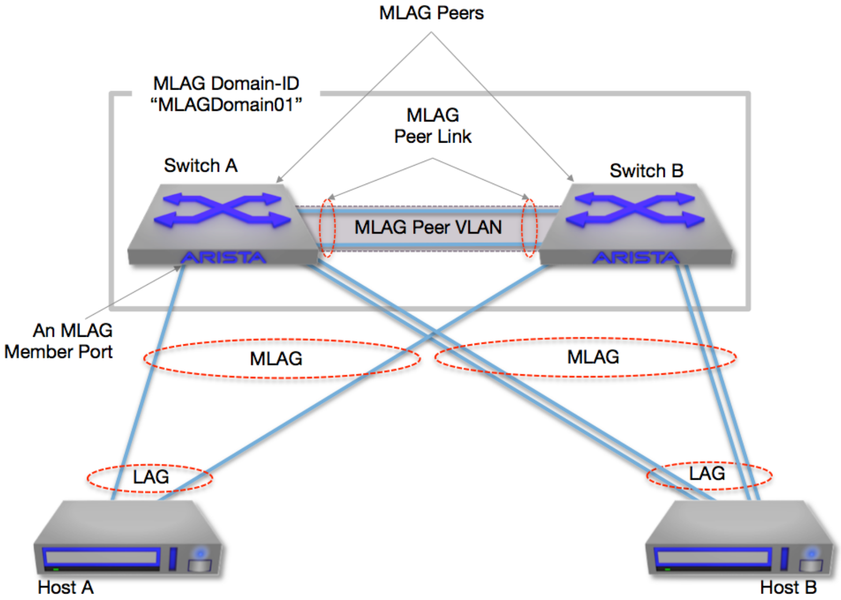

Figure 5: Arista Multi-Chassis Link Aggregation Architecture

In a conventional topology, where dual-attached devices connect to multiple layer-2 switches for redundancy, Spanning Tree Protocol (STP) blocks half of the switch-device links. In the MLAG topology, STP does not block any portion because it views the MLAG Domain as a single switch and each MLAG as a single link. The MLAG protocol facilitates the balancing of device traffic between the peer switches.

When MLAG is disabled, peer switches revert to their independent state. MLAG is automatically disabled by any of the following conditions:

- The MLAG configuration is changed
- The TCP connection between MLAG peers fails
- The peer link or local interface goes down
- A switch does not receive a response to a keep-alive message from its peer within a specified period

As Arista MLAG is standards-based, two MLAG domains can be connected together to form a "bow-tie" topology as shown in Figure 6. It is also possible to connect other vendors' implementation of multi-chassis link aggregation to an Arista MLAG pair in the same way, provided the links conform to either the IEEE 802.1AX LACP specification or can be configured as a static LAG.

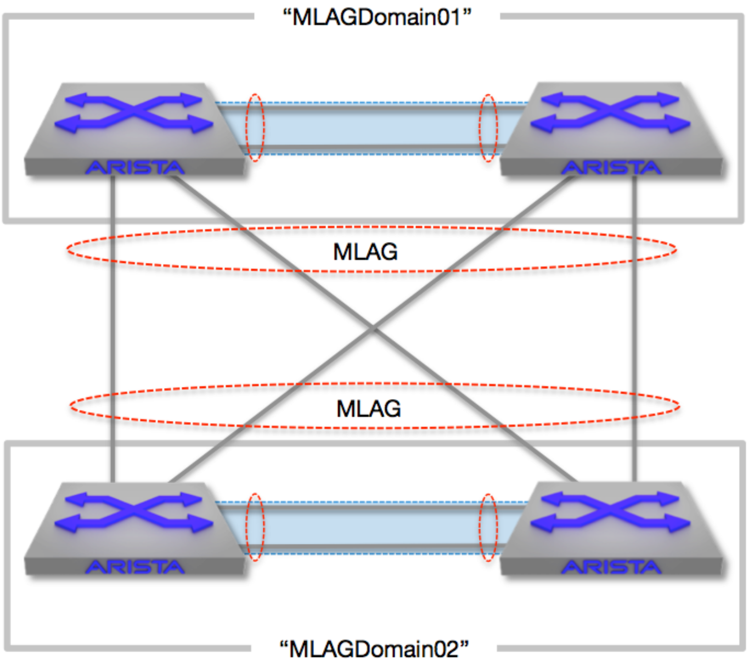

Figure 6: "Bow-Tie" MLAG Topology

ARISTA VXLAN+MLAG SOLUTION

Arista's implementation of VXLAN can span a pair of switches interconnected with MLAG. This allows for the implementation of a VTEP that operates on two separate switches simultaneously, effectively creating a "logical" VTEP. This doubles performance as well as providing an active-active, fully redundant highly available system.

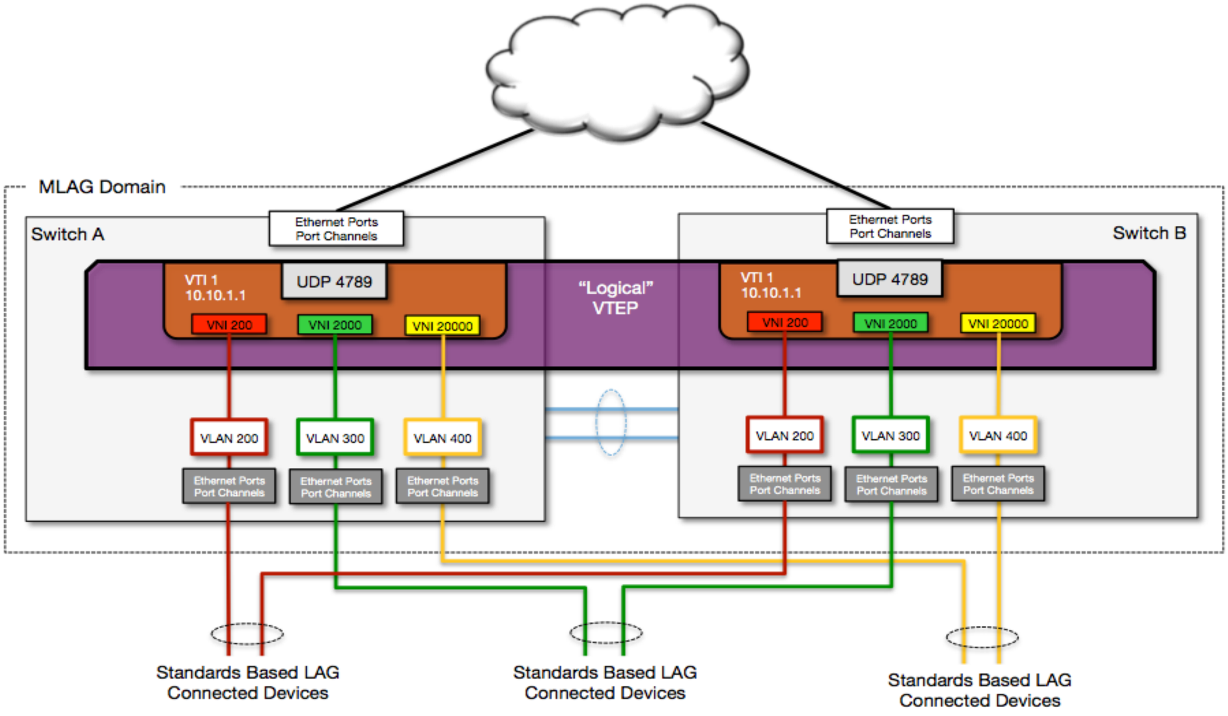

Figure 7: Arista VXLAN+MLAG Hardware VTEP Architecture

DEPLOYING THE ARISTA VXLAN DCI SOLUTION

There are a wide variety of topologies that can be supported with the Arista VXLAN Data Center Interconnect solution, including dual-site or 3 or more site configurations. At each site, the DCI connection can be implemented as a highly available VXLAN+MLAG configuration or as a single VXLAN-enabled switch. With EOS 4.14, Arista supports VXLAN+MLAG bridging on the 7150 and 7280SE switches. Standalone VXLAN bridging is also supported on the 7050X family of switches. The wide variety of Arista platforms supporting VXLAN bridging in hardware allows users to select the best product for the any given deployment scenario. For example, smaller locations can utilize the cost effective 7150S-24 24-port 10GbE switch, while larger sites with more demanding traffic workloads can deploy the 7280SE-64 with 48 x 10GbE and 4 x 40GbE ports, which due to its 9GB of deep buffering would be a ideal choice. The 7280SE is especially suited for scenarios where the oversubscription ratio from LAN to WAN may be very high or where traffic profiles have significant but transient bursts.

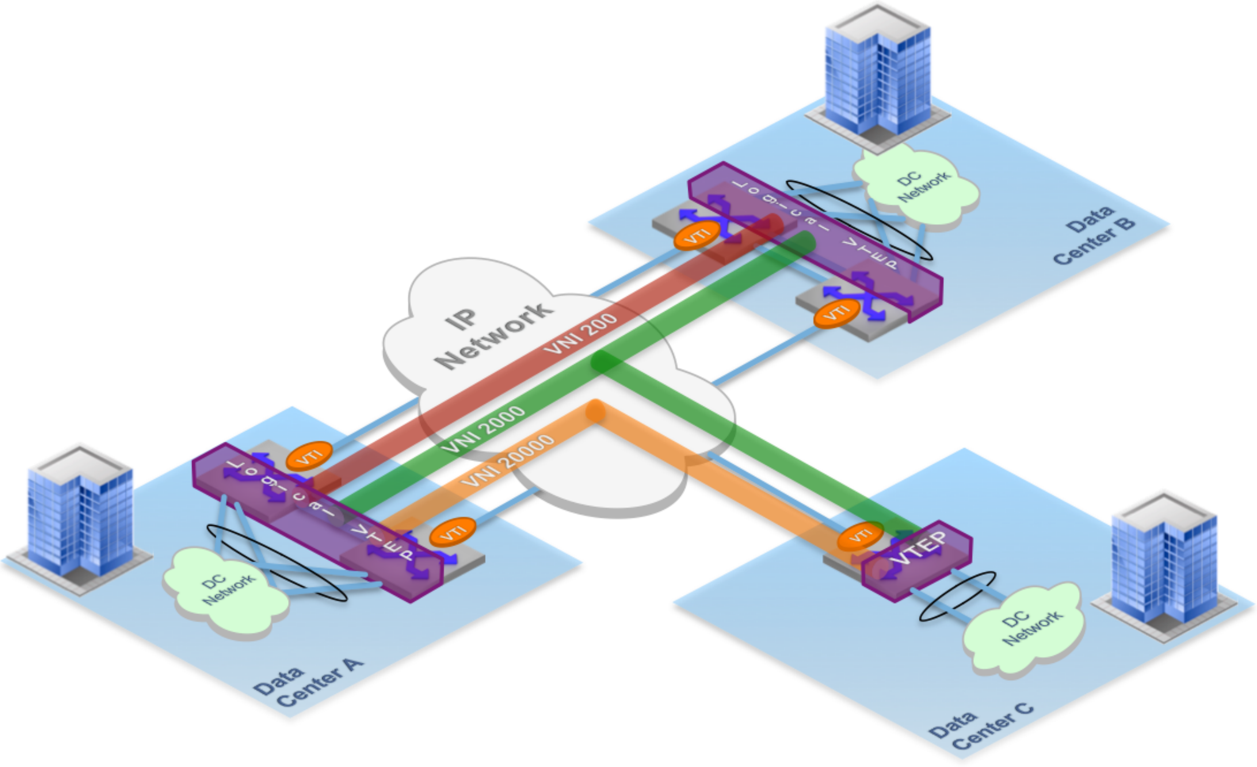

Figure 8: The Industry's First Truly Flexible, Scalable, High-Performance and Cost-Effective DCI Solution

INTEGRATING THE ARISTA DCI SOLUTION WITH EXISTING NETWORKS

As VXLAN is standards-based, it is relatively simple to connect existing network equipment to the Arista VXLAN DCI solution. Any device capable of link aggregation or port channels —either statically configured or using 802.1AX Link Aggregation Control Protocol (LACP) —can connect to the Arista switches as an MLAG pair. This includes, of course, other Arista devices, or other vendors' solutions such as Cisco VPC or VSS, or equivalent technologies from other vendors.

The most common deployment scenario for the Arista VXLAN DCI solution a dual-site DC environment with active-active VXLAN+MLAG DCI configurations at each site. Typically the pair of VXLAN+MLAG switches to be used for data center interconnection are deployed as a dedicated "DCI module", which is connected to the existing leaf-spine data center network (DCN) via a “network edge leaf” using a “bow-tie” MLAG configuration (as shown in Figure 9). DC network edge leaf can consist of any switch capable of supporting standards-based single- or multi-chassis link aggregation, either using static port channels or IEEE 802.1AX-2008 LACP based link aggregation. The inter-site connection can be a direct, point-to-point layer-3 link, or a routed transport network. Any dynamic routing protocol supported on the Arista switches can be used (e.g. OSPF, BGP, IS-IS etc.) or static routing can be used if preferred.

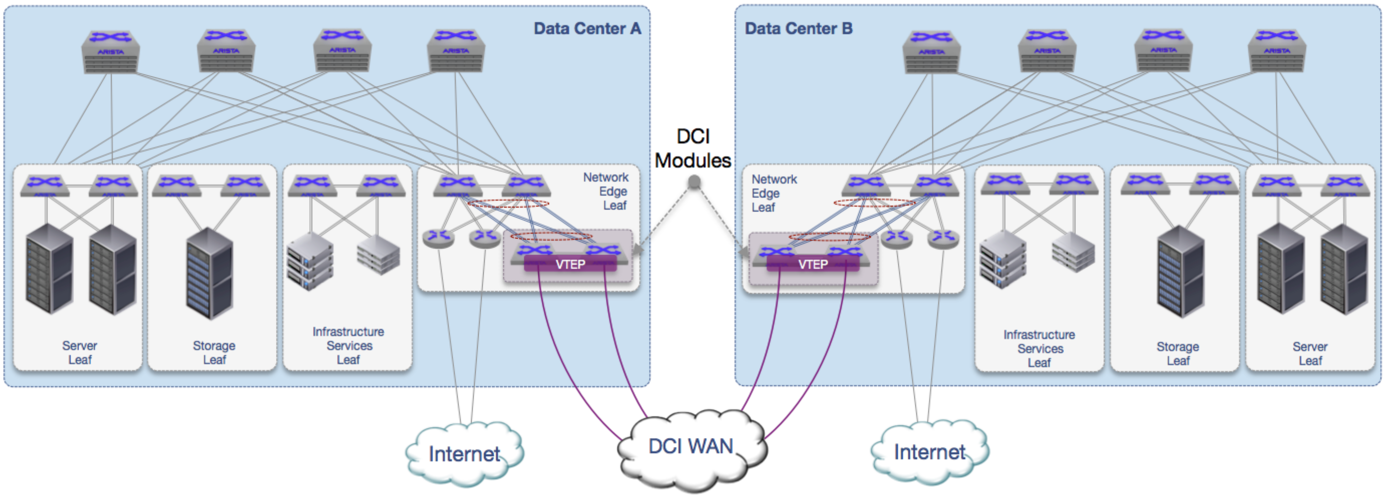

Figure 9: Integration of the DCI module with a Leaf-Spine Data Center Network

Arista’s leaf-spine architecture can be built with either a layer-2, MLAG-based spine or with a layer-3 routed spine. With a layer-2 leaf-spine topology, VLANs can be extended from any leaf node to the Arista DCI module. In the case of a layer-3 leaf-spine topology, layer-2 network virtualization with VXLAN and Arista’s hardware VTEP allows VLANs to span different leaf blocks, allowing VLANs to be extended to the DCI module as required.

Of course, not all data center networks are built using leaf-spine architectures and there are still many DC networks in operation today based on hierarchical 3-tier designs. These 3-tier designs are often built with layer-2 access and distribution tiers, and with layer-3 interfaces in the core tier. The core provides inter-VLAN routing as well as access to external WAN routers and infrastructure services such as firewalls, load balancers, etc. In this scenario, the logical place to connect the Arista DCI module is to the core switches, with the inter-DC interfaces on the Arista switches either routed via existing external routers or dedicated DCI routers, or routed directly to the corresponding Arista DCI switches at the remote site. In this scenario, servers, storage etc. have layer-2 connectivity to the Arista DCI module.

The VLANs that are identified as needing to be relayed between data centers are trunked via the MLAG to the DCI module switches. Within the DCI modules, these VLANs are mapped, one-to-one, to VXLAN VNIs, and the remote VTEPs are configured for each VNI to be carried for the purposes of Head End Replication. A logical VXLAN VTEP that spans both MLAG member switches is configured on each of the DCI modules. Logical VTEPs are configured in the exact same way as those on standalone switches, with each member of the logical VTEP being configured with the same VTI address.

Depending on the specific deployment scenario, careful consideration will need to be given to the intervening transport network, especially with respect to the maximum transfer unit (MTU) size. As VXLAN is a MAC-in-IP encapsulation technology, the encapsulation process will add 50 additional bytes to each frame being relayed across a VXLAN connection. Networking devices carrying VXLAN-encapsulated traffic, both within the DC and across the wide area network transport, need to be capable of supporting the resulting larger packets. If, as in some cases, some older router interfaces are incapable of supporting MTUs of greater than 1500 bytes, it may be necessary to modify the default MTU size on some end devices to ensure this limit isn't exceeded.

CONFIGURATION EXAMPLE

The following example describes a simple VXLAN data center interconnection deployment scenario, with the following characteristics:

- Dual data centers
- A DCI module per site, consisting of a pair of switches configured as an MLAG pair connected to a single LAG attached to a "downstream" switch (representing the local DC network)
- Parallel, point-to-point layer-3 links between adjacent MLAG members at each DC
- A logical VXLAN VTEP configured on each DCI MLAG pair
- OSPF routing between sites
- Two VLANs (100 & 200) to relay between sites, each mapped to a corresponding VXLAN VNI as follows:
    - VLAN100 mapped to VNI 10000
    - VLAN200 mapped to VNI 200

The following diagrams show, in turn, the physical and logical topologies of this network. The relevant configuration file sections are also included for reference.

THE PHYSICAL TOPOLOGY

Figure 10 shows the physical topology used for this configuration example. All links are 10GbE in this example.

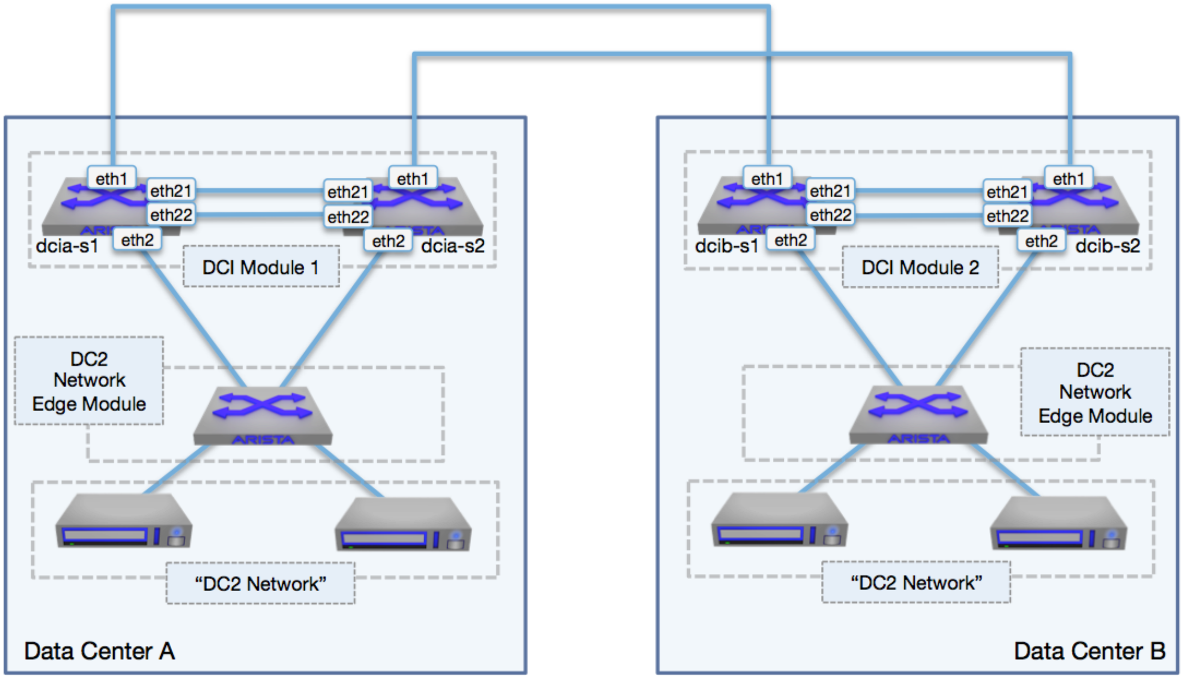

Figure 10: Example DCI Solution – Physical Topology

Switch Configurations – Interfaces

The following configuration examples show the non-default interface configuration required to build the above topology, i.e.,

- Inter-site links configured as layer-3 routed interfaces
- MTU size adjusted on these interfaces, as the default for layer-3 interfaces is 1500 bytes (i.e., increase the MTU to the maximum of 9124 bytes)

All other interfaces use default configuration.

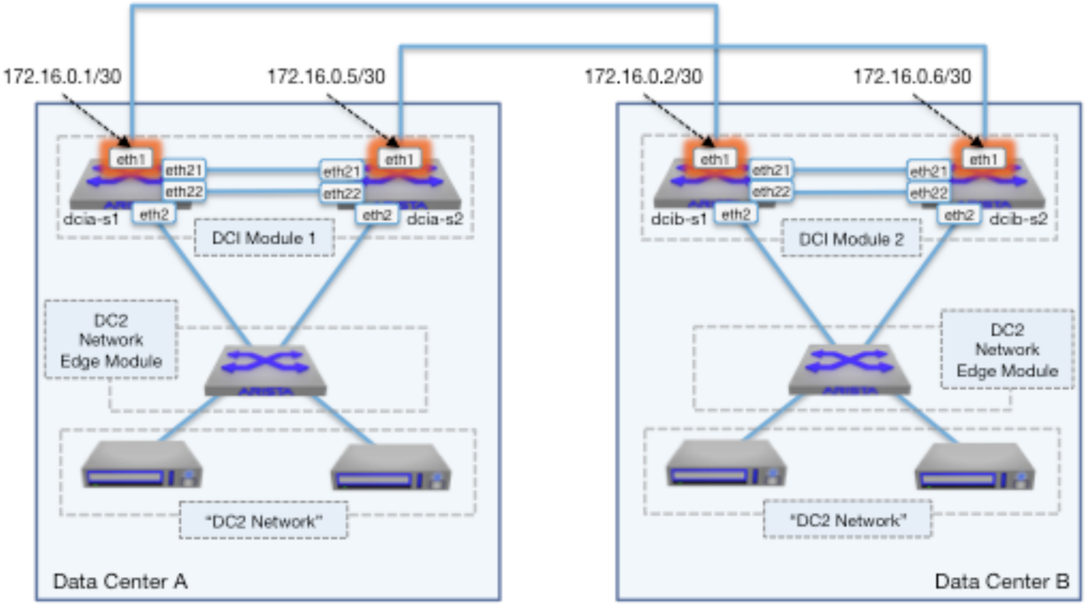

Figure 11: Example DCI Solution – Layer-3 Interfaces

The following configuration file sections show the relevant interface configuration for each of the DCI switches.

Switch "dcia-s1"

```
!
interface Ethernet1
   description P2P L3 link to dcib-s1
   mtu 9214
   no switchport
   ip address 172.16.0.1/30
!
```

Switch "dcia-s2"

```
!
interface Ethernet1
   description P2P L3 link to dcib-s2
   no switchport
   mtu 9214
   ip address 172.16.0.5/30
!
```

Switch “dcib-s1”

```
!
interface Ethernet1
   description P2P L3 link to dcia-s1
   no switchport
   mtu 9214
   ip address 172.16.0.2/30
!
```

Switch “dcib-s2”

```
!
interface Ethernet1
   description P2P L3 link to dcia-s2
   no switchport
   mtu 9214
   ip address 172.16.0.6/30
!
```

THE MLAG TOPOLOGY

Figure 12 shows the MLAG configuration and topology details used for this example configuration.

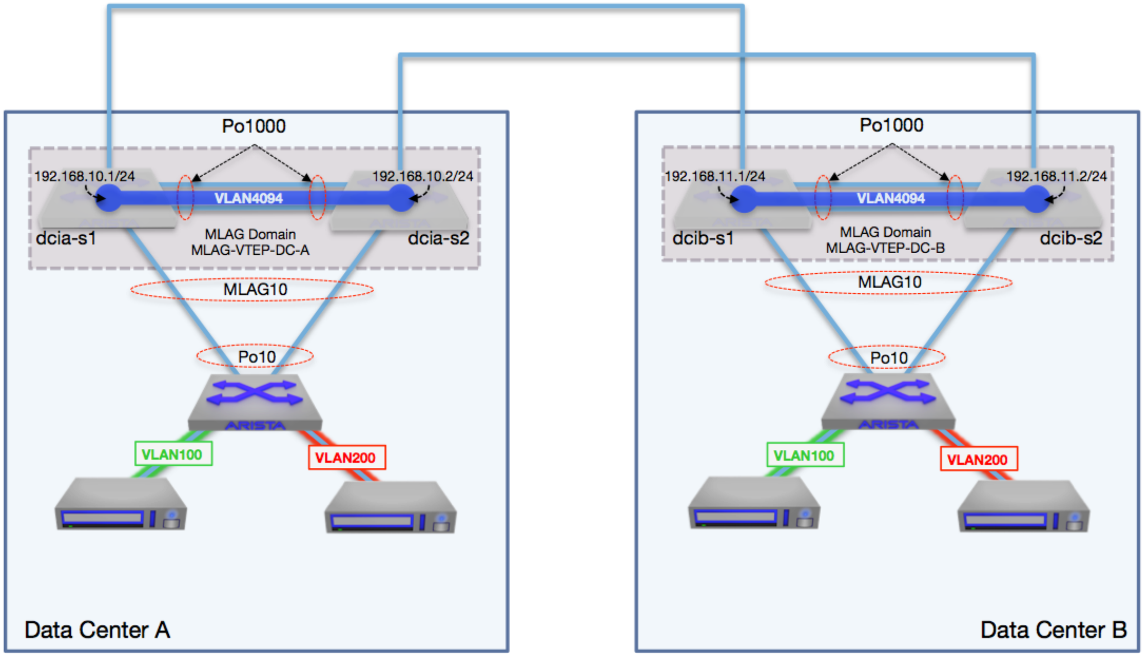

Figure 12: Example DCI Solution – MLAG Topology

Switch Configurations - MLAG

The following configuration examples show the non-default configuration required to build the above topology, i.e.,

- MLAG peer VLAN (i.e., VLAN 4094) created and assigned to the MLAG trunk group (i.e., “MLAGPEER”)
- IP address assigned to interface VLAN 4094 (the MLAG peer VLAN)
- IP routing enabled (to allow for control traffic to be exchanged between MLAG peers)
- Spanning tree disabled for VLAN4094
- Port channel 1000 created and used as MLAG Peer Link, mode set to “trunk” and assigned to the MLAG trunk group (i.e., “MLAGPEER”)
- Peer link Ethernet interfaces added to port channel 1000
- MLAG configuration created, including:
    - MLAG domain ID
    - Local interface (i.e., VLAN 4094)
    - Peer IP address
    - Peer link (i.e., port channel 1000)
- A port channel created to connect to the DC Network switches (i.e., port channel 10”), mode set to “trunk” and assigned as part of an MLAG (i.e., MLAG 10)

The following configuration file sections show the relevant MLAG configuration for each of the DCI switches.

Switch “dcia-s1”

```
no spanning-tree vlan 4094
!
vlan 4094
   trunk group MLAGPEER
!
interface Port-Channel10
   switchport mode trunk
   mlag 10
!
interface Port-Channel1000
   description MLAG-PeerLink
   switchport mode trunk
   switchport trunk group MLAGPEER
!
interface Ethernet21
   channel-group 1000 mode active
!
interface Ethernet22
   channel-group 1000 mode active
!
interface Ethernet3
   channel-group 10 mode active
   lacp rate fast
!
interface Vlan4094
   ip address 192.168.10.1/24
!
ip routing
```

```
!
mlag configuration
   domain-id MLAG-VTEP-DC-A
   local-interface Vlan4094
   peer-address 192.168.10.2
   peer-link Port-Channel1000
!
```

Switch “dcia-s2”

```
no spanning-tree vlan 4094
!
vlan 4094
   trunk group MLAGPEER
!
interface Port-Channel10
   switchport mode trunk
   mlag 10
!
interface Port-Channel1000
   description MLAG-PeerLink
   switchport mode trunk
   switchport trunk group MLAGPEER
!
interface Ethernet21
   channel-group 1000 mode active
!
interface Ethernet22
   channel-group 1000 mode active
!
interface Ethernet3
   channel-group 10 mode active
   lacp rate fast
!
interface Vlan4094
   ip address 192.168.10.2/24
!
ip routing
!
mlag configuration
   domain-id MLAG-VTEP-DC-A
   local-interface Vlan4094
   peer-address 192.168.10.1
   peer-link Port-Channel1000
!
```

Switch “dcib-s1”

```
no spanning-tree vlan 4094
!
vlan 4094
   trunk group MLAGPEER
!
interface Port-Channel10
   switchport mode trunk
   mlag 10
!
interface Port-Channel1000
   description MLAG-PeerLink
   switchport mode trunk
   switchport trunk group MLAGPEER
!
interface Ethernet21
   channel-group 1000 mode active
!
interface Ethernet22
   channel-group 1000 mode active
!
interface Ethernet3
   channel-group 10 mode active
   lacp rate fast
!
interface Vlan4094
   ip address 192.168.11.1/24
!
ip routing
!
mlag configuration
   domain-id MLAG-VTEP-DC-B
   local-interface Vlan4094
   peer-address 192.168.11.2
   peer-link Port-Channel1000
!
```

Switch “dcib-s2”

```
no spanning-tree vlan 4094
!
vlan 4094
   trunk group MLAGPEER
!
interface Port-Channel10
   switchport mode trunk
   mlag 10
!
interface Port-Channel1000
   description MLAG-PeerLink
   switchport mode trunk
   switchport trunk group MLAGPEER
!
interface Ethernet21
   channel-group 1000 mode active
!
interface Ethernet22
   channel-group 1000 mode active
!
interface Ethernet3
   channel-group 10 mode active
   lacp rate fast
!
interface Vlan4094
   ip address 192.168.11.2/24
!
ip routing
!
mlag configuration
   domain-id MLAG-VTEP-DC-B
   local-interface Vlan4094
   peer-address 192.168.11.1
   peer-link Port-Channel1000
!
```

VXLAN AND LAYER-3 TOPOLOGIES

Figure 13 shows the VXLAN and layer-3 configuration and topology details used for this example configuration.

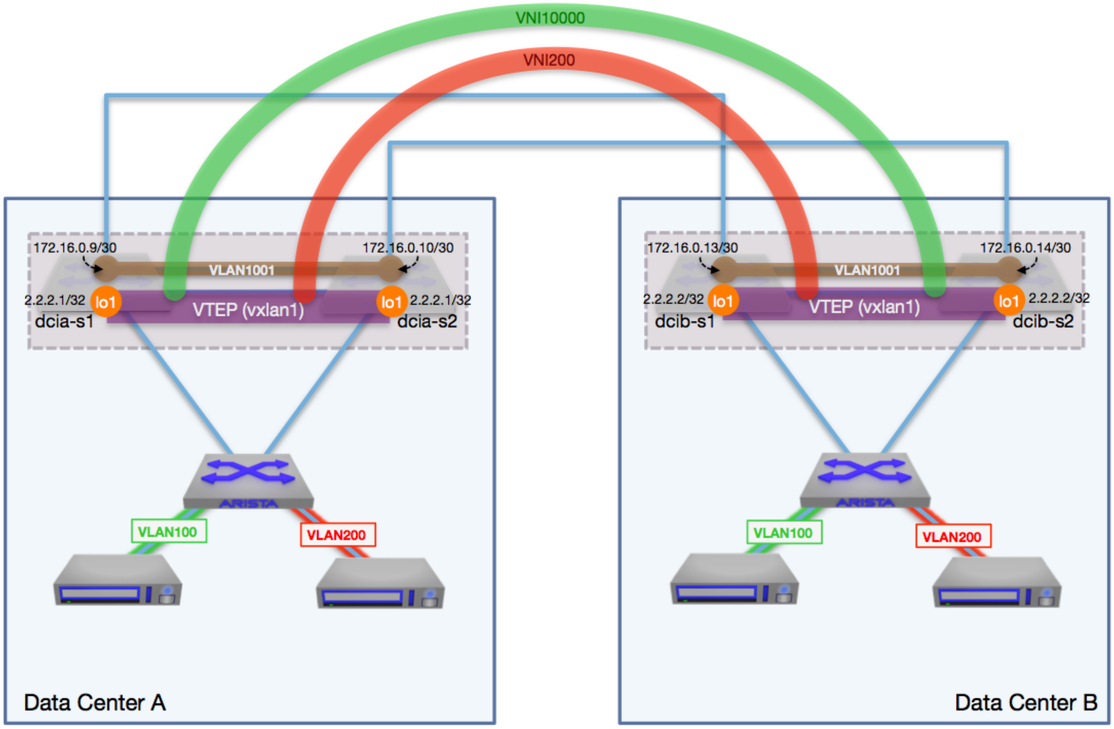

Figure 13: Example DCI Solution – VLAN & VXLAN Topology

Switch Configurations – VXLAN & Layer-3 Forwarding

The following configuration examples show the non-default configuration required to build the above topology, including the implementation of a VXLAN logical VTEP spanning both MLAG members:

- VLANs requiring layer-2 DCI connectivity (e.g. VLAN100 and VLAN200) are configured on both switches and are allowed on the MLAG interfaces.
- A loopback interface (i.e. "Loopback1") is created and assigned an IP address —this will become the VTI address. This address must be the same on both switches in the MLAG pair. Ensure that this subnet is routed and reachable.
- Create a layer-3 routed path between both switches in the MLAG pair in order to provide a back-up forwarding path for VXLAN-encapsulated traffic in case of an inter-site link failure:
  - Create a VLAN (i.e., VLAN1001)
  - Assign an IP address to this VLAN
  - Ensure that this subnet is routed and reachable
- Create interface VXLAN1
- Assign Loopback1 as the source interface
- Change the default UDP port number for VXLAN traffic if required (default is 4789)
- Map each VLAN requiring DCI connectivity to a VNI, e.g.,

    - VLAN100 mapped to VNI10000
    - VLAN200 mapped to VNI200
- Specify the IP address of any remote VTEPs that should be part of the DCI connection for the purpose of receiving unknown unicast, broadcast and multicast flooded traffic via the Head End Replication mechanism
- Configure IP routing (e.g., OSPF, BGP etc.) for all subnets requiring layer-3 reachability (see above)

The following configuration file sections show the relevant VLAN, VXLAN and routing configuration for each of the DCI switches.

Switch Configurations – VLANs, VXLAN and Routing

Switch "dcia-s1"

```
!
vlan 100
   name DCI-VLAN100
!
vlan 200
   name DCI-VLAN200
!
vlan 1001
   name VXLAN-l3-failover-path
!
interface Loopback1
   ip address 2.2.2.1/32
!
interface Vlan1001
   ip address 172.16.0.9/30
!
interface Vxlan1
   vxlan source-interface Loopback1
   vxlan udp-port 4789
   vxlan vlan 100 vni 10000
   vxlan vlan 200 vni 200
   vxlan flood vtep 2.2.2.2
!
router ospf 1
   router-id 10.0.0.1
   network 2.2.2.0/24 area 0.0.0.0
   network 172.16.0.0/24 area 0.0.0.0
   max-lsa 12000
!
```

Switch “dcia-s2”

```
!
vlan 100
   name DCI-VLAN100
!
vlan 200
   name DCI-VLAN200
!
vlan 1001
   name VXLAN-l3-failover-path
!
interface Loopback1
   ip address 2.2.2.1/32
!
interface Vlan1001
   ip address 172.16.0.10/30
!
interface Vxlan1
   vxlan source-interface Loopback1
   vxlan udp-port 4789
   vxlan vlan 100 vni 10000
   vxlan vlan 200 vni 200
   vxlan flood vtep 2.2.2.2
!
router ospf 1
   router-id 10.0.0.1
   network 2.2.2.0/24 area 0.0.0.0
   network 172.16.0.0/24 area 0.0.0.0
   max-lsa 12000
!
```

Switch “dcib-s1”

```
!
vlan 100
   name DCI-VLAN100
!
vlan 200
   name DCI-VLAN200
!
vlan 1001
   name VXLAN-l3-failover-path
!
interface Loopback1
   ip address 2.2.2.2/32
!
interface Vlan1001
   ip address 172.16.0.13/30
!
interface Vxlan1
   vxlan source-interface Loopback1
   vxlan udp-port 4789
   vxlan vlan 100 vni 10000
   vxlan vlan 200 vni 200
   vxlan flood vtep 2.2.2.1
!
router ospf 1
   router-id 10.0.0.1
   network 2.2.2.0/24 area 0.0.0.0
   network 172.16.0.0/24 area 0.0.0.0
   max-lsa 12000
!
```

Switch “dcib-s2”

```
!
vlan 100
   name DCI-VLAN100
!
vlan 200
   name DCI-VLAN200
!
vlan 1001
   name VXLAN-l3-failover-path
!
interface Loopback1
   ip address 2.2.2.2/32
!
interface Vlan1001
   ip address 172.16.0.14/30
!
interface Vxlan1
   vxlan source-interface Loopback1
   vxlan udp-port 4789
   vxlan vlan 100 vni 10000
   vxlan vlan 200 vni 200
   vxlan flood vtep 2.2.2.1
!
router ospf 1
   router-id 10.0.0.1
   network 2.2.2.0/24 area 0.0.0.0
   network 172.16.0.0/24 area 0.0.0.0
   max-lsa 12000
!
```

VERIFYING & MONITORING THE ARISTA VXLAN DCI SOLUTION

It is of course possible to verify basic connectivity between data centers using the ping command. There are also a number of commands available to check and verify the status of the Arista VXLAN DCI solution.

1. Verify MLAG is operational:

```
dcia-s1#show mlag
MLAG Configuration:
domain-id            :      MLAG-VTEP-DC-A
local-interface      :            Vlan4094
peer-address         :        192.168.10.2
peer-link            :    Port-Channel1000

MLAG Status:
state                :              Active
negotiation status   :           Connected
peer-link status     :                  Up
local-int status     :                  Up
system-id            :   02:1c:73:00:44:d6

MLAG Ports:
Disabled             :                   0
Configured           :                   0
Inactive             :                   0
Active-partial       :                   0
Active-full          :                   1
dcia-s1#
```

2. Check IP reachability (i.e., use ping, show ip routes etc. to verify that VTEPs, default gateways and end devices are reachable):

```
dcia-s1#show ip route
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP,
       R - RIP, I - ISIS, A B - BGP Aggregate, A O - OSPF Summary,
       NG - Nexthop Group Static Route

Gateway of last resort:
 S      0.0.0.0/0 [1/0] via 192.168.1.254, Management1

 C      2.2.2.1/32 is directly connected, Loopback1
 O      2.2.2.2/32 [110/20] via 172.16.0.2, Ethernet1
 C      172.16.0.0/30 is directly connected, Ethernet1
 O      172.16.0.4/30 [110/20] via 172.16.0.10, Vlan1001
 C      172.16.0.8/30 is directly connected, Vlan1001
 O      172.16.0.12/30 [110/20] via 172.16.0.2, Ethernet1
 C      192.168.0.0/22 is directly connected, Management1
 C      192.168.10.0/24 is directly connected, Vlan4094

dcia-s1#
```

3. Verify that VLANs requiring inter-DC connectivity are configured correctly:

```
dcia-s1#show vlan
VLAN  Name                             Status    Ports
----- -------------------------------- --------- -------------------------------
1     default                          active    Po10, Po1000
100   DCI-VLAN100                      active    Po10, Po1000, Vx1
200   DCI-VLAN200                      active    Po10, Po1000, Vx1
1001  VXLAN-l3-failover-path           active    Cpu, Po10, Po1000
4094  VLAN4094                         active    Cpu, Po1000

dcia-s1#
```

4. Verify that basic layer-2 forwarding is as expected (i.e.., MAC addresses are being learned correctly):

```
dcia-s1#show mac address-table
          Mac Address Table
------------------------------------------------------------------

Vlan    Mac Address       Type        Ports      Moves   Last Move
----    -----------       ----        -----      -----   ---------
 100    001c.730c.1074    DYNAMIC     Po10       1       0:06:29 ago
 100    001c.7310.3d1c    DYNAMIC     Vx1        1       0:06:29 ago
 200    001c.730c.1074    DYNAMIC     Po10       1       0:06:43 ago
 200    001c.7310.3d1c    DYNAMIC     Vx1        1       0:06:43 ago
1001    001c.731e.e5ee    STATIC      Po1000
4094    001c.731e.e5ee    STATIC      Po1000
Total Mac Addresses for this criterion: 6

          Multicast Mac Address Table
------------------------------------------------------------------

Vlan    Mac Address       Type        Ports
----    -----------       ----        -----
Total Mac Addresses for this criterion: 0

dcia-s1#
```

5. Verify layer-2 learning on VXLAN interfaces:

```
dcia-s1#show vxlan address-table
          Vxlan Mac Address Table
----------------------------------------------------------------------

Vlan  Mac Address     Type     Prt  Vtep             Moves   Last Move
----  -----------     ----     ---  ----             -----   ---------
 100  001c.7310.3d1c  DYNAMIC  Vx1  2.2.2.2          1       0:06:35 ago
 200  001c.7310.3d1c  DYNAMIC  Vx1  2.2.2.2          1       0:06:49 ago
Total Remote Mac Addresses for this criterion: 2
s7151#sho vxlan vtep
Remote vteps for Vxlan1:
2.2.2.2
Total number of remote vteps:  1
dcia-s1#
```

6. Verify that remote VTEPs are visible:

```
dcia-s1#show vxlan vtep
Remote vteps for Vxlan1:
2.2.2.2
Total number of remote vteps:  1
dcia-s1#
```

7. Ensure that Head End Replication is being performed for the required VLANs:

```
dcia-s1#show vxlan flood vtep
          Vxlan Flood Vtep Table
--------------------------------------------------------

Vlan   Ip Address
----   -------------------------------------------------
100    2.2.2.2
200    2.2.2.2
dcia-s1#
```

OPERATION OF THE VXLAN DCI SOLUTION WITH FAILURE CONDITIONS

In this section we will consider the impact on traffic forwarding for a range of different failure conditions. The example solution described in the previous section has in-built redundancy and is capable of rapid and graceful failover (and failback) for a number of possible failure conditions. Redundancy features include:

- Highly available and reliable switching hardware with redundant power supplies and fans
- Arista's EOS, a robust, modular switch operating system supporting process restart and fault containment
- Multi-chassis link aggregation (MLAG) for active-active layer-2 forwarding with 2 switches
- Arista VXLAN + MLAG, a highly available, high-performance hardware VTEP implementation delivering active-active VXLAN bridging across a pair of switches at line rate and low latency

For the architecture described in the preceding example design, there are two possible failure scenarios that are worth considering in more detail:

1. The failure of an MLAG member link, i.e., a link or interface failure between the DC network edge leaf and the DCI module.
2. The failure of an inter-site link, i.e., a link or interface failure between the DCI modules.

For each of the above scenarios, the effect of the failure and the impact on traffic forwarding is considered in terms of both a local failure (i.e., impact on traffic sourced from the same DC where the failure occurred) and a remote failure (i.e., impact on traffic sourced from the DC not directly connected to failed links/switches).

Before considering the various failover scenarios, it is worthwhile reviewing the forwarding decision mechanisms for normal operation.

NORMAL OPERATION

The normal MLAG+VXLAN traffic forwarding mechanism is as follows (as shown in see Figure 14):

1. Traffic from the leaf destined for DCI forwarding is balanced based on port channel LAG policy and forwarded to one of the MLAG peer switches.

2. The local VTEP on the switch receiving layer-2 frames to be forwarded to the remote location encapsulates them with a VXLAN header based on the VLAN-to-VNI mapping. The source IP address is the local VTEP address; the destination IP address is that of the remote VTEP.
3. The VXLAN-encapsulated packet will be forwarded on the interface identified by the IP routing table entry as providing the best path, i.e., under normal conditions, the Ethernet link to remote site.
4. The remote VTEP de-encapsulates traffic and forwards it on the local MLAG port channel member to the downstream switch.

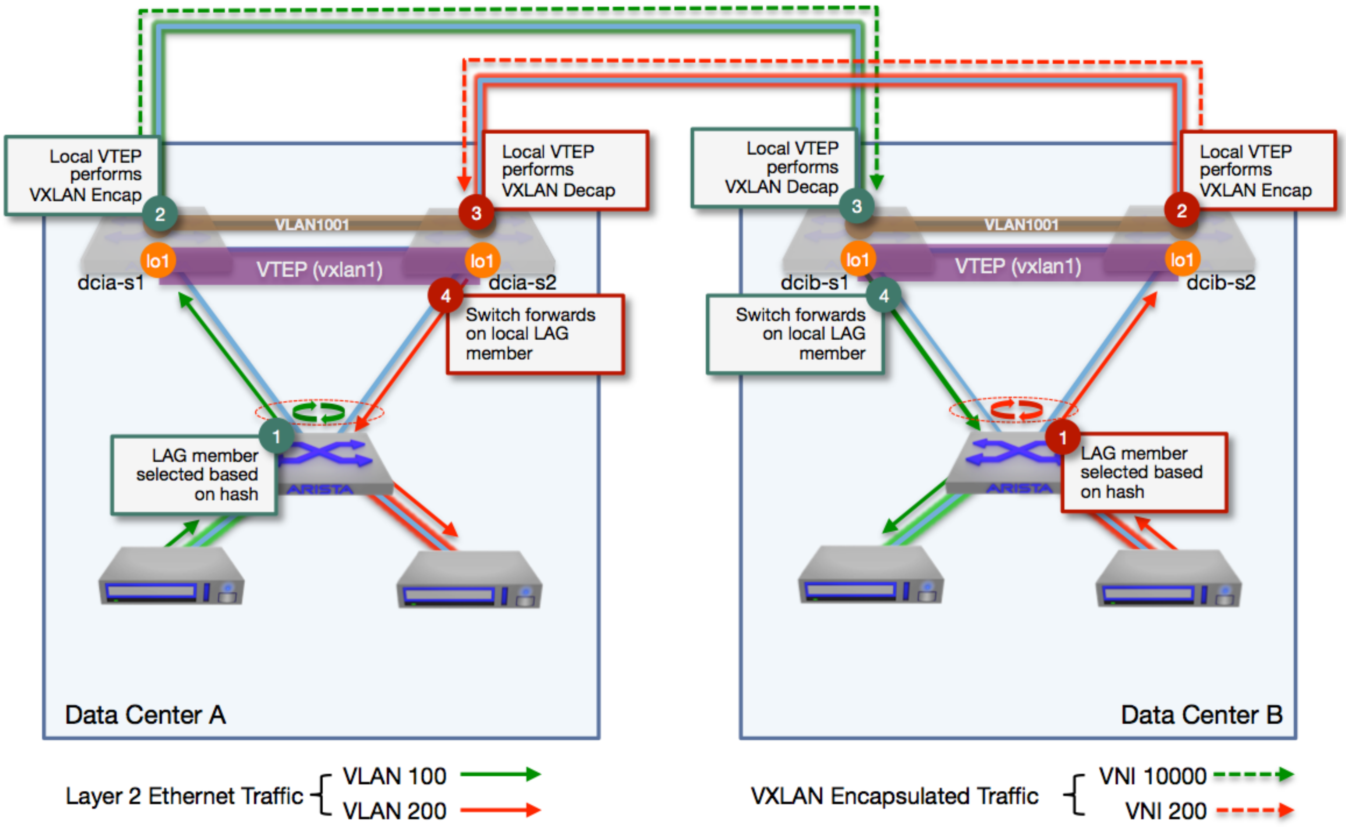
Local VTEP performs VXLAN Encap
2
Local VTEP performs VXLAN Decap
3
Local VTEP performs VXLAN Decap
3
Local VTEP performs VXLAN Encap
2
VLAN1001
lo1
VTEP (vxlan1)
lo1
dcia-s1
dcia-s2
4
Switch forwards on local LAG member
VLAN1001
lo1
VTEP (vxlan1)
lo1
dcib-s1
dcib-s2
4
Switch forwards on local LAG member
1
LAG member selected based on hash
ARISTA
1
LAG member selected based on hash
ARISTA
Data Center A
Data Center B
Layer 2 Ethernet Traffic
VLAN 100
VLAN 200
VXLAN Encapsulated Traffic
VNI 10000
VNI 200

Figure 14: Normal VXLAN+MLAG forwarding Mechanism

AN MLAG MEMBER LINK FAILURE

In the situation where a member port of the LAG or the MLAG is disabled or fails (leaving only a single path from DC network edge leaf to the DCI module), traffic will be re-balanced onto the available member port. For local traffic (i.e., traffic sourced from the same site as the failure), the following failover process occurs (as shown in Figure 15):

1. The sending DC network edge module leaf switch detects LAG member failure and rebalances sent traffic to the available link (i.e., to “dcia-s2” in the example below).
2. The VTEP on “dcia-s2” will now encapsulate all traffic with a VXLAN header based on the VLAN-to-VNI mapping. The source IP address is the local VTEP address; the destination IP address is that of the remote VTEP.
3. The VXLAN-encapsulated packet will be forwarded on the layer-3 path determined by the routing table, i.e., the Ethernet link to the remote site.

4. Remote VTEP de-encapsulates traffic and forwards it on the local MLAG port channel member to the downstream switch.

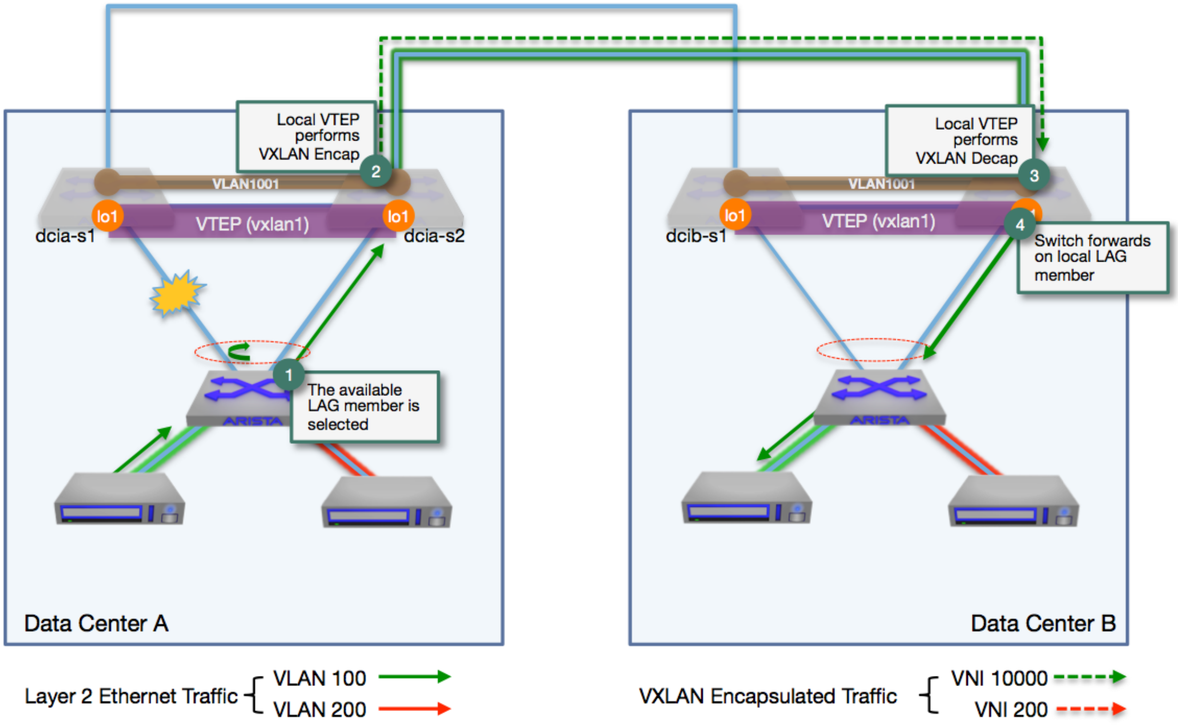

Figure 15: Effect of Local MLAG Member Port Failure

In the case of a remote MLAG port failure, the following failover mechanism comes into effect (see Figure 16):

1. The remote DC network edge leaf switch is unaware of the remote MLAG member failure, so it continues to forward traffic as before (i.e., based on the port channel load balance hash).
2. The remote DCI switches are also unaware of the MLAG member failure, so the VTEPs will encapsulate the traffic with a VXLAN header and continue to forward traffic as before (i.e., based on the layer-3 routing table). In the example below, VTEP on "dcib-s1" forwards to the VTEP on "dcia-s2".
3. The VTEP (e.g., on switch "dcia-s1") receiving the VXLAN-encapsulated traffic removes the VXLAN and maps the appropriate VLAN.
4. Switch "dcia-s1" is aware that the local MLAG member port is down and therefore uses the peer link to relay the traffic to "dcia-s2".
5. Switch "dcia-s2" forwards as per the MAC address table, i.e., on the remaining MLAG member port.

For each case, failure detection and failover performance is based on LAG and MLAG link failure detection mechanisms. In lab testing, failover was performed in approximately 100-200msecs.

With this type of failure, both inter-DC links continue to carry traffic. This is deemed to be the preferred mechanism as it is likely that DCI bandwidth will be lower (and probably more costly), whereas the MLAG peer link bandwidth can be easily provisioned to provide ample failover capacity.

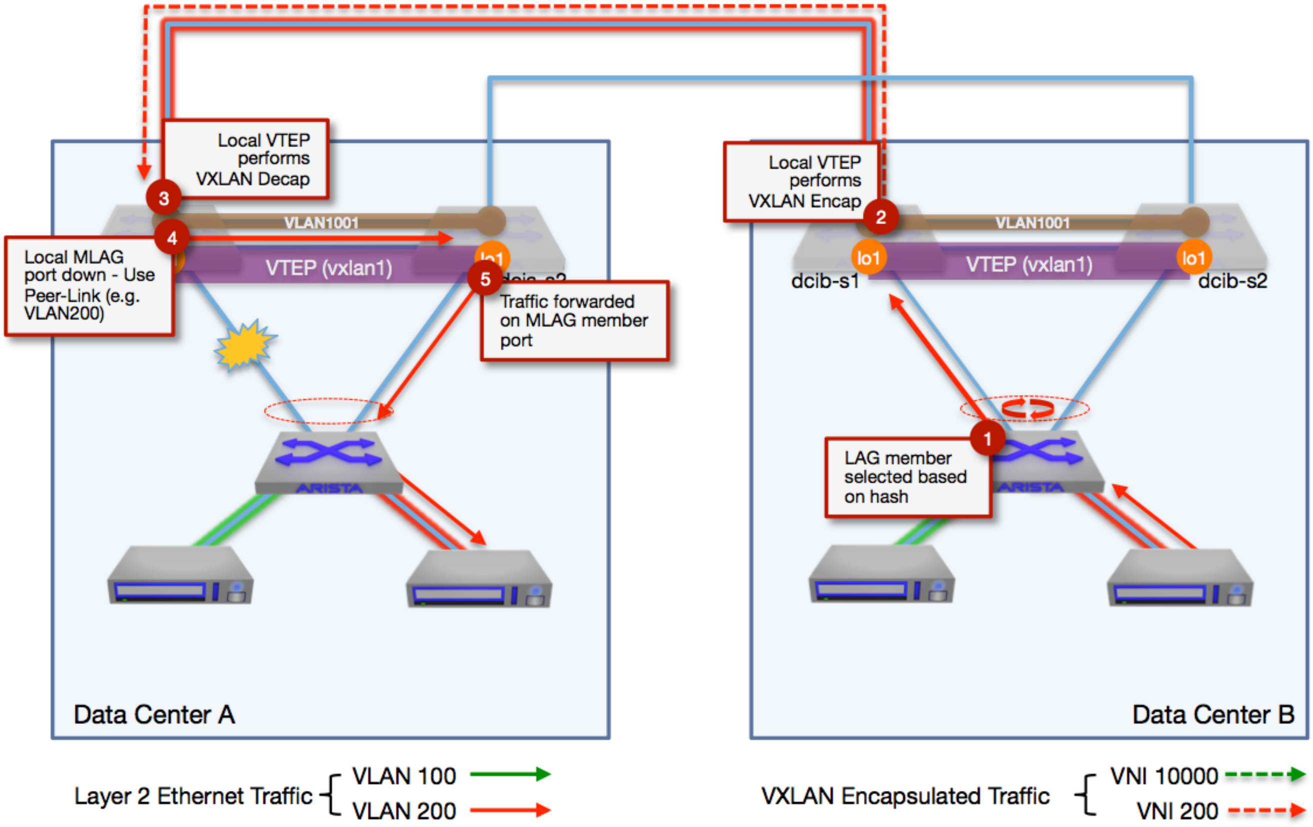
Local VTEP performs VXLAN Decap
3
4
VLAN1001
Local MLAG port down - Use Peer-Link (e.g. VLAN200)
VTEP (vxlan1)
lo1
dcia-s2
5
Traffic forwarded on MLAG member port
ARISTA
Data Center A
Local VTEP performs VXLAN Encap
2
VLAN1001
lo1
VTEP (vxlan1)
lo1
dcib-s1
dcib-s2
1
LAG member selected based on hash
ARISTA
Data Center B
Layer 2 Ethernet Traffic
VLAN 100
VLAN 200
VXLAN Encapsulated Traffic
VNI 10000
VNI 200

Figure 16: Effect of Remote MLAG Member Port Failure

AN INTER-SITE LINK FAILURE

In the situation where an inter-site link is disabled or fails (leaving only a single path between DCs), traffic will be re-routed using the next lowest cost layer-3 path, i.e., failover path selection is performed after the DCI traffic has been encapsulated with the VXLAN header. For local traffic (i.e., traffic sourced from the same site as the failure), the failover process occurs as follows (see Figure 17):

1. The DC network edge leaf switches are unaware of the DCI link failure and will forward based on LAG hashing mechanism, e.g., to “dcia-s1”.
2. The local VTEP will encapsulate all traffic with a VXLAN header based on the VLAN-to-VNI mapping. The source IP address is the local VTEP address, and the destination IP address is that of the remote VTEP.
3. The VXLAN-encapsulated packet will be forwarded on the layer-3 path based on the routing table entry, i.e., via VLAN1001. In the example below “dcia-s1” forwards VXLAN-encapsulated frames via VLAN1001 to switch “dcia-s2”, which in turn forwards via its inter-site line to “dcib-s2”.
4. The remote VTEP de-encapsulates traffic and forwards it on the local MLAG port channel member to the downstream switch.

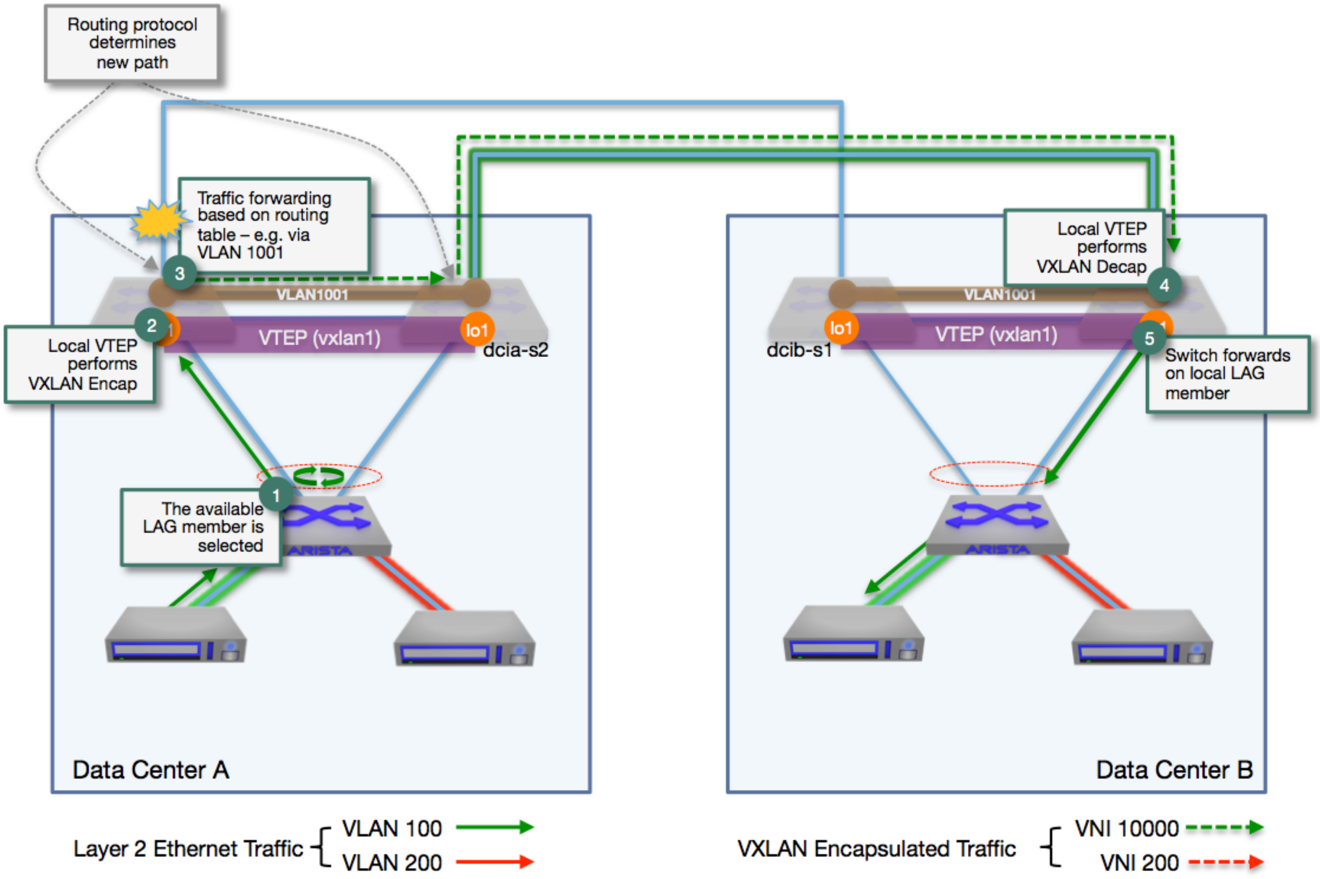
Routing protocol determines new path
Traffic forwarding based on routing table – e.g. via VLAN 1001
3
VLAN1001
2
Local VTEP performs VXLAN Encap
VTEP (vxlan1)
lo1
dcia-s2
1
The available LAG member is selected
ARISTA
Data Center A
Local VTEP performs VXLAN Decap
4
VLAN1001
lo1
dcib-s1
VTEP (vxlan1)
5
Switch forwards on local LAG member
ARISTA
Data Center B
Layer 2 Ethernet Traffic
VLAN 100
VLAN 200
VXLAN Encapsulated Traffic
VNI 10000
VNI 200

Figure 17: Effect of Local DCI Inter-site Link Failure

Within the example topology used within this design guide, the situation for a remote failure of a DCI inter-site link is virtually identical to the local failure scenario (as shown in Figure 18), i.e.:

1. The DC network edge switches are unaware of the DCI link failure and will forward based on the LAG hashing mechanism e.g., to “dcib-s1”.
2. The local VTEP will encapsulate all traffic with a VXLAN header based on the VLAN-to-VNI mapping. The source IP address is the local VTEP address and the destination IP address is that of the remote VTEP.
3. The VXLAN-encapsulated packet will be forwarded on layer-3 based on the routing table entry, i.e., via VLAN1001. In the example below, “dcib-s1” forwards VXLAN-encapsulated frames via VLAN1001 to switch “dcib-s2”, which in turn forwards via its inter-site link to “dcia-s2”.
4. The remote VTEP de-encapsulates traffic and forwards it on the local MLAG port channel member to the downstream switch.

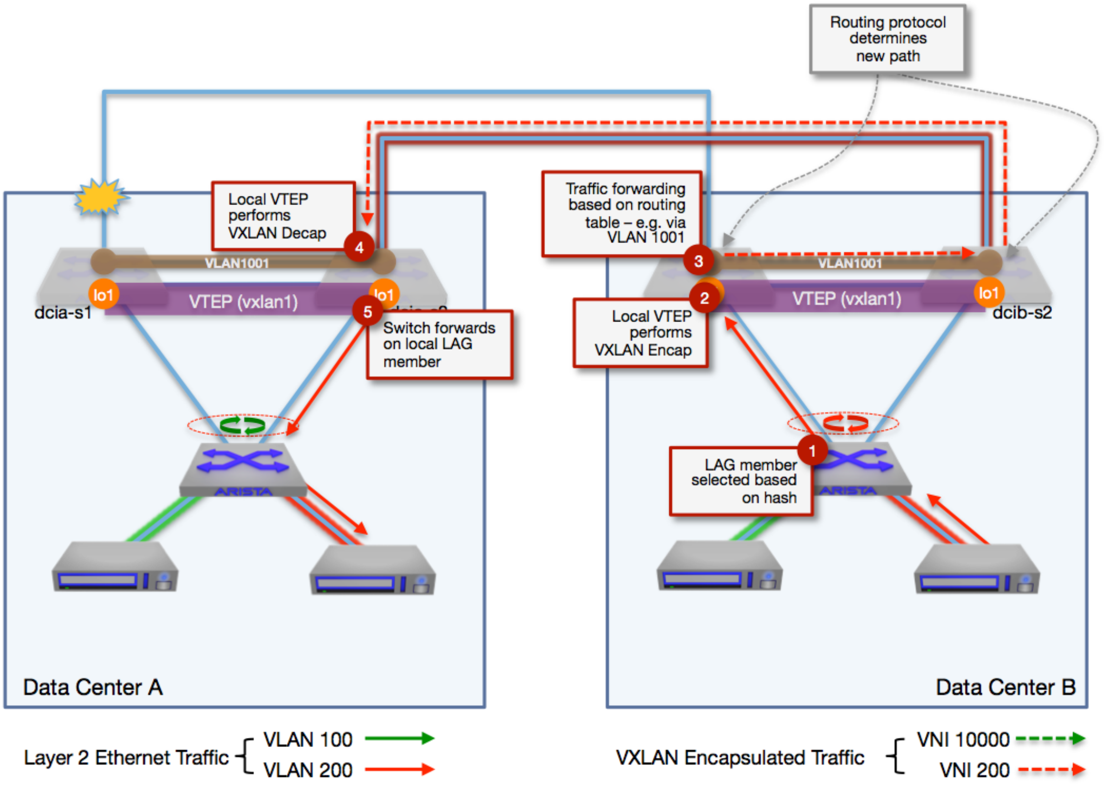
Routing protocol determines new path
Local VTEP performs VXLAN Decap
4
Traffic forwarding based on routing table – e.g. via VLAN 1001
3
VLAN1001
VLAN1001
lo1
lo1
lo1
dcia-s1
VTEP (vxlan1)
5
Switch forwards on local LAG member
2
Local VTEP performs VXLAN Encap
VTEP (vxlan1)
dcib-s2
ARISTA
1
LAG member selected based on hash
ARISTA
Data Center A
Data Center B
Layer 2 Ethernet Traffic
VLAN 100
VLAN 200
VXLAN Encapsulated Traffic
VNI 10000
VNI 200

Figure 18: Effect of Remote DCI Inter-Site Link Failure

In both the situations, the failover performance is wholly dependent on layer-3 routing convergence. In the example used in this guide, lab testing showed failover took less the 10msec

CONCLUSION

The Arista VXLAN DCI solution provides the first truly open and cost-effective solution for providing layer-2 connectivity between data centers over a layer-3 transport network. It utilizes standard switching hardware and is based on technologies being deployed throughout modern virtualized data centers, requiring no special expertise, management tools or significant changes to operating procedures. It can be deployed as a permanent facility to enable layer-2 adjacency for server or storage clustering as well as VM mobility and high availability. It can also be deployed as a temporary solution for data center migration purposes, with the option to redeploy the hardware as standard DC switches after the migration is completed.

Finally, the Arista VXLAN DCI solution is built on Arista EOS, the world's most robust and extensible switch operating system, offering advanced programmability and comprehensive protocol support, all backed up by a world-class support organization.

Santa Clara—Corporate Headquarters
5453 Great America Parkway
Santa Clara, CA 95054
Tel: 408-547-5500
www.arista.com

Ireland—International Headquarters
4130 Atlantic Avenue
West park Business Campus
Shannon
Co. Clare, Ireland

Singapore—APAC Administrative Office
9 Tease Boulevard
#29-01, Sundeck Tower Two
Singapore 038989