CHAPTER 1

OTV Technology Introduction and Deployment Considerations

This document introduces a Cisco innovative LAN extension technology called Overlay Transport Virtualization (OTV). OTV is an IP-based functionality that has been designed from the ground up to provide Layer 2 extension capabilities over any transport infrastructure: Layer 2 based, Layer 3 based, IP switched, label switched, and so on. The only requirement from the transport infrastructure is providing IP connectivity between remote data center sites. In addition, OTV provides an overlay that enables Layer 2 connectivity between separate Layer 2 domains while keeping these domains independent and preserving the fault-isolation, resiliency, and load-balancing benefits of an IP-based interconnection.

As of this writing, the Nexus 7000 is the only Cisco platform supporting OTV. All the technology and deployment considerations contained in this paper focus on positioning the Nexus 7000 platforms inside the data center to establish Layer 2 connectivity between remote sites. OTV support on Nexus 7000 platforms has been introduced from the NX-OS 5.0(3) software release. When necessary, available OTV features will be identified in the current release or mentioned as a future roadmap function. This document will be periodically updated every time a software release introduces significant new functionality.

OTV Technology Primer

Before discussing OTV in detail, it is worth differentiating this technology from traditional LAN extension solutions such as EoMPLS and VPLS.

OTV introduces the concept of "MAC routing," which means a control plane protocol is used to exchange MAC reachability information between network devices providing LAN extension functionality. This is a significant shift from Layer 2 switching that traditionally leverages data plane learning, and it is justified by the need to limit flooding of Layer 2 traffic across the transport infrastructure. As emphasized throughout this document, Layer 2 communications between sites resembles routing more than switching. If the destination MAC address information is unknown, then traffic is dropped (not flooded), preventing waste of precious bandwidth across the WAN.

OTV also introduces the concept of dynamic encapsulation for Layer 2 flows that need to be sent to remote locations. Each Ethernet frame is individually encapsulated into an IP packet and delivered across the transport network. This eliminates the need to establish virtual circuits, called Pseudowires, between the data center locations. Immediate advantages include improved flexibility when adding or removing sites to the overlay, more optimal bandwidth utilization across the WAN (specifically when the transport infrastructure is multicast enabled), and independence from the transport characteristics (Layer 1, Layer 2 or Layer 3).

Finally, OTV provides a native built-in multi-homing capability with automatic detection, critical to increasing high availability of the overall solution. Two or more devices can be leveraged in each data center to provide LAN extension functionality without running the risk of creating an end-to-end loop that would jeopardize the overall stability of the design. This is achieved by leveraging the same control plane protocol used for the exchange of MAC address information, without the need of extending the Spanning-Tree Protocol (STP) across the overlay.

The following sections detail the OTV technology and introduce alternative design options for deploying OTV within, and between, data centers.

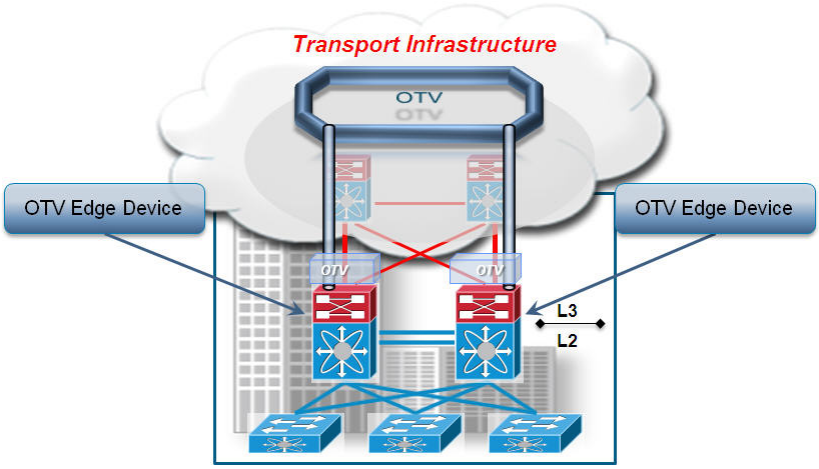
OTV Terminology

Before learning how OTV control and data planes work, you must understand OTV specific terminology.

Edge Device

The edge device (Figure 1-1) performs OTV functions: it receives the Layer 2 traffic for all VLANs that need to be extended to remote locations and dynamically encapsulates the Ethernet frames into IP packets that are then sent across the transport infrastructure.

Figure 1-1 OTV Edge Device



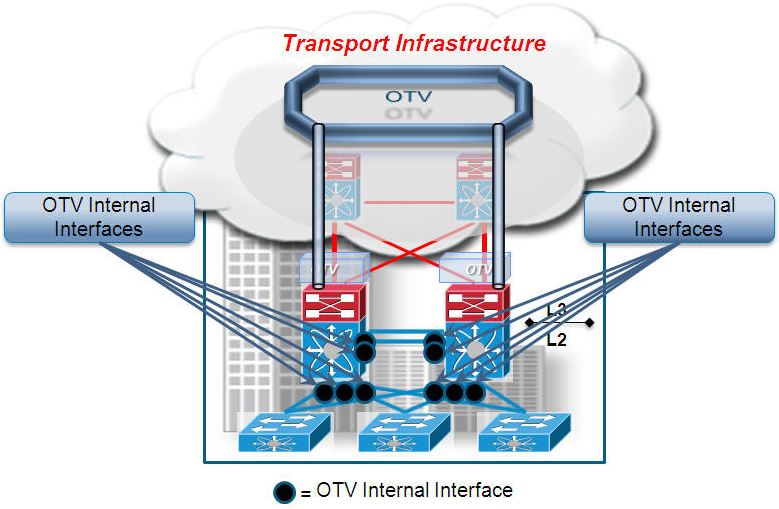
It is expected that at least two OTV edge devices are deployed at each data center site to improve the resiliency, as discussed more fully in Multi-Homing, page 1-21.

Finally, the OTV edge device can be positioned in different parts of the data center. The choice depends on the site network topology. Figure 1-1 shows edge device deployment at the aggregation layer.

Internal Interfaces

To perform OTV functionality, the edge device must receive the Layer 2 traffic for all VLANs that need to be extended to remote locations. The Layer 2 interfaces, where the Layer 2 traffic is usually received, are named internal interfaces (Figure 1-2).

Figure 1-2 OTV Internal Interfaces

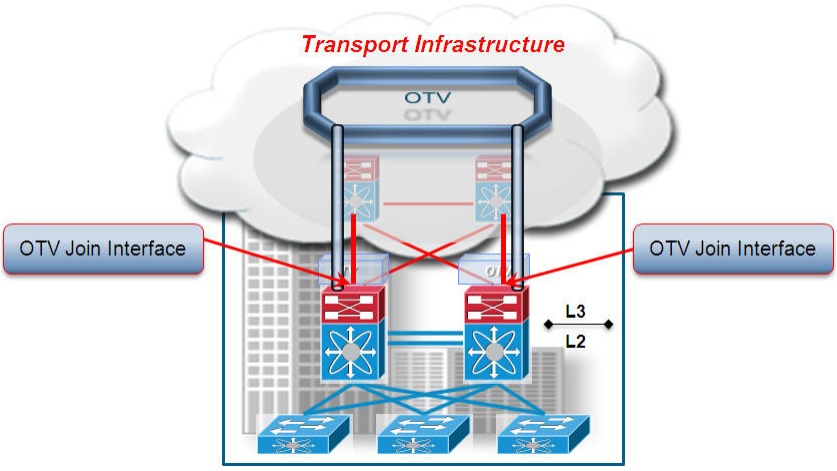


Internal interfaces are regular Layer 2 interfaces configured as access or trunk ports. Trunk configuration is typical given the need to concurrently extend more than one VLAN across the overlay. There is no need to apply OTV-specific configuration to these interfaces. Also, typical Layer 2 functions (like local switching, spanning-tree operation, data plane learning, and flooding) are performed on the internal interfaces. Figure 1-2 shows Layer 2 trunks that are considered internal interfaces which are usually deployed between the edge devices also.

Join Interface

The Join interface (Figure 1-3) is used to source the OTV encapsulated traffic and send it to the Layer 3 domain of the data center network.

Figure 1-3 OTV Join Interface



The Join interface is a Layer 3 entity and with the current NX-OS release can only be defined as a physical interface (or subinterface) or as a logical one (i.e. Layer 3 port channel or Layer 3 port channel subinterface). A single Join interface can be defined and associated with a given OTV overlay. Multiple overlays can also share the same Join interface.

Note Support for loopback interfaces as OTV Join interfaces is planned for a future NX-OS release.

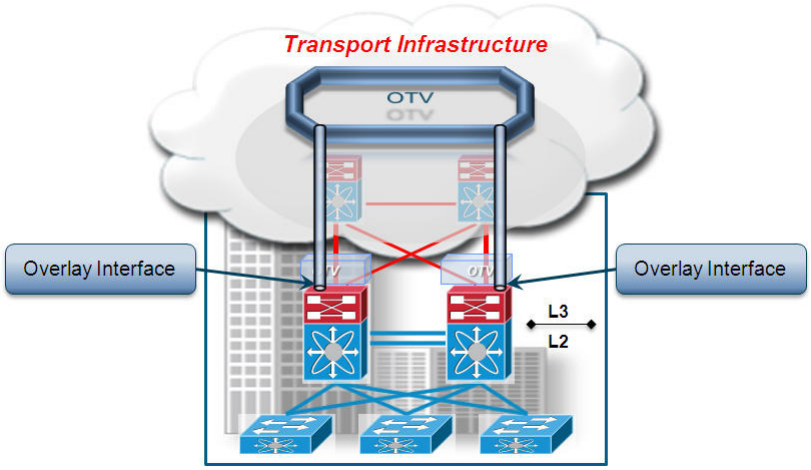
The Join interface is used by the edge device for different purposes:

- "Join" the Overlay network and discover the other remote OTV edge devices.
- Form OTV adjacencies with the other OTV edge devices belonging to the same VPN.
- Send/receive MAC reachability information.
- Send/receive unicast and multicast traffic.

Overlay Interface

The Overlay interface (Figure 1-4) is a logical multi-access and multicast-capable interface that must be explicitly defined by the user and where the entire OTV configuration is applied.

Figure 1-4 OTV Overlay Interface



Every time the OTV edge device receives a Layer 2 frame destined for a remote data center site, the frame is logically forwarded to the Overlay interface. This instructs the edge device to perform the dynamic OTV encapsulation on the Layer 2 packet and send it to the Join interface toward the routed domain.

Control Plane Considerations

As mentioned, one fundamental principle on which OTV operates is the use of a control protocol running between the OTV edge devices to advertise MAC address reachability information instead of using data plane learning. However, before MAC reachability information can be exchanged, all OTV edge devices must become "adjacent" to each other from an OTV perspective. This can be achieved in two ways, depending on the nature of the transport network interconnecting the various sites:

- If the transport is multicast enabled, a specific multicast group can be used to exchange the control protocol messages between the OTV edge devices.
- If the transport is not multicast enabled, an alternative deployment model is available starting from NX-OS release 5.2(1), where one (or more) OTV edge device can be configured as an "Adjacency Server" to which all other edge devices register and communicates to them the list of devices belonging to a given overlay.

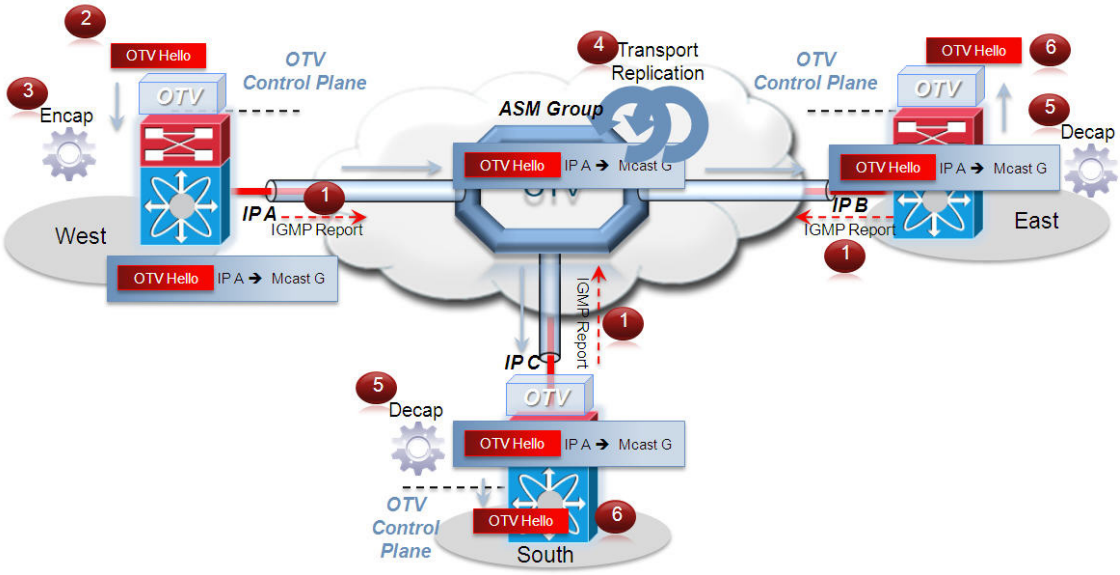
Multicast Enabled Transport Infrastructure

Assuming the transport is multicast enabled, all OTV edge devices can be configured to join a specific ASM (Any Source Multicast) group where they simultaneously play the role of receiver and source. If the transport is owned by a Service Provider, for example, the Enterprise will have to negotiate the use of this ASM group with the SP.

.

Figure 1-5 shows the overall sequence of steps leading to the discovery of all OTV edge devices belonging to the same overlay.

Figure 1-5 OTV Neighbor Discovery



Step 1 Each OTV edge device sends an IGMP report to join the specific ASM group used to carry control protocol exchanges (group G in this example). The edge devices join the group as hosts, leveraging the Join interface. This happens without enabling PIM on this interface. The only requirement is to specify the ASM group to be used and associate it with a given Overlay interface.

Step 2 The OTV control protocol running on the left OTV edge device generates Hello packets that need to be sent to all other OTV edge devices. This is required to communicate its existence and to trigger the establishment of control plane adjacencies.

Step 3 The OTV Hello messages need to be sent across the logical overlay to reach all OTV remote devices. For this to happen, the original frames must be OTV-encapsulated, adding an external IP header. The source IP address in the external header is set to the IP address of the Join interface of the edge device, whereas the destination is the multicast address of the ASM group dedicated to carry the control protocol. The resulting multicast frame is then sent to the Join interface toward the Layer 3 network domain.

Step 4 The multicast frames are carried across the transport and optimally replicated to reach all the OTV edge devices that joined that multicast group G.

Step 5 The receiving OTV edge devices decapsulate the packets.

Step 6 The Hellos are passed to the control protocol process.

The same process occurs in the opposite direction and the end result is the creation of OTV control protocol adjacencies between all edge devices. The use of the ASM group as a vehicle to transport the Hello messages allows the edge devices to discover each other as if they were deployed on a shared LAN segment. The LAN segment is basically implemented via the OTV overlay.

Two important considerations for OTV control protocol are as follows:

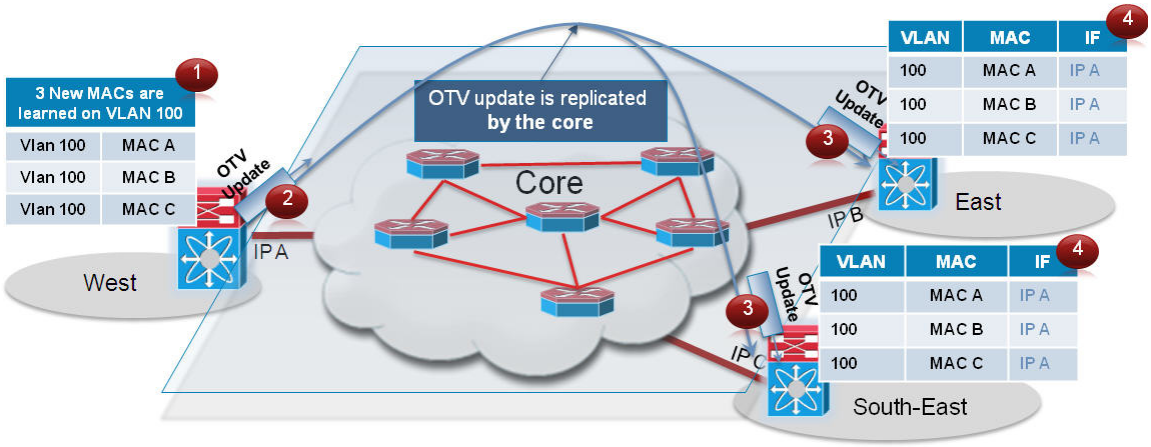
1. This protocol runs as an "overlay" control plane between OTV edge devices which means there is no dependency with the routing protocol (IGP or BGP) used in the Layer 3 domain of the data center, or in the transport infrastructure.
2. The OTV control plane is transparently enabled in the background after creating the OTV Overlay interface and does not require explicit configuration. Tuning parameters, like timers, for the OTV protocol is allowed, but this is expected to be more of a corner case than a common requirement.

Note The routing protocol used to implement the OTV control plane is IS-IS. It was selected because it is a standard-based protocol, originally designed with the capability of carrying MAC address information in the TLV. In the rest of this document, the control plane protocol will be generically called "OTV protocol".

From a security perspective, it is possible to leverage the IS-IS HMAC-MD5 authentication feature to add an HMAC-MD5 digest to each OTV control protocol message. The digest allows authentication at the IS-IS protocol level, which prevents unauthorized routing message from being injected into the network routing domain. At the same time, only authenticated devices will be allowed to successfully exchange OTV control protocol messages between them and hence to become part of the same Overlay network.

Once OTV edge devices have discovered each other, it is then possible to leverage the same mechanism to exchange MAC address reachability information, as shown in Figure 1-6.

Figure 1-6 MAC Address Advertisement



Step 1 The OTV edge device in the West data center site learns new MAC addresses (MAC A, B and C on VLAN 100) on its internal interface. This is done via traditional data plan learning.

Step 2 An OTV Update message is created containing information for MAC A, MAC B and MAC C. The message is OTV encapsulated and sent into the Layer 3 transport. Once again, the IP destination address of the packet in the outer header is the multicast group G used for control protocol exchanges.

Step 3 The OTV Update is optimally replicated in the transport and delivered to all remote edge devices which decapsulate it and hand it to the OTV control process.

Step 4 The MAC reachability information is imported in the MAC Address Tables (CAMs) of the edge devices. As noted in Figure 1-6, the only difference with a traditional CAM entry is that instead of having associated a physical interface, these entries refer the IP address of the Join interface of the originating edge device.

Note MAC table content shown in Figure 1-6 is an abstraction used to explain OTV functionality.

The same control plane communication is also used to withdraw MAC reachability information. For example, if a specific network entity is disconnected from the network, or stops communicating, the corresponding MAC entry would eventually be removed from the CAM table of the OTV edge device. This occurs by default after 30 minutes on the OTV edge device. The removal of the MAC entry triggers an OTV protocol update so that all remote edge devices delete the same MAC entry from their respective tables.

Unicast-Only Transport Infrastructure (Adjacency-Server Mode)

Starting with NX-OS 5.2(1) release, OTV can be deployed with unicast-only transport. As previously described, a multicast enabled transport infrastructure lets a single OTV update or Hello packet reach all other OTV devices by virtue of leveraging a specific multicast control group address.

The OTV control plane over a unicast-only transport works exactly the same way as OTV with multicast mode. The only difference is that each OTV devices would need to create multiple copies of each control plane packet and unicast them to each remote OTV device part of the same logical overlay. Because of this head-end replication behavior, leveraging a multicast enabled transport remains the recommended way of deploying OTV in cases where several DC sites are involved. At the same time, the operational simplification brought by the unicast-only model (removing the need for multicast deployment) can make this deployment option very appealing in scenarios where LAN extension connectivity is required only between few (2-3) DC sites.

To be able to communicate with all the remote OTV devices, each OTV node needs to know a list of neighbors to replicate the control packets to. Rather than statically configuring in each OTV node the list of all neighbors, a simple dynamic means is used to provide this information. This is achieved by designating one (or more) OTV Edge device to perform a specific role, named Adjacency Server. Every OTV device wishing to join a specific OTV logical overlay, needs to first "register" with the Adjacency Server (by start sending OTV Hello messages to it). All other OTV neighbor addresses are discovered dynamically through the Adjacency Server. Thereby, when the OTV service needs to be extended to a new DC site, only the OTV edge devices for the new site need to be configured with the Adjacency Server addresses. No other sites need additional configuration.

The reception of the Hello messages from all the OTV edge devices helps the Adjacency Server to build up the list of all the OTV devices that should be part of the same overlay (named unicast-replication-list). This list is periodically sent in unicast fashion to all the listed OTV devices, so that they can dynamically be aware about all the OTV neighbors in the network.

In Figure 1-7, the OTV Edge device on Site-1 is configured as Adjacency Server. All other OTV edge devices register to this Adjacency Server, which in turn sends the entire neighbor list to each OTV client periodically by means of OTV Hellos.

Figure 1-7 Adjacency Server Functionality

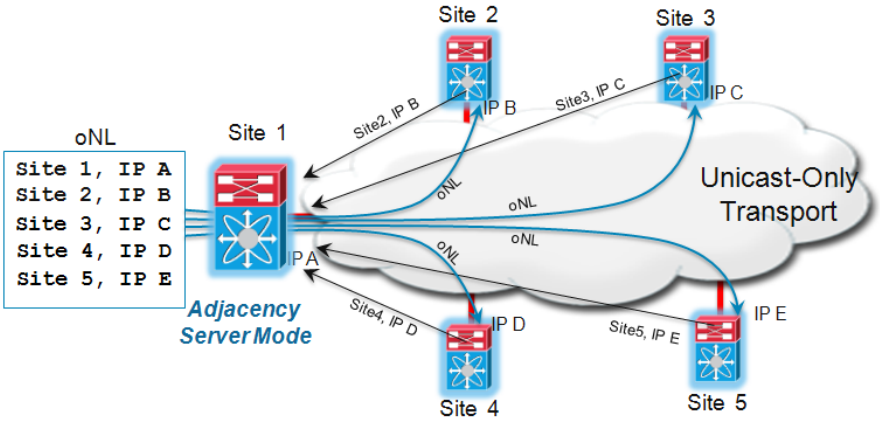
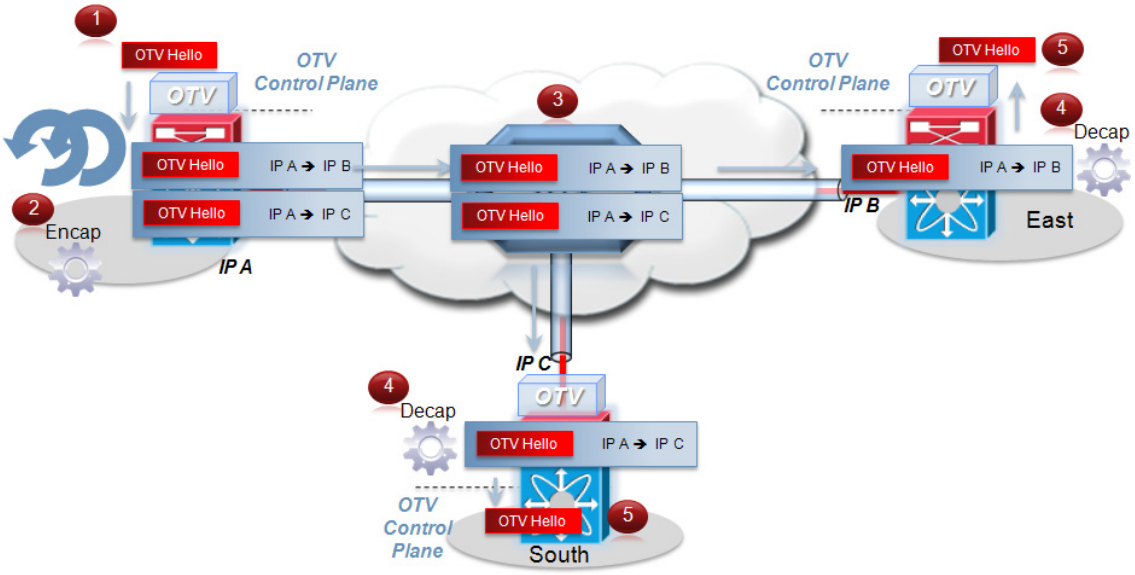


Figure 1-8 shows the overall sequence of steps leading to the establishment of OTV control plane adjacencies between all the OTV edge devices belonging to the same overlay.

Figure 1-8 Creation of OTV Control Plane Adjacencies (Unicast Core)



Step 1 The OTV control protocol running on the left OTV edge device generates Hello packets that need to be sent to all other OTV edge devices. This is required to communicate its existence and to trigger the establishment of control plane adjacencies.

Step 2 The OTV Hello messages need to be sent across the logical overlay to reach all OTV remote devices. For this to happen, the left OTV device must perform head-end replication, creating one copy of the Hello message for each remote OTV device part of the unicast-replication-list previously received from the Adjacency Server. Each of these frames must then be OTV-encapsulated, adding an external IP header. The source IP address in the external header is set to the IP address of the Join interface of the local edge device, whereas the destination is the Join interface address of a remote OTV edge device. The resulting unicast frames are then sent out the Join interface toward the Layer 3 network domain.

Step 3 The unicast frames are routed across the unicast-only transport infrastructure and delivered to their specific destination sites.

Step 4 The receiving OTV edge devices decapsulate the packets.

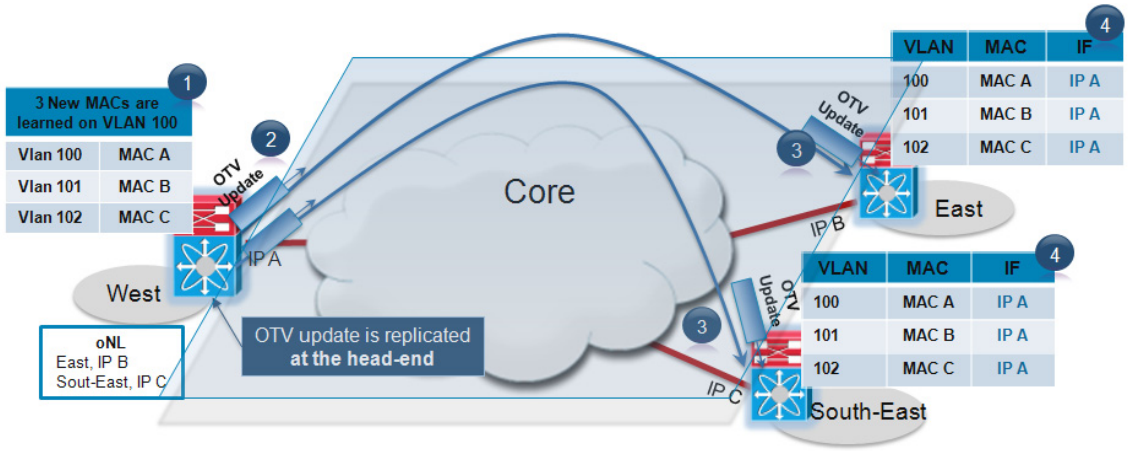
Step 5 The Hellos are passed to the control protocol process.

The same process occurs on each OTV edge device and the end result is the creation of OTV control protocol adjacencies between all edge devices.

The same considerations around the OTV control protocol characteristics already discussed for the multicast transport option still hold valid here (please refer to the previous section for more details).

Once the OTV edge devices have discovered each other, it is then possible to leverage a similar mechanism to exchange MAC address reachability information, as shown in Figure 1-9.

Figure 1-9 MAC Address Advertisement (Unicast Core)



Step 1 The OTV edge device in the West data center site learns new MAC addresses (MAC A, B and C on VLAN 100, 101 and 102) on its internal interface. This is done via traditional data plan learning.

Step 2 An OTV Update message containing information for MAC A, MAC B and MAC C is created for each remote OTV edge device (head-end replication). These messages are OTV encapsulated and sent into the Layer 3 transport. Once again, the IP destination address of the packet in the outer header is the Join interface address of each specific remote OTV device.

Step 3 The OTV Updates are routed in the unicast-only transport and delivered to all remote edge devices which decapsulate them and hand them to the OTV control process.

Step 4 The MAC reachability information is imported in the MAC Address Tables (CAMs) of the edge devices. As noted above, the only difference with a traditional CAM entry is that instead of having associated a physical interface, these entries refer the IP address (IP A) of the Join interface of the originating edge device.

A pair of Adjacency Servers can be deployed for redundancy purposes. These Adjacency Server devices are completely stateless between them, which implies that every OTV edge device (OTV clients) should register its existence with both of them. For this purpose, the primary and secondary Adjacency Servers are configured in each OTV edge device. However, an OTV client will not process an alternate server's replication list until it detects that the primary Adjacency Server has timed out. Once that happens, each OTV edge device will start using the replication list from the secondary Adjacency Server and push the difference to OTV. OTV will stale the replication list entries with a timer of 10 minutes. If the Primary Adjacency Server comes back up within 10 mins, OTV will always revert back to the primary replication list. In case the Primary Adjacency Server comes back up after replication list is deleted, a new replication list will be pushed by the Primary after learning all OTV neighbors by means of OTV Hellos that are sent periodically.

OTV also uses graceful exit of Adjacency Server. When a Primary Adjacency Server is de-configured or is rebooted, it can let its client know about it and can exit gracefully. Following this, all OTV clients can start using alternate Server's replication list without waiting for primary Adjacency Server to time out.

For more information around Adjacency Server configuration, please refer to "OTV Configuration" section on page 1-44.

Data Plane: Unicast Traffic

Once the control plane adjacencies between the OTV edge devices are established and MAC address reachability information is exchanged, traffic can start flowing across the overlay. Focusing initially on unicast traffic, it is worthwhile to distinguish between intra-site and inter-site Layer 2 communication (Figure 1-10).

Figure 1-10 Intra Site Layer 2 Unicast Traffic

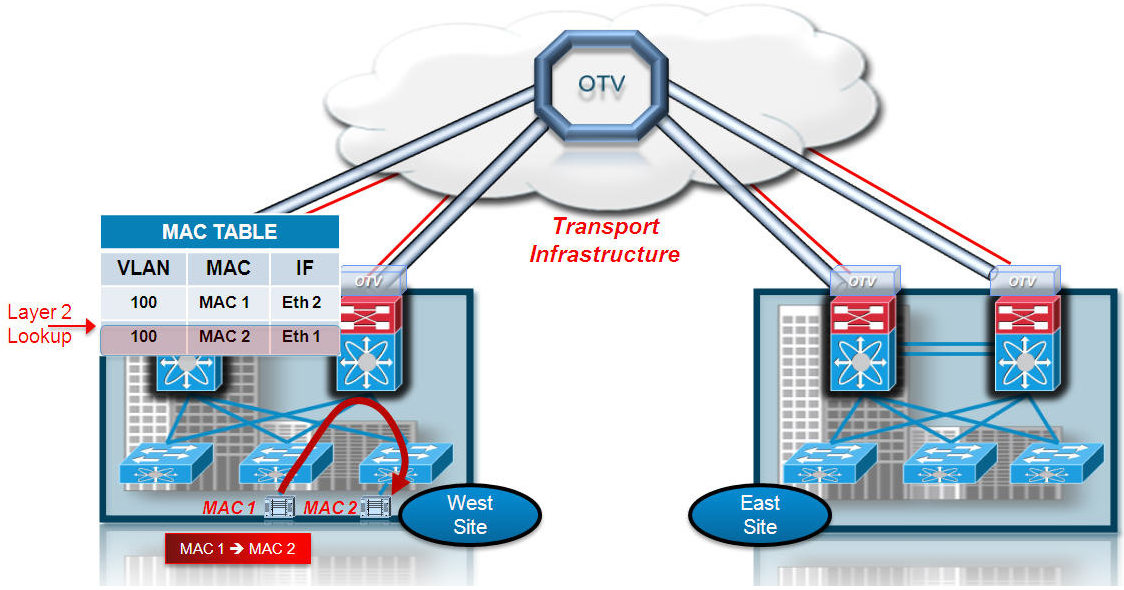
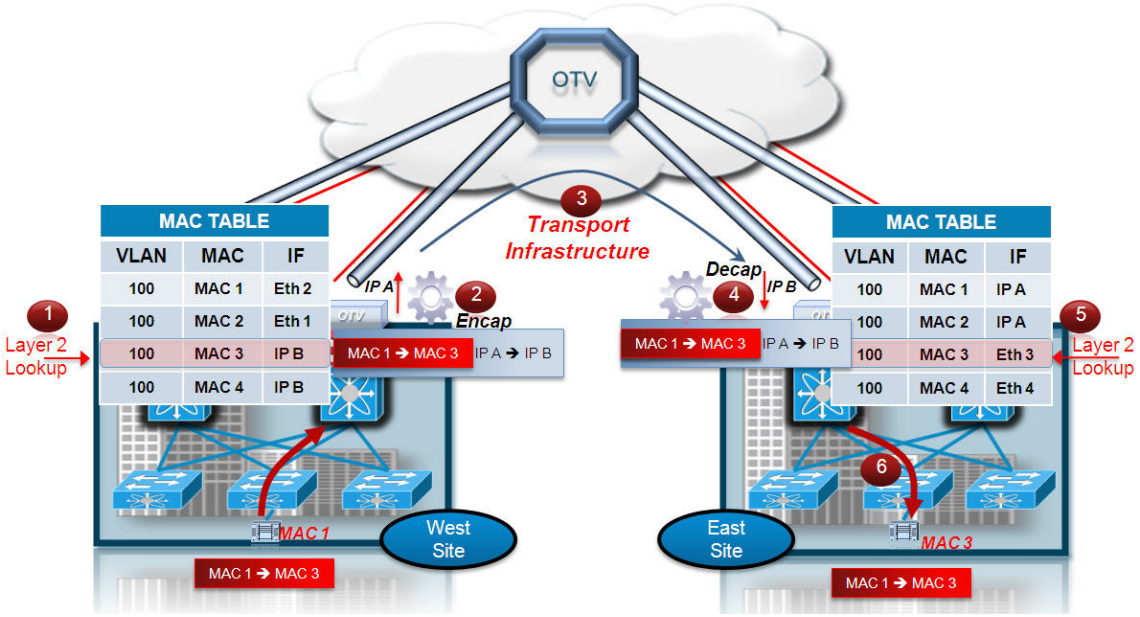


Figure 1-10 depicts intra-site unicast communication: MAC 1 (Server 1) needs to communicate with MAC 2 (Server 2), both belonging to the same VLAN. When the frame is received at the aggregation layer device (which in this case is also deployed as the OTV edge device), the usual Layer 2 lookup is performed to determine how to reach the MAC 2 destination. Information in the MAC table points out a local interface (Eth 1), so the frame is delivered by performing classical Ethernet local switching. A different mechanism is required to establish Layer 2 communication between remote sites (Figure 1-11).

Figure 1-11 Inter Site Layer 2 Unicast Traffic



The following procedure details that relationship:

Step 1 The Layer 2 frame is received at the aggregation layer, or OTV edge device. A traditional Layer 2 lookup is performed, but this time the MAC 3 information in the MAC table does not point to a local Ethernet interface but to the IP address of the remote OTV edge device that advertised the MAC reachability information.

Step 2 The OTV edge device encapsulates the original Layer 2 frame: the source IP of the outer header is the IP address of its Join interface, whereas the destination IP is the IP address of the Join interface of the remote edge device.

Step 3 The OTV encapsulated frame (a regular unicast IP packet) is carried across the transport infrastructure and delivered to the remote OTV edge device.

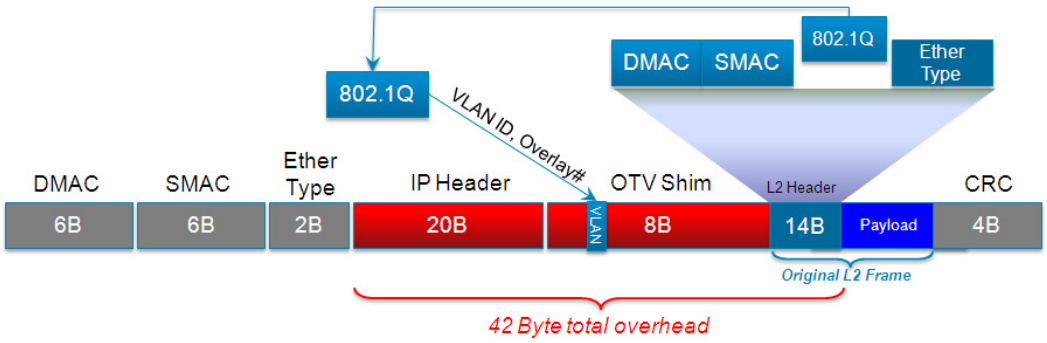
Step 4 The remote OTV edge device decapsulates the frame exposing the original Layer 2 packet.

Step 5 The edge device performs another Layer 2 lookup on the original Ethernet frame and discovers that it is reachable through a physical interface, which means it is a MAC address local to the site.

Step 6 The frame is delivered to the MAC 3 destination.

Given that Ethernet frames are carried across the transport infrastructure after being OTV encapsulated, some considerations around MTU are necessary. Figure 1-12 highlights OTV Data Plane encapsulation performed on the original Ethernet frame.

Figure 1-12 OTV Data Plane Encapsulation



In the first implementation, the OTV encapsulation increases the overall MTU size of 42 bytes. This is the result of the operation of the Edge Device that removes the CRC and the 802.1Q fields from the original Layer 2 frame and adds an OTV Shim (containing also the VLAN and Overlay ID information) and an external IP header.

Also, all OTV control and data plane packets originate from an OTV Edge Device with the “Don't Fragment” (DF) bit set. In a Layer 2 domain the assumption is that all intermediate LAN segments support at least the configured interface MTU size of the host. This means that mechanisms like Path MTU Discovery (PMTUD) are not an option in this case. Also, fragmentation and reassembly capabilities are not available on Nexus 7000 platforms. Consequently, increasing the MTU size of all the physical interfaces along the path between the source and destination endpoints to account for introducing the extra 42 bytes by OTV is recommended.

Note This is not an OTV specific consideration, since the same challenge applies to other Layer 2 VPN technologies, like EoMPLS or VPLS.

Data Plane: Multicast Traffic

In certain scenarios there may be the requirement to establish Layer 2 multicast communication between remote sites. This is the case when a multicast source sending traffic to a specific group is deployed in a given VLAN A in site 1, whereas multicast receivers belonging to the same VLAN A are placed in remote sites 2 and 3 and need to receive traffic for that same group.

Similarly to what is done for the OTV control plane, we need to distinguish the two scenarios where the transport infrastructure is multicast enabled, or not, for the data plane.

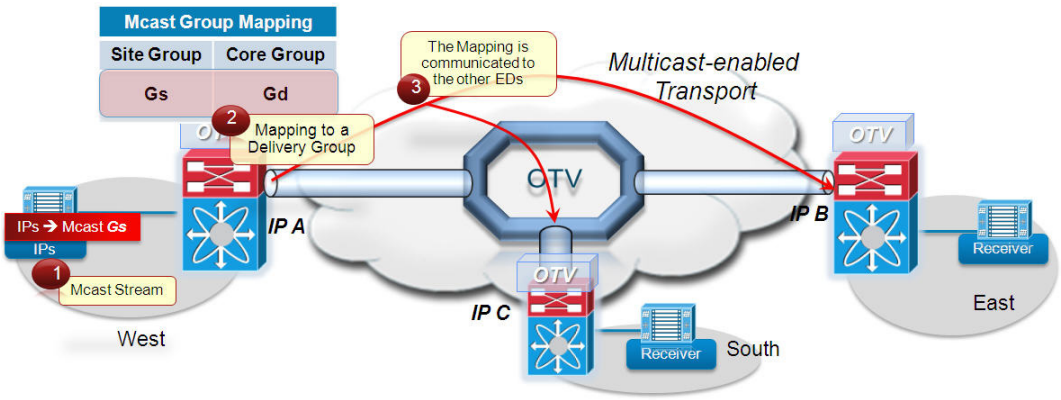
Multicast Enabled Transport Infrastructure

The Layer 2 multicast traffic must flow across the OTV overlay, and to avoid suboptimal head-end replication, a specific mechanism is required to ensure that multicast capabilities of the transport infrastructure can be leveraged.

The idea is to use a set of Source Specific Multicast (SSM) groups in the transport to carry these Layer 2 multicast streams. These groups are independent from the ASM group previously introduced to transport the OTV control protocol between sites.

Figure 1-13 shows the steps occurring once a multicast source is activated in a given data center site.

Figure 1-13 Multicast Source Streaming to Group Gs



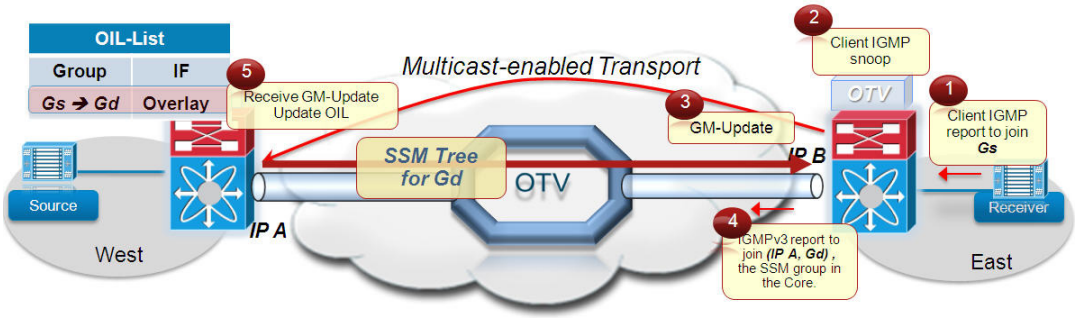
Step 1 A multicast source is activated on the West side and starts streaming traffic to the group Gs.

Step 2 The local OTV edge device receives the first multicast frame and creates a mapping between the group Gs and a specific SSM group Gd available in the transport infrastructure. The range of SSM groups to be used to carry Layer 2 multicast data streams are specified during the configuration of the Overlay interface. Refer to OTV Configuration, page 1-44 for details.

Step 3 The OTV control protocol is used to communicate the Gs-to-Gd mapping to all remote OTV edge devices. The mapping information specifies the VLAN (VLAN A) to which the multicast source belongs and the IP address of the OTV edge device that created the mapping.

Figure 1-14 shows steps that occur once a receiver, deployed in the same VLAN A of the multicast source, decides to join the multicast stream Gs.

Figure 1-14 Receiver Joining the Multicast Group Gs



Step 1 The client sends an IGMP report inside the East site to join the Gs group.

Step 2 The OTV edge device snoops the IGMP message and realizes there is an active receiver in the site interested in group Gs, belonging to VLAN A.

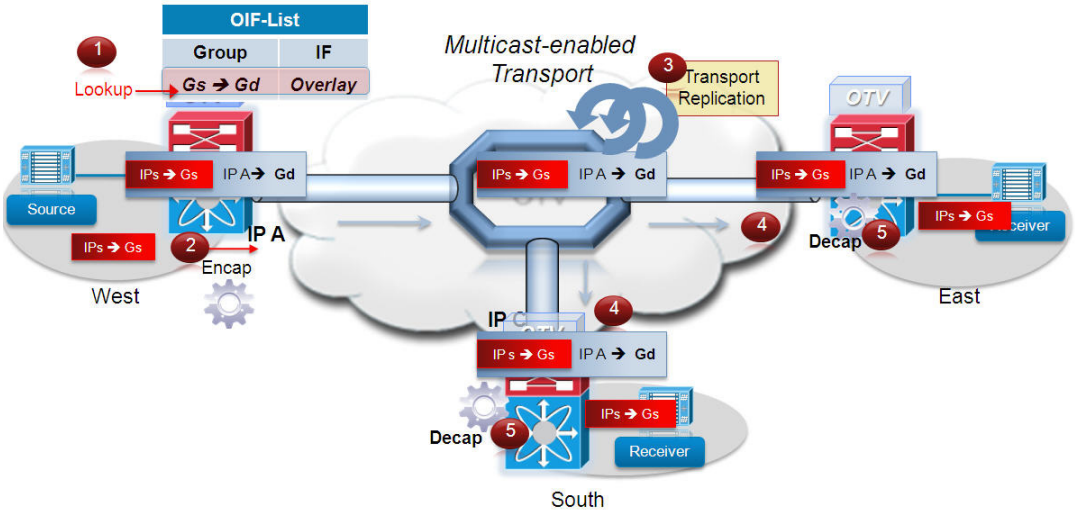
Step 3 The OTV Device sends an OTV control protocol message to all the remote edge devices to communicate this information.

Step 4 The remote edge device in the West side receives the GM-Update and updates its Outgoing Interface List (OIL) with the information that group Gs needs to be delivered across the OTV overlay.

Step 5 Finally, the edge device in the East side finds the mapping information previously received from the OTV edge device in the West side identified by the IP address IP A. The East edge device, in turn, sends an IGMPv3 report to the transport to join the (IP A, Gd) SSM group. This allows building an SSM tree (group Gd) across the transport infrastructure that can be used to deliver the multicast stream Gs.

Figure 1-15 finally shows how multicast traffic is actually delivered across the OTV overlay.

Figure 1-15 Delivery of the Multicast Stream Gs



Step 1 The OTV edge device receives the stream Gs (sourced by IPs) and determines by looking at the OIL that there are receivers interested in group Gs that are reachable through the overlay.

Step 2 The edge device encapsulates the original multicast frame. The source in the outer IP header is the IP A address identifying itself, whereas the destination is the Gd SSM group dedicated to the delivery of multicast data.

Step 3 The multicast stream Gd flows across the SSM tree previously built across the transport infrastructure and reaches all the remote sites with receivers interested in getting the Gs stream.

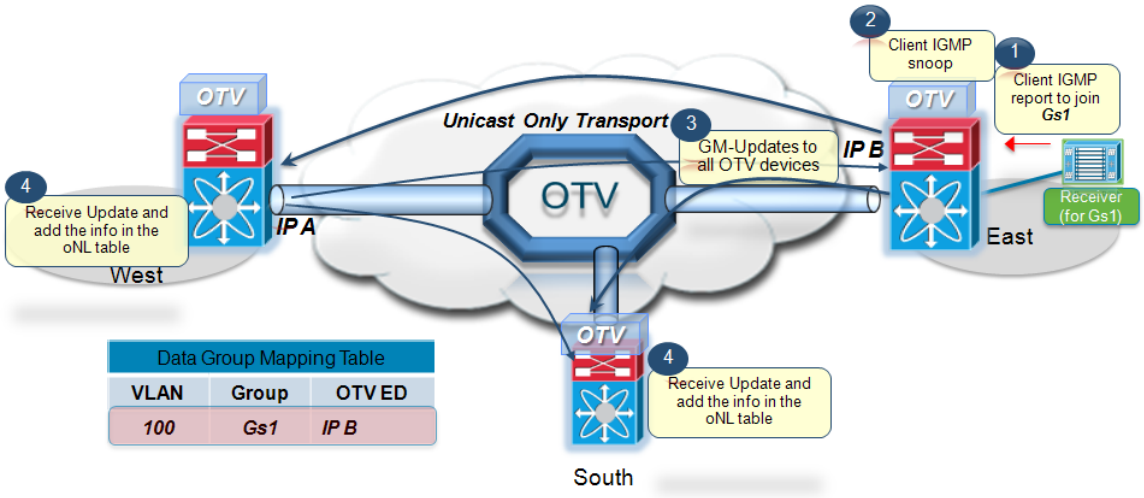
Step 4 The remote OTV edge devices receive the packets.

Step 5 The packets are decapsulated and delivered to the interested receivers belonging to each given site.

Unicast-Only Transport Infrastructure (Adjacency-Server Mode)

When multicast capabilities are not available in the transport infrastructure, Layer 2 multicast traffic can be sent across the OTV overlay by leveraging head-end replication from the OTV device deployed in the DC site where the multicast source is located. However, similarly to what discussed above for the multicast transport scenario, a specific mechanism based on IGMP Snooping is still available to ensure Layer 2 multicast packets are sent only to remote DC sites where active receivers interested in that flow are connected. This behavior allows reducing the amount of required head-end replication, and it is highlighted in Figure 1-16.

Figure 1-16 Receiver Joining the Multicast Group Gs (Unicast Core)



Step 1 The client sends an IGMP report inside the East site to join the Gs group.

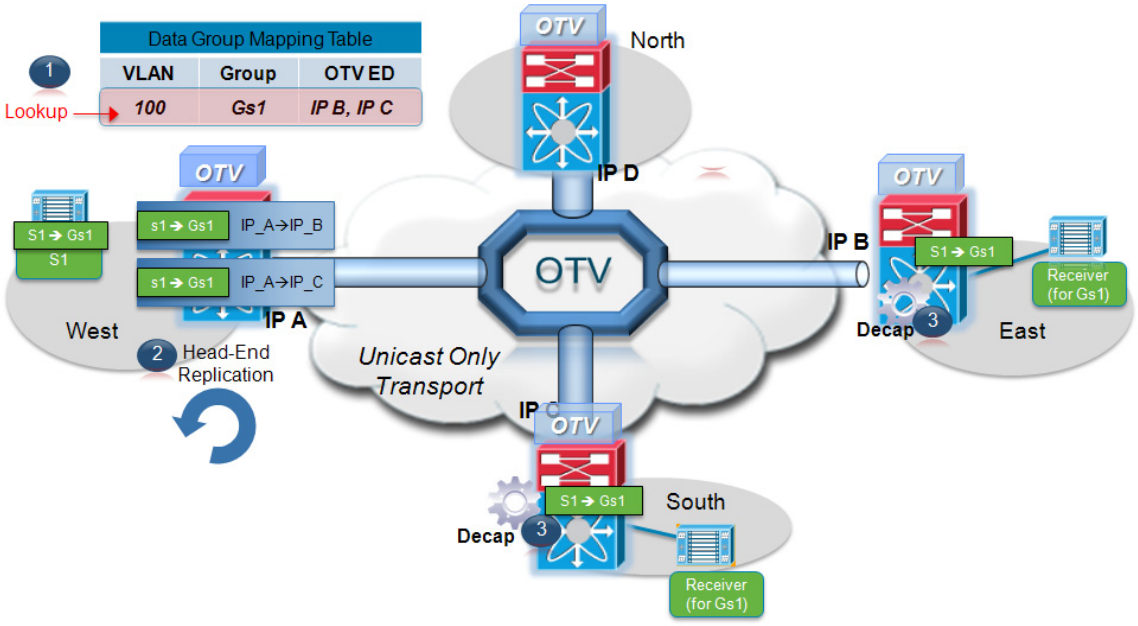
Step 2 The OTV edge device snoops the IGMP message and realizes there is an active receiver in the site interested in group Gs, belonging to VLAN 100.

Step 3 The OTV Device sends an OTV control protocol message (GM-Update) to each remote edge devices (belonging to the unicast list) to communicate this information.

Step 4 The remote edge devices receive the GM-Update message and update their "Data Group Mapping Table" with the information that a receiver interested in multicast group Gs1 is now connected to the site reachable via the OTV device identified by the IP B address.

Figure 1-17 highlights how Layer 2 multicast traffic is actually delivered across the OTV overlay.

Figure 1-17 Delivery of the Layer 2 Multicast Stream Gs (Unicast Core)



Step 1 The multicast traffic destined to Gs1 and generated by a source deployed in the West site reaches the left OTV Edge Device. An OIF lookup takes place in the "Data Group Mapping Table". The table shows that there are receivers across the Overlay (in this example connected to the East and South sites).

Step 2 The left edge device performs head-end replication and create two unicast IP packets (by encapsulating the original Layer 2 multicast frame). The source in the outer IP header is the IP A address identifying itself, whereas the destinations are the IP addresses identifying the Join interfaces of the OTV devices in East and South sites (IP B and IP C).

Step 3 The unicast frames are routed across the transport infrastructure and properly delivered to the remote OTV devices, which decapsulate the frames and deliver them to the interested receivers belonging to the site. Notice how the Layer 2 multicast traffic delivery is optimized, since no traffic is sent to the North site (since no interested receivers are connected there).

Step 4 The remote OTV edge devices receive the packets.

Data Plane: Broadcast Traffic

Finally, it is important to highlight that a mechanism is required so that Layer 2 broadcast traffic can be delivered between sites across the OTV overlay. Failure Isolation, page 1-17 details how to limit the amount of broadcast traffic across the transport infrastructure, but some protocols, like Address Resolution Protocol (ARP), would always mandate the delivery of broadcast packets.

In the current OTV software release, When a multicast enabled transport infrastructure is available, the current NX-OS software release broadcast frames are sent to all remote OTV edge devices by leveraging the same ASM multicast group in the transport already used for the OTV control protocol. Layer 2 broadcast traffic will then be handled exactly the same way as the OTV Hello messages shown in Figure 1-5.

For unicast-only transport infrastructure deployments, head-end replication performed on the OTV device in the site originating the broadcast would ensure traffic delivery to all the remote OTV edge devices part of the unicast-only list.

Failure Isolation

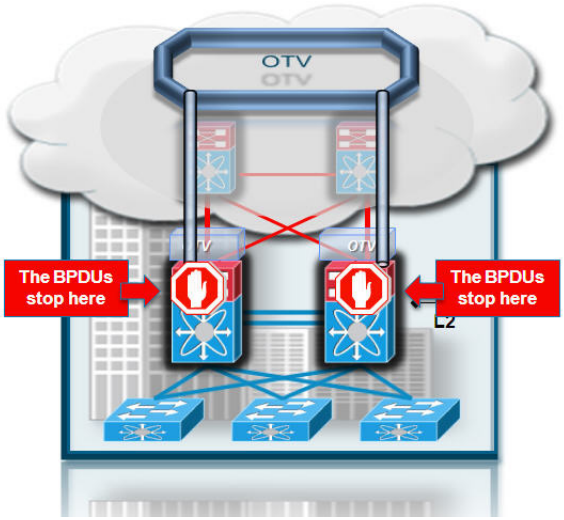
One of the main requirements of every LAN extension solution is to provide Layer 2 connectivity between remote sites without giving up the advantages of resiliency, stability, scalability, and so on, obtained by interconnecting sites through a routed transport infrastructure.

OTV achieves this goal by providing four main functions: Spanning Tree (STP) isolation, Unknown Unicast traffic suppression, ARP optimization, and broadcast policy control.

STP Isolation

Figure 1-18 shows how OTV, by default, does not transmit STP Bridge Protocol Data Units (BPDUs) across the overlay. This is a native function that does not require the use of an explicit configuration, such as BPDU filtering, and so on. This allows every site to become an independent STP domain: STP root configuration, parameters, and the STP protocol flavor can be decided on a per-site basis.

Figure 1-18 OTV Spanning Tree Isolation



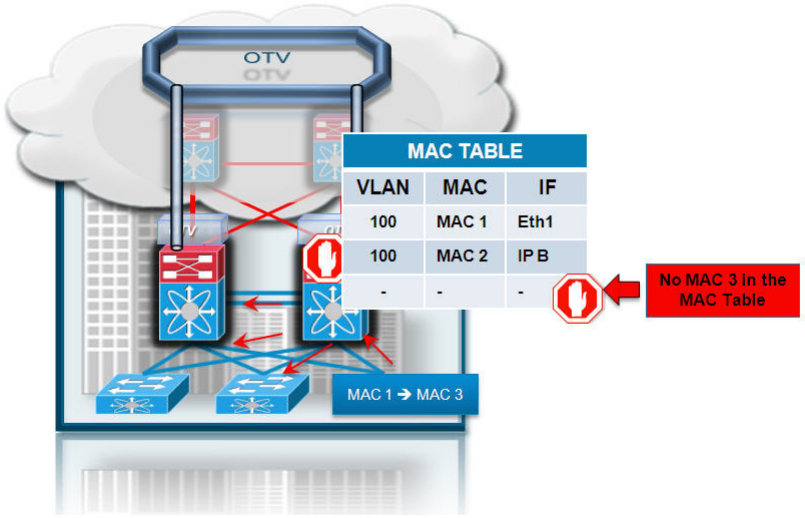
This fundamentally limits the fate sharing between data center sites: a STP problem in the control plane of a given site would not produce any effect on the remote data centers.

Limiting the extension of STP across the transport infrastructure potentially creates undetected end-to-end loops that would occur when at least two OTV edge devices are deployed in each site, inviting a common best practice to increase resiliency of the overall solution. Multi-Homing, page 1-21 details how OTV prevents the creation of end-to-end loops without sending STP frames across the OTV overlay.

Unknown Unicast Handling

The introduction of an OTV control protocol allows advertising MAC address reachability information between the OTV edge devices and mapping MAC address destinations to IP next hops that are reachable through the network transport. The consequence is that the OTV edge device starts behaving like a router instead of a Layer 2 bridge, since it forwards Layer 2 traffic across the overlay if it has previously received information on how to reach that remote MAC destination. Figure 1-19 shows this behavior.

Figure 1-19 OTV Unknown Unicast Handling



When the OTV edge device receives a frame destined to MAC 3, it performs the usual Layer 2 lookup in the MAC table. Since it does not have information for MAC 3, Layer 2 traffic is flooded out the internal interfaces, since they behave as regular Ethernet interfaces, but not via the overlay.

Note This behavior of OTV is important to minimize the effects of a server misbehaving and generating streams directed to random MAC addresses. This could occur as a result of a DoS attack as well.

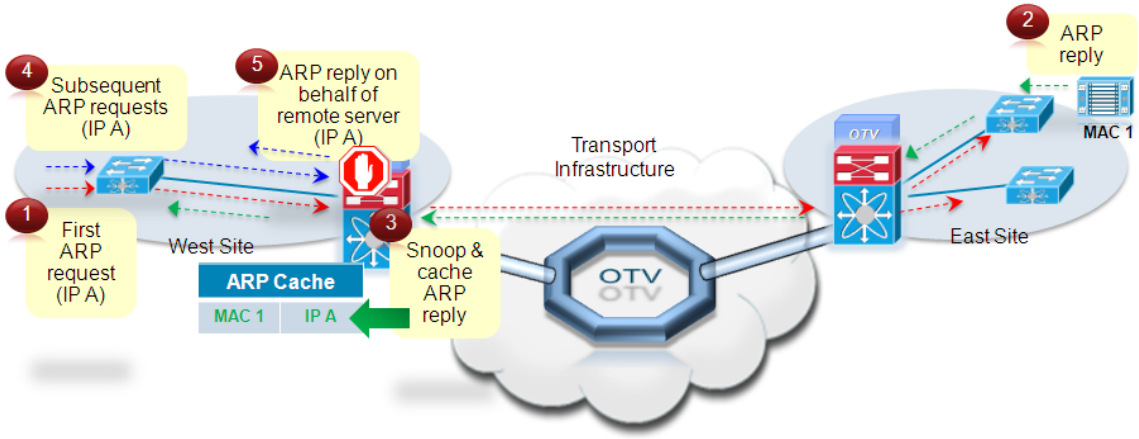
The assumption is that there are no silent or unidirectional devices in the network, so sooner or later the local OTV edge device will learn an address and communicate it to the remaining edge devices through the OTV protocol. To support specific applications, like Microsoft Network Load Balancing Services (NLBS) which require the flooding of Layer 2 traffic to function, a configuration knob is provided to enable selective flooding. Individual MAC addresses can be statically defined so that Layer 2 traffic destined to them can be flooded across the overlay, or broadcast to all remote OTV edge devices, instead of being dropped. The expectation is that this configuration would be required in very specific corner cases, so that the default behavior of dropping unknown unicast would be the usual operation model.

Note Behavior with the current NX-OS release
The configuration knob that allows for selective unicast flooding is not available in the current OTV software release. Consequently, all unknown unicast frames will not be forwarded across the logical overlay.

ARP Optimization

Another function that reduces the amount of traffic sent across the transport infrastructure is ARP optimization. Figure 1-20 depicts the OTV ARP optimization process:

Figure 1-20 OTV ARP Optimization



Step 1 A device in the West site sources an ARP request to determine the MAC of the host with address IP A.

Step 2 The ARP request is a Layer 2 broadcast frame and it is sent across the OTV overlay to all remote sites eventually reaching the machine with address IP A, which creates an ARP reply message. The ARP reply is sent back to the originating host in the West data center.

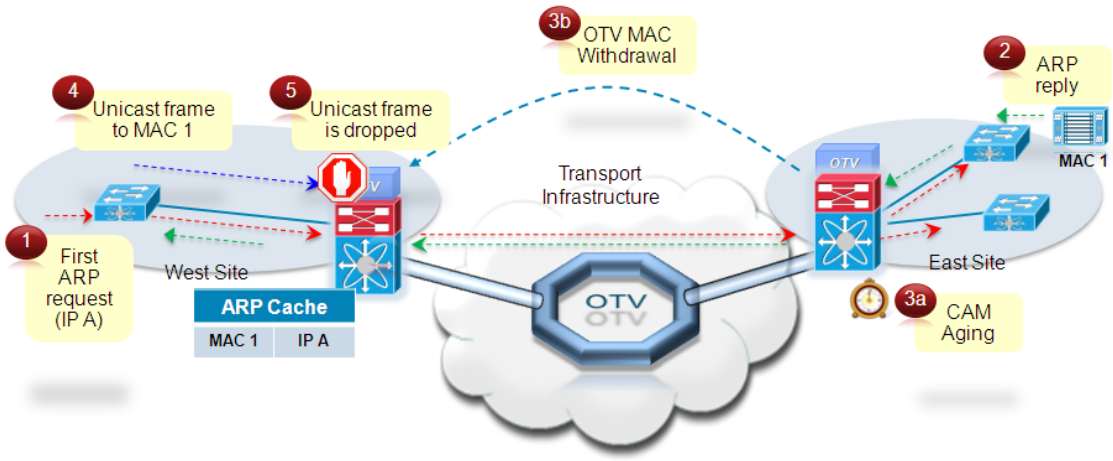
Step 3 The OTV edge device in the original West site is capable of snooping the ARP reply and caches the contained mapping information (MAC 1, IP A) in a local data structure named ARP Neighbor-Discovery (ND) Cache.

Step 4 A subsequent ARP request is originated from the West site for the same IP A address.

Step 5 The request is not forwarded to the remote sites but is locally answered by the local OTV edge device on behalf of the remote device IP A.

Because of this ARP caching behavior, you should consider the interactions between ARP and CAM table aging timers, since incorrect settings may lead to black-holing traffic. Figure 1-21 shows the ARP aging timer is longer than the CAM table aging timer. This is also a consequence of the OTV characteristic of dropping unknown unicast frames.

Figure 1-21 Traffic Black-Holing Scenario



The following steps explain the traffic black-holing scenario.

Step 1 A device in the West site sources an ARP request to determine the MAC of the host with address IP A.

Step 2 The ARP request is a Layer 2 broadcast frame and it is sent across the OTV overlay to all remote sites eventually reaching the machine with address IP A, which creates an ARP reply message. The ARP reply is sent back to the originating host in the West data center. This information is snooped and added to the ARP cache table in the OTV edge device in the West side.

Step 3 MAC 1 host stops communicating, hence the CAM aging timer for the MAC 2 entry expires on the East OTV edge device. This triggers an OTV Update sent to the edge device in the West site, so that it can remove the MAC 1 entry as well. This does not affect the entry in the ARP cache, since we assume the ARP aging timer is longer than the CAM.

Step 4 A Host in the West Site sends a unicast frame directed to the host MAC 1.

Step 5 The unicast frame is received by the West OTV edge device, which has a valid ARP entry in the cache for that destination. However, the lookup for MAC 1 in the CAM table does not produce a hit, resulting in a dropped frame.

The ARP aging timer on the OTV edge devices should always be set lower than the CAM table aging timer. The defaults on Nexus 7000 platforms for these timers are shown below:

- OTV ARP aging-timer: 480 seconds / 8 minutes
- MAC aging-timer: 1800 seconds / 30 minutes

Note It is worth noting how the ARP aging timer on the most commonly found operating systems (Win2k, XP, 2003, Vista, 2008, 7, Solaris, Linux, MacOSX) is actually lower than the 30 minutes default MAC aging-timer. This implies that the behavior shown in Figure 1-21 would never occur in a real deployment scenario, because the host would re-ARP before aging out an entry which would trigger an update of the CAM table, hence maintaining the OTV route.

In deployments where the hosts default gateway is placed on a device different than the Nexus 7000 it is important to set the ARP aging-timer of the device to a value lower than its MAC aging-timer.

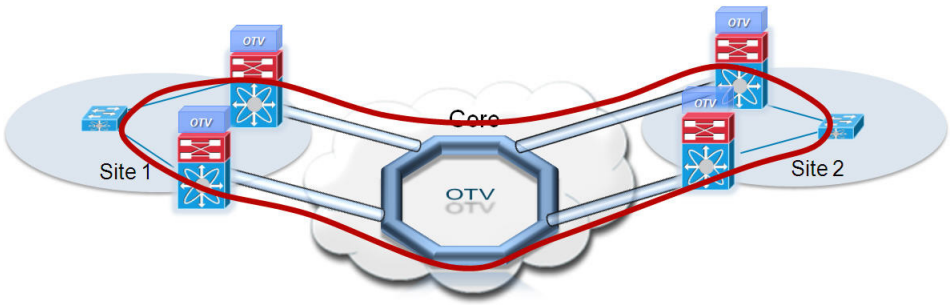
Broadcast Policy Control

In addition to the previously described ARP optimization, OTV will provide additional functionality such as broadcast suppression, broadcast white-listing, and so on, to reduce the amount of overall Layer 2 broadcast traffic sent across the overlay. Details will be provided upon future functional availability.

Multi-Homing

One key function built in the OTV protocol is multi-homing where two (or more) OTV edge devices provide LAN extension services to a given site. As mentioned, this redundant node deployment, combined with the fact that STP BPDUs are not sent across the OTV overlay, may lead to the creation of an end-to-end loop, Figure 1-22.

Figure 1-22 Creation of an End-to-End STP Loop

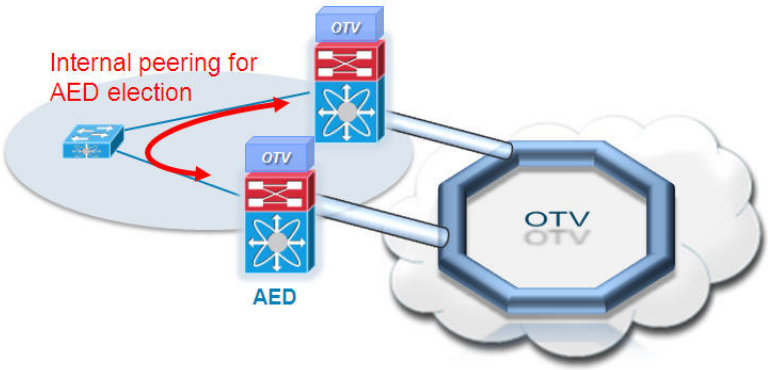


The concept of Authoritative edge device (AED) is introduced to avoid the situation depicted in Figure 1-22. The AED has two main tasks:

1. Forwarding Layer 2 traffic (unicast, multicast and broadcast) between the site and the overlay (and vice versa).
2. Advertising MAC reachability information to the remote edge devices.

The AED role is negotiated, on a per-VLAN basis, between all the OTV edge devices belonging to the same site (that is, characterized by the same Site ID). Prior to NX-OS release 5.2(1), OTV used a VLAN called "Site VLAN" within a site to detect and establish adjacencies with other OTV edge devices as shown in Figure 1-23. OTV used this site adjacencies as an input to determine Authoritative Edge devices for the VLANS being extended from the site.

Figure 1-23 Establishment of Internal Peering



The Site VLAN should be carried on multiple Layer 2 paths internal to a given site, to increase the resiliency of this internal adjacency (including vPC connections eventually established with other edge switches). However, the mechanism of electing Authoritative Edge device (AED) solely based on the communication established on the site VLAN may create situations (resulting from connectivity issues or misconfiguration), where OTV edge devices belonging to the same site can fail to detect one another and thereby ending up in an "active/active" mode (for the same data VLAN). This could ultimately result in the creation of a loop scenario.

To address this concern, starting with 5.2 (1) NX-OS release, each OTV device maintains dual adjacencies with other OTV edge devices belonging to the same DC site. OTV edge devices continue to use the site VLAN for discovering and establishing adjacency with other OTV edge device in a site. This adjacency is called Site Adjacency.

In addition to the Site Adjacency, OTV devices also maintain a second adjacency, named "Overlay Adjacency", established via the Join interfaces across the Layer 3 network domain. In order to enable this new functionality, it is now mandatory to configure each OTV device also with a site-identifier value. All edge devices that are in the same site must be configured with the same site-identifier. This site-identifier is advertised in IS-IS hello packets sent over both the overlay as well as on the site VLAN. The combination of the site-identifier and the IS-IS system-id is used to identify a neighbor edge device in the same site.

Note The Overlay interface on an OTV edge device is forced in a "down" state until a site-identifier is configured. This must be kept into consideration when performing an ISSU upgrade to 5.2(1) from a pre-5.2(1) NX-OS software release, because that would result in OTV not being functional anymore once the upgrade is completed.

The dual site adjacency state (and not simply the Site Adjacency established on the site VLAN) is now used to determine the Authoritative Edge Device role for each extended data VLAN. Each OTV edge device can now proactively inform their neighbors in a local site about their capability to become Authoritative Edge Device (AED) and its forwarding readiness. In other words, if something happens on an OTV device that prevents it from performing its LAN extension functionalities, it can now inform its neighbor about this and let itself excluded from the AED election process.

An explicit AED capability notification allows the neighbor edge devices to get a fast and reliable indication of failures and to determine AED status accordingly in the consequent AED election, rather than solely depending on the adjacency creation and teardown. The forwarding readiness may change due to local failures such as the site VLAN or the extended VLANs going down or the join-interface going down, or it may be intentional such as when the edge device is starting up and/or initializing.

Hence, the OTV adjacencies may be up but OTV device may not be ready to forward traffic. The edge device also triggers a local AED election when its forwarding readiness changes. As a result of its AED capability going down, it will no longer be AED for its VLANs.

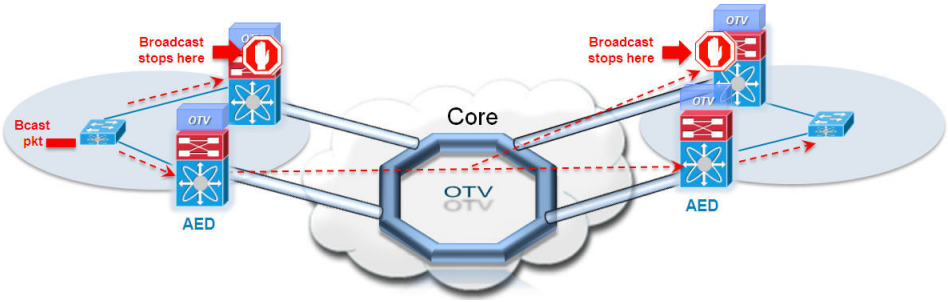
The AED capability change received from a neighboring edge device in the same site influences the AED assignment, and hence will trigger an AED election. If a neighbor indicates that it is not AED capable, it will not be considered as active in the site. An explicit AED capability down notification received over either the site or the overlay adjacency will bring down the neighbor's dual site adjacency state into inactive state and the resulting AED election will not assign any VLANs to that neighbor.

As mentioned above, the single site adjacency (pre 5.2(1) releases) or dual site adjacencies (from 5.2(1) release) are used to negotiate the Authoritative Edge Device role. A deterministic algorithm is implemented to split the AED role for Odd and Evan VLANs between two OTV Edge Devices. More specifically, the Edge Device identified by a lower System-ID will become Authoritative for all the even extended VLANs, whereas the device with higher System-ID will "own" the odd extended VLANs. This behavior is hardware enforced and cannot be tuned in the current NX-OS release.

Note The specific OTV edge device System-ID can be visualized using the **show otv site** CLI command.

Figure 1-24 shows how the definition of the AED role in each site prevents the creation of end-to-end STP loops.

Figure 1-24 Prevention of End-to-End STP Loops



Assume, for example, a Layer 2 broadcast frame is generated in the left data center. The frame is received by both OTV edge devices however, only the AED is allowed to encapsulate the frame and send it over the OTV overlay. All OTV edge devices in remote sites will also receive the frame via the overlay, since broadcast traffic is delivered via the multicast group used for the OTV control protocol, but only the AED is allowed to decapsulate the frame and send it into the site.

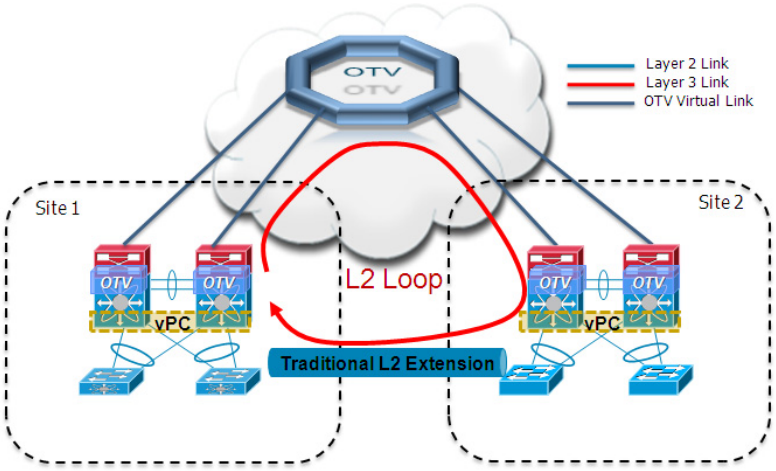
If an AED re-election was required in a specific failure condition or misconfiguration scenario, the following sequence of events would be triggered to re-establish traffic flows both for inbound and outbound directions:

1. The OTV control protocol hold timer expires or an OTV edge device receives an explicit AED capability notification from the current AED.
2. One of the two edge devices becomes an AED for all extended VLANs.
3. The newly elected AED imports in its CAM table all the remote MAC reachability information. This information is always known by the non-AED device, since it receives it from the MAC advertisement messages originated from the remote OTV devices. However, the MAC reachability information for each given VLAN is imported into the CAM table only if the OTV edge device has the AED role for that VLAN.

4. The newly elected AED starts learning the MAC addresses of the locally connected network entities and communicates this information to the remote OTV devices by leveraging the OTV control protocol.

The same loop avoidance mechanism discussed above would be used in site merging scenarios, where a back-door Layer 2 connection exists between data centers. This could happen for example when migrating the LAN extension solution from a traditional one to OTV. Figure 1-25 highlights the fact that during the migration phase it may be possible to create an end-to-end loop between sites. This is a consequence of the design recommendation of preventing STP BPDUs from being sent across the DCI connection.

Figure 1-25 Creation of an End-to-End Loop



To avoid the creation of the end-to-end loop, depicted in Figure 1-25, and minimize the outage for Layer 2 traffic during the migration phase from a traditional DCI solution to OTV, the following step-by-step procedure should be followed (refer to the "OTV Configuration" section for more configuration details):

- Ensure that the same site VLAN is globally defined on the OTV devices deployed in the two data center sites.
- Make sure that the site VLAN is added to the set of VLANs carried via the traditional LAN extension solution. This is critical, because it will allow OTV to detect the already existent Layer 2 connection. For this to happen, it is also important to ensure that the OTV edge devices are adjacent to each other on the site VLAN (i.e. a Layer 2 path exists between these devices through the existing DCI connection).
- From 5.2(1) release, makes sure also that the same site-identifier is configured for all the OTV devices belonging to the same site.
- Create the Overlay configuration on the first edge device in site 1, but do not extend any VLAN for the moment. Do not worry about enabling OTV on the second edge device belonging to the same site yet.
- Create the Overlay configuration on the first edge device in site 2, but do not extend any VLAN for the moment. Do not worry about enabling OTV on the second edge device belonging to the same site yet.
- Make sure the OTV edge devices in site 1 and site 2 establish an internal adjacency on the site VLAN (use the “show otv site” CLI command for that). Assuming the internal adjacency is established, OTV will consider the two sites as merged in a single one. The two edge devices will then negotiate the AED role, splitting between them odd and even VLANs (as previously discussed).

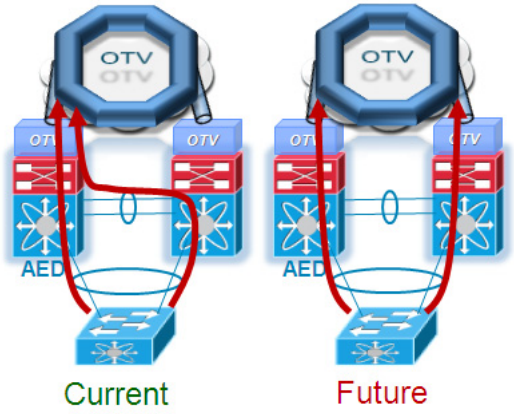
- Configure the VLANs that need to be extended through the OTV Overlay (using the “otv extend-vlan” command) on the OTV edge devices in site 1 and site 2. Notice that even if the two OTV devices are now adjacent also via the overlay (this can be verified with the “show otv adjacency” command) the VLAN extension between sites is still happening only via the traditional Layer 2 connection.
- Disable the traditional Layer 2 extension solution. This will cause the OTV devices to lose the internal adjacency established via the site VLAN and to detect a “site partition” scenario. After a short convergence window where Layer 2 traffic between sites is briefly dropped, VLANs will start being extended only via the OTV overlay. It is worth noticing that at this point each edge device will assume the AED role for all the extended VLANs.
- OTV is now running in single-homed mode (one edge device per site) and it is then possible to improve the resiliency of the solution by enabling OTV on the second edge device existing in each data center site.

Traffic Load Balancing

As mentioned, the election of the AED is paramount to eliminating risk of creating end-to-end loops. The first immediate consequence is that all Layer 2 multicast and broadcast streams need to be handled by the AED device, leading to a per-VLAN load-balancing scheme for these traffic flows.

The exact same considerations are valid for unicast traffic when considering the current NX-OS release of OTV. Only the AED is allowed to forward unicast traffic in and out a given site. However, an improvement behavior is planned for a future release to provide per-flow load-balancing of unicast traffic. Figure 1-26 shows that this behavior can be achieved every time Layer 2 flows are received by the OTV edge devices over a vPC connection.

Figure 1-26 Unicast Traffic Load Balancing

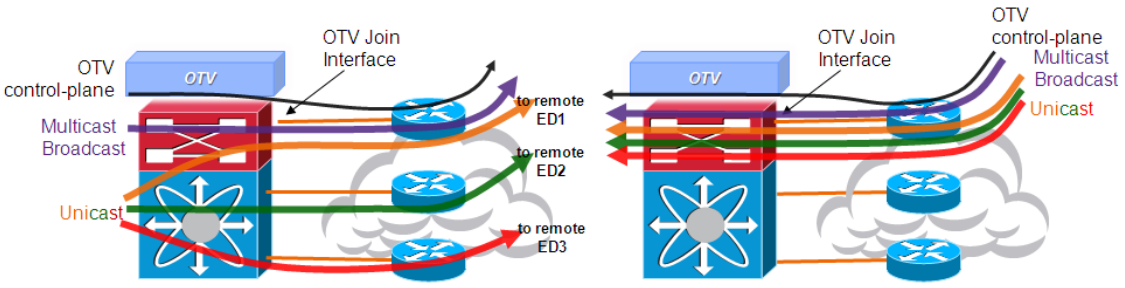


The vPC peer-link is leveraged in the initial release to steer the traffic to the AED device. In future releases where the edge device does not play the AED role, a given VLAN will be allowed to forward unicast traffic to the remote sites via the OTV overlay, providing a desired per-flow load-balancing behavior.

Another consideration for traffic load balancing is on a per-device level: to understand the traffic behavior it is important to clarify that in the current Nexus 7000 hardware implementation, the OTV edge device encapsulates the entire original Layer 2 frame into an IP packet. This means that there is no

Layer 4 (port) information available to compute the hash determining which link to source the traffic from. This consideration is relevant in a scenario where the OTV edge device is connected to the Layer 3 domain by leveraging multiple routed uplinks, as highlighted in Figure 1-27.

Figure 1-27 Inbound and Outbound Traffic Paths



In the example above, it is important to distinguish between egress/ingress direction and unicast/multicast (broadcast) traffic flows.

- Egress unicast traffic: this is destined to the IP address of a remote OTV edge device Join interface. All traffic flows sent by the AED device to a given remote site are characterized by the same source_IP and dest_IP values in the outer IP header of the OTV encapsulated frames. This means that hashing performed by the Nexus 7000 platform acting as the edge device would always select the same physical uplink, even if the remote destination was known with the same metric (equal cost path) via multiple Layer 3 links (notice that the link selected may not necessarily be the Join Interface). However, traffic flows sent to different remote sites will use a distinct dest_IP value, hence it is expected that a large number of traffic flows will be load-balanced across all available Layer 3 links, as shown on the left in Figure 1-27.

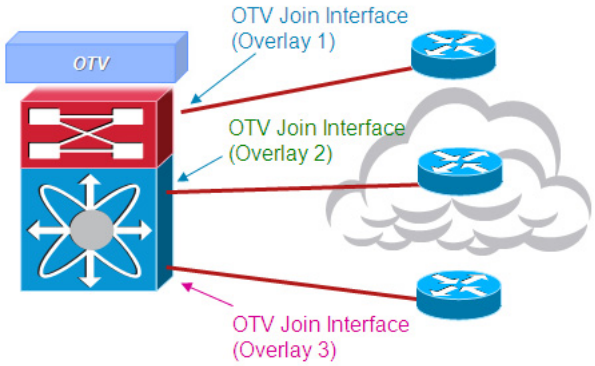
Note The considerations above are valid only in the presence of equal cost routes. If the remote OTV edge device was reachable with a preferred metric out of a specific interface, all unicast traffic directed to that site would use that link, independently from where the Join Interface is defined. Also, the same considerations apply also to deployment scenarios where a Layer 3 Port-channel is defined as OTV Join Interface.

- Egress multicast/broadcast and control plane traffic: independently from the number of equal cost paths available on a given edge device, multicast, broadcast and control plane traffic is always going to be sourced from the defined Join interface.
- Ingress unicast traffic: all the incoming unicast traffic will always be received on the Join interface, since it is destined to this interface IP address, as shown on the right in Figure 1-27.
- Ingress multicast/broadcast and control plane traffic: independently from the number of equal cost paths available on a given edge device, multicast, broadcast and control plane traffic must always be received on the defined Join interface. If in fact control plane messages were delivered to a different interface, they would be dropped and this would prevent OTV from becoming fully functional.

Load balancing behavior will be modified once loopback interfaces (or multiple physical interfaces) will be supported as OTV Join interfaces allowing the load-balance of unicast and multicast traffic across multiple links connecting the OTV edge device to the routed network domain.

In the meantime, a possible workaround to improve the load-balancing of OTV traffic consists in leveraging multiple OTV overlays on the same edge device and spread the extended VLANs between them. This concept is highlighted in Figure 1-28.

Figure 1-28 Use of Multiple TV Overlays



In the scenario above, all traffic (unicast, multicast, broadcast, control plane) belonging to a given Overlay will always be sent and received on the same physical link configured as Join Interface, independent from the remote site it is destined to. This means that if the VLANs that need to be extended are spread across the 3 defined Overlays, an overall 3 way load-balancing would be achieved for all traffic even in point-to-point deployments where all the traffic is sent to a single remote data center site.

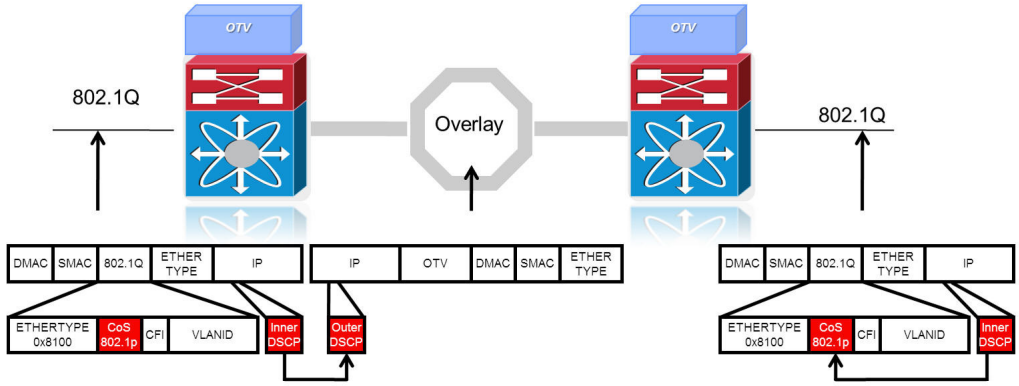
QoS Considerations

To clarify the behavior of OTV from a QoS perspective, we should distinguish between control and data plane traffic.

- Control Plane: The control plane frames are always originated by the OTV edge device and statically marked with a CoS = 6/DSCP = 48.
- Data Plane: The assumption is that the Layer 2 frames received by the edge device to be encapsulated have already been properly marked (from a CoS and DCSP perspective).

Figure 1-29 shows the default behavior of the OTV edge device when encapsulating and decapsulating Layer 2 frames if running pre-5.2(1) NX-OS releases.

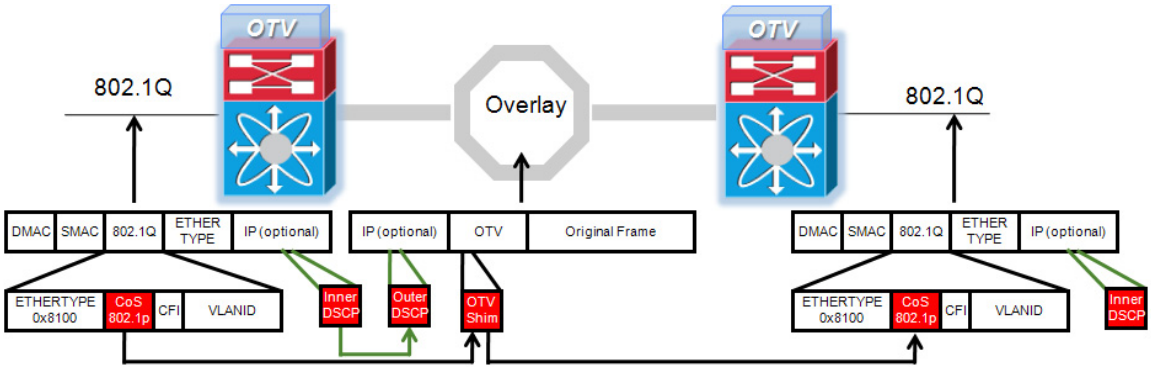
Figure 1-29 Default QoS Behavior of the OTV Edge Device



By default, the original (inner) DSCP value is copied to the outer IP header of the OTV encapsulated frame. This allows QoS decisions to be applied to the encapsulated frame across the transport infrastructure. Once the packet reaches the destination site, it is decapsulated at which time the original DSCP value is used to populate the CoS field of the Layer 2 frame sent into the site.

Release 5.2(1) added functionality to allow for the preservation of the CoS and DSCP values across an OTV overlay, as shown in Figure 1-30.

Figure 1-30 CoS and DSCP preservation in 5.2(1) NX-OS Release



Before the Layer 2 frame is encapsulated, the OTV edge device copies the CoS bits (802.1p) from the original Layer 2 header to the OTV shim header. Also, if the original frame is an IP packet, the original (inner) DSCP value is also copied to the "outer" DSCP. This would allow to apply consistent QoS policies to OTV traffic across the transport infrastructure.

Once the packet is received on the remote OTV device and decapsulated, the CoS value is recovered from the OTV shim and added to the 802.1Q header, allowing for preservation of both original CoS and DSCP values. Note that if the DSCP value in the outer IP header were to be modified in the transport infrastructure, these changes would not be reflected in the inner DSCP value exposed after decapsulation.

In addition to the functionality described above, from the initial OTV release it is possible to give the user the ability to map the CoS of the original packets to the DSCP value of the outer IP header for OTV packets by applying a specific policy map. This is important to uniquely mark and identify the OTV packets and apply QoS policy accordingly.

FHRP Isolation

The last capability introduced by OTV is to filter First Hop Redundancy Protocol (FHRP—HSRP, VRRP, and so on) messages across the logical overlay. This is required to allow for the existence of the same default gateway in different locations and optimize the outbound traffic flows (server to client direction). Figure 1-31 highlights the root of the problem.

Given that the same VLAN/IP subnet is available in different sites, the free exchange of FHRP messages across the OTV connection would lead to the election of a single default gateway. This would force traffic to follow a suboptimal path to reach the default gateway (in the site where it is deployed) each time it is required to be routed outside the subnet and the server is located in a different site.

Figure 1-31 Suboptimal Outbound Routing

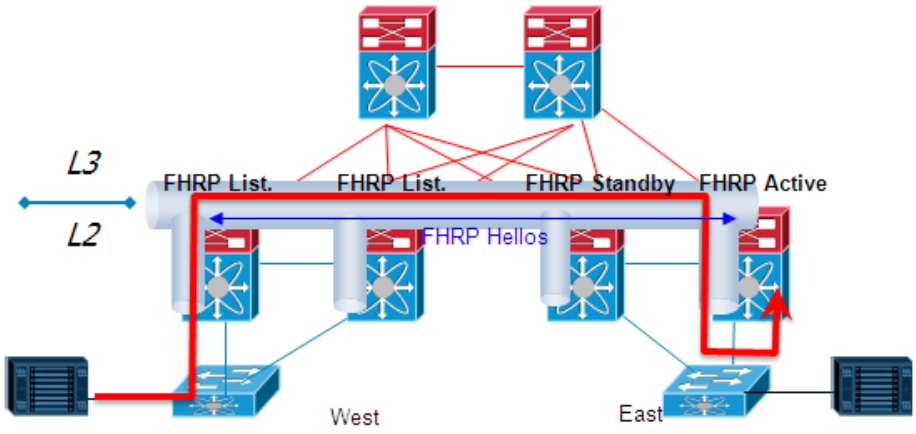
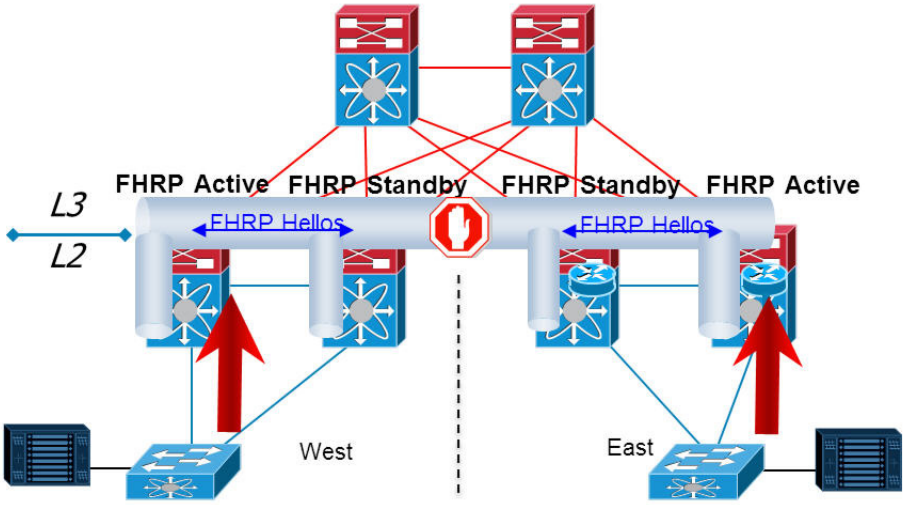


Figure 1-32 shows the deployment of independent default gateways in each data center site, to optimize and localize routing of outbound traffic flows.

Figure 1-32 FHRP Isolation with OTV



It is critical that you enable the filtering of FHRP messages across the overlay because it allows the use of the same FHRP configuration in different sites. The end result is that the same default gateway is available, and characterized by the same virtual IP and virtual MAC addresses, in each data center. This means that the outbound traffic will be able to follow the optimal and shortest path, always leveraging the local default gateway, as it will be discussed in the "OTV Configuration" section on page 1-44.

Note Behavior with the current NX-OS release
OTV can provide a single command to enable the FHRP filtering functionality. However, this is not available in the current OTV software release An alternative configuration (leveraging MAC access-control lists) can be implemented in the interim to achieve the same result.

It is important to stress how this outbound path optimization functionality should be deployed in conjunction with an equivalent one optimizing inbound traffic flows to avoid asymmetric traffic behavior (this would be highly undesirable especially in deployments leveraging stateful services across data centers).

Note Discussing inbound traffic optimization solutions is out of the scope of this document. For more information, refer to the Virtualized Workload Mobility design guide available at: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/4.0/EMC/EMC.pdf

OTV and SVIs Coexistence

The current OTV implementation on the Nexus 7000 enforces the separation between SVI routing and OTV encapsulation for a given VLAN. This separation can be achieved with the traditional workaround of having two separate network devices to perform these two functions.

An alternative, cleaner and less intrusive solution is proposed here by introducing the use of Virtual Device Contexts (VDCs) available with Nexus 7000 platforms. Two VDCs would be deployed: an OTV VDC dedicated to perform the OTV functionality and a Routing VDC used to provide SVI routing support.

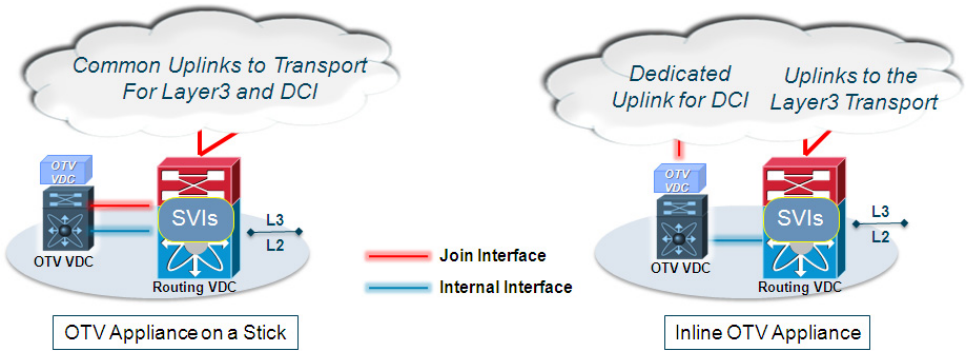
Note More discussion on the VDC requirements for OTV deployment can be found in the "OTV Deployment Options" section on page 1-34.

Two different deployment models are considered for the OTV VDC based on the availability of uplinks to the DCI Transport:

- OTV Appliance on a Stick: where a common set of uplinks from the Routing VDC are used for both the routing and DCI extension
- Inline OTV Appliance: where a dedicated link from the OTV VDC is used for the DCI extension

From an OTV perspective there is no difference between the two models (Figure 1-33).

Figure 1-33 OTV VDC Models

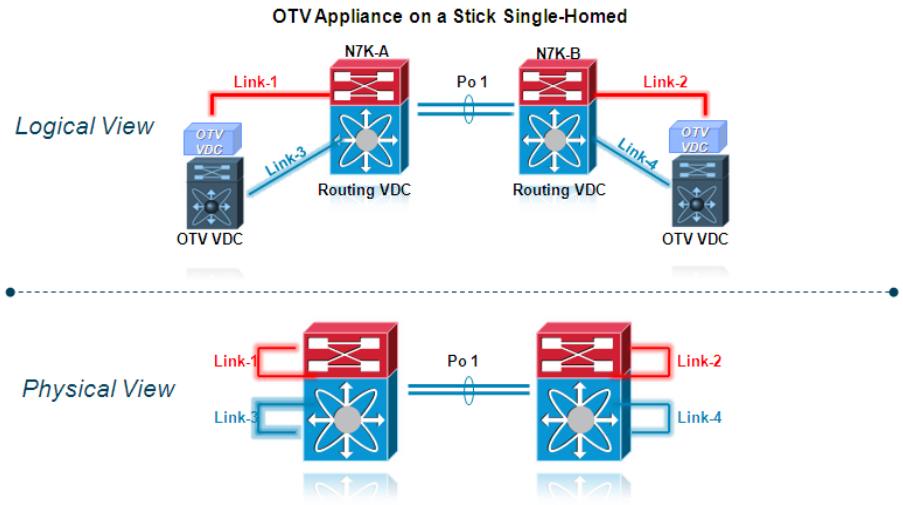


The main advantage of the Appliance on a Stick model is that no changes to the design or to the physical connections would be required once the dependency from the OTV VDC is removed. The only migration steps at that point would be to move the OTV configuration from the OTV VDC to the Routing VDC and deactivate the OTV VDC. This would be transparent to the rest of the data center network.

Dimensioning the OTV VDC

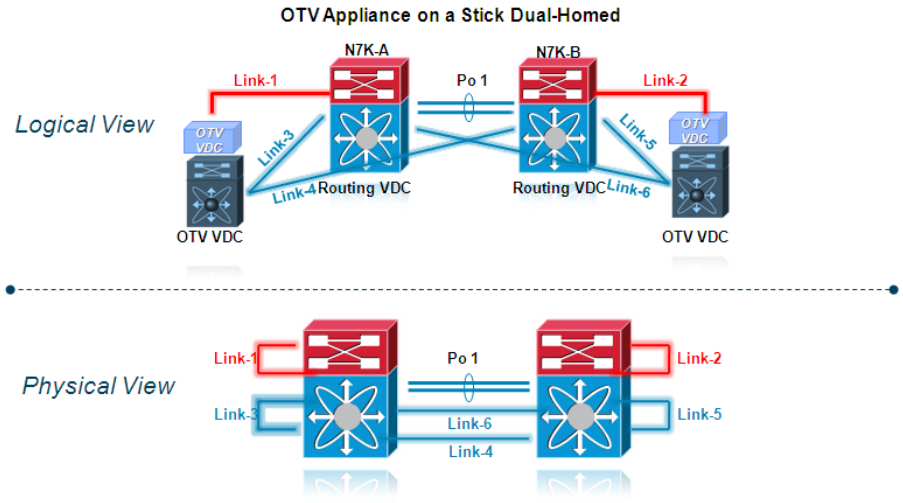
When creating the OTV VDC, the minimum ports to be allocated to the new VDC are two: one join-interface and one internal interface, as shown in Figure 1-34.

Figure 1-34 Single-homed OTV VDC



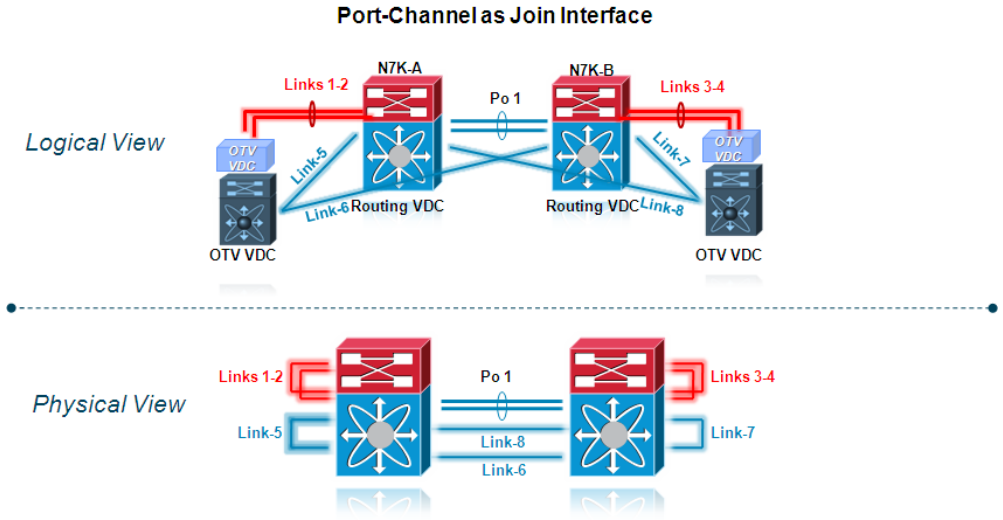
To achieve higher resiliency it is possible to dual-home the OTV VDC to both the aggregation VDCs as shown in Figure 1-35.

Figure 1-35 OTV Appliance on a Stick



Finally, the most resilient deployment removing all possible single point of failures from the design is shown in Figure 1-36.

Figure 1-36 Use of Port Channel as Join Interface



The idea is to leverage a Layer 3 Port-Channel as Join interface for each OTV VDC, by bundling two Layer 3 links connecting each OTV device to its own aggregation layer VDC. As shown in Figure 1-36, deploying redundant links for both the internal and the Join interfaces represent a trade-off between cost (given the use of additional physical ports to interconnect VDCs) and overall resiliency and convergence characteristics of the design (since an AED re-election would now be required to restore traffic flows only in a dual failure scenario). For achieving the most resilient configuration and remove any single point of failure from the design, it is recommended to distribute the physical interfaces bundled together in the Internal or Join interface port-channels across different line-cards (when possible).

OTV Scalability Considerations

From a scalability perspective, there are several elements to be considered. The following values are supported in the current 5.2(1) NX-OS release of OTV will increase in future releases and will be reflected is corresponding documentation.

- 10 Overlays
- 6 sites
- 2 Edge Devices per site (12 total)
- 256 VLANs extended via OTV
- 16000 MAC addresses across all the extended VLANs
- 500 (*,G) and 1500 (S,G) for all OTV connected sites

The max number of VLANs and MAC addresses is considered across all defined Overlays (and not on a per-Overlay basis).

Note The scalability values highlighted above are mostly dependent on internal support considerations and not driven by HW or SW limitations. Directly contact the Cisco account team for information on the currently supported scalability figures.

OTV Hardware Support and Licensing Information

The only Cisco platform supporting OTV functionality at this time is the Nexus 7000. Figure 1-37 shows all existing M1 linecards OTV fully supports.

Figure 1-37 M1 Linecards Supporting OTV

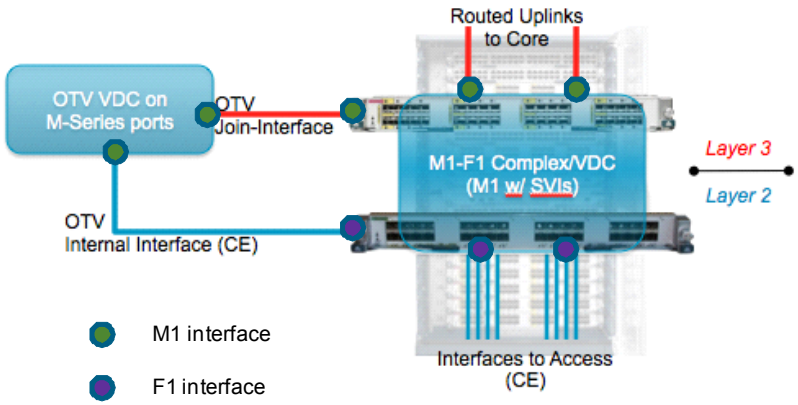
I/O Module	Photo	OTV Support
N7K-M132XP-12 N7K-M132XP-12L		✓
N7K-M148GT-11 N7K-M148GT-11L		✓
N7K-M148GS-11 N7K-M148GS-11L		✓
N7K-M108X2-12L		✓

Full support essentially means that both Join and internal interfaces can be configured on one of the M1 modules mentioned. OTV on F1 modules is not supported on NX-OS releases shipping at the time of writing of this document (5.2 and 6.0 releases) and will be introduced in a future software release but restricted only to internal interfaces (an M1 linecard will always be required on the OTV edge device to source the OTV encapsulated traffic out of the Join interface).

From a licensing perspective, it is important to note how OTV support requires the use of the new Transport Services (TRS) license. Depending on the specifics of the OTV deployment, the Advanced License may be required as well to provide Virtual Device Contexts (VDCs) support. More information on VDC requirements for OTV can be found in the following "OTV Deployment Options" section on page 1-34.

Also, in the specific scenarios where OTV is deployed on a dedicated VDC, it is possible to leverage F1 interfaces on the default VDC to connect to the internal interfaces of the OTV edge device, as highlighted in Figure 1-38.

Figure 1-38 Use of M1 and F1 Interfaces for OTV Deployment



OTV Deployment Options

OTV employs several deployment models. It is difficult to provide a universal response to where the OTV edge devices should be deployed in the data center, given design variations and permutations found in real network data center deployments. The following sections highlight three specific OTV deployment scenarios, which are expected to cover most of real life OTV design options.

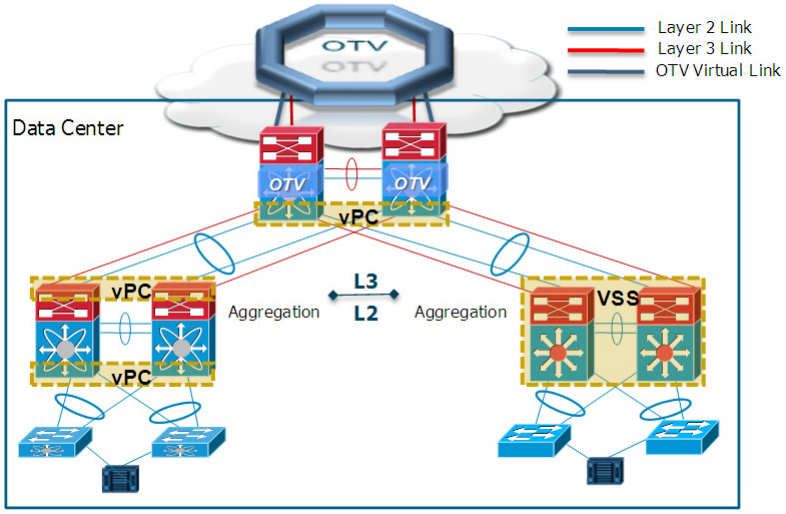
OTV in the DC Core

The first deployment model is targeted to network designs large enough to justify the use of a dedicated data center Core layer to interconnect different aggregation blocks. The term POD will be used in the context of this document as synonymous of aggregation block. A POD is therefore represented by the server, access and aggregation layers. This specific design proposes the deployment of OTV edge devices in the data center Core layer. Two flavors of this design are possible, depending on where the demarcation line between Layer 2 and Layer 3 network domains exists.

Layer 2-Layer 3 Boundary at Aggregation

This deployment model is shown in Figure 1-39.

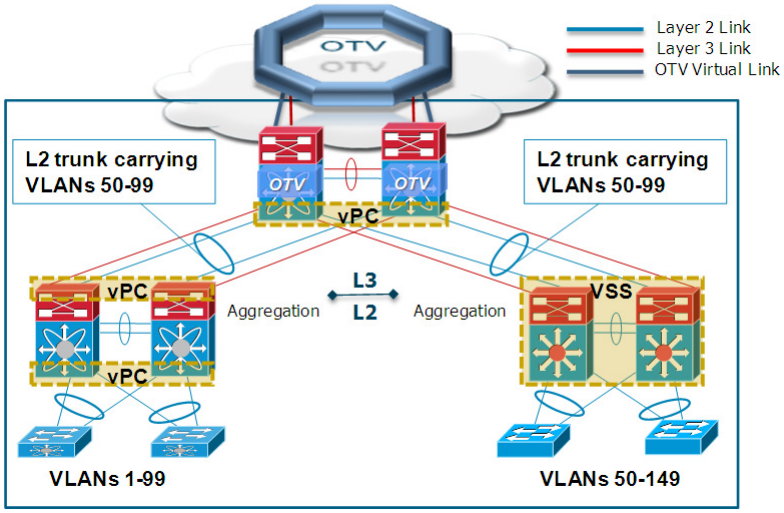
Figure 1-39 OTV in the Core (Layer 2-Layer 3 Boundary at the Aggregation)



In this design, the DC Core devices perform Layer 3 and OTV functions. In a traditional design, each DC POD is connected to the transport layer via routed links. Building a routed network infrastructure is a best practice for increasing the resiliency and stability of the overall design. For core devices to start functioning as OTV edge devices, traffic for all the VLANs that need to be extended to remote sites must now be carried to the DC core.

Figure 1-40 proposes a design where Layer 2 connectivity for a set of VLANs is required not only with the remote sites, but also between the aggregation blocks belonging to a same data center location. 100 VLANs defined in each aggregation block where Layer 2 connectivity is only required for a subset of them (50 VLANs belonging to the range 50-99 in this example).

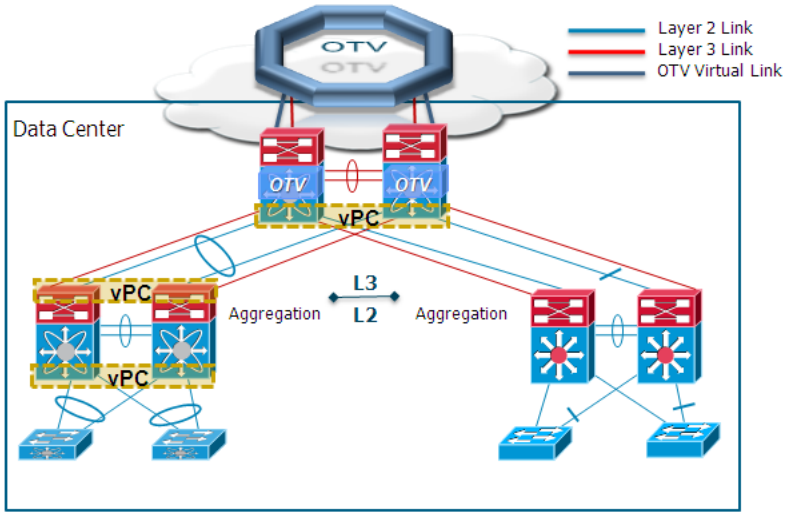
Figure 1-40 Layer 2 Connectivity Intra- and Inter- Data Center



The recommended method of modifying the traditional design of extending Layer 2 connectivity to the DC transport suggests deploying a dedicated set of fiber links to connect each POD to the transport devices. Whenever possible, these links should be configured as part of a logical EtherChannel bundle configured as a Layer 2 trunk carrying all VLANs that require LAN extension services (intra- and inter-data center). This creates a topology that is Layer 2 loop-free, emphasizing the need to avoid jeopardizing overall design stability. It is necessary to create the logical hub-and-spoke Layer 2 topology using separate physical cables to carry Layer 2 traffic because of vPC-specific lack of support for establishing route peering on a dedicated VLAN carried on a Layer 2 trunk.

Figure 1-41 shows what happens when devices deployed at the aggregation layer of the right POD do not have MCEC capabilities.

Figure 1-41 OTV in the Core with non MCEC Capable POD Devices



The disadvantage in this case is that not all the Layer 2 links can be used to forward traffic to the OTV edge devices, since one of the Layer 2 links would normally be blocked by Spanning-Tree. Also, a topology change notification (TCN) would be created every time a failure causes a STP convergence event. The notification is sent to all the PODs where the common VLANs are defined. At the same time,

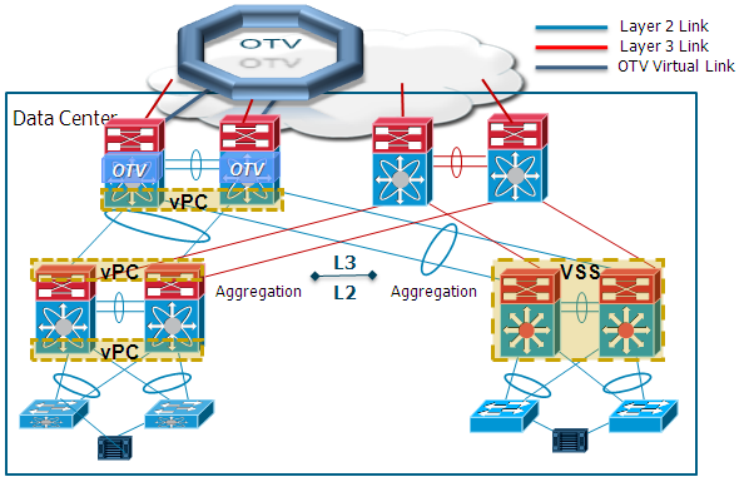
since this design does not require the use of vPC, it would be possible to share a common set of physical links between each POD and the OTV edge devices to establish Layer 2 and Layer 3 connectivity. Figure 1-41 shows a separate set of connections.

Some additional considerations for the recommended design shown in Figure 1-40 are as follows:

- Routing for all the VLANs defined at the access layer still occurs at the aggregation layer. Traffic is then routed to the core layer by leveraging the pre-existing routed links (shown as red lines).
- The STP domain for the extended VLANs spans across the entire data center network (aggregation blocks and core). The use of STP BPDU filtering is not recommended inside the same data center physical location, because STP should always be enabled to detect loops created via configuration errors or cabling mistakes. Consequently, it is important to modify the positioning of the STP root and backup bridges to deploy them in the DC core. The use of MCEC functionality allows building a STP domain that, even if larger, is more stable since no single failure would cause the generation of a TCN notification.
- Storm-control configuration is recommended on the Layer 2 trunk connections between aggregation and transport to limit the exposure to broadcast, multicast, or unknown unicast storms across different aggregation blocks.
- OTV is leveraged in this design only to provide Layer 2 connectivity to remote data center sites. LAN extension between the different aggregation blocks deployed in the same physical location is achieved by traditional Layer 2 switching occurring at the core.
- If PIM is part of the configuration of the core uplink interfaces used as OTV join-interfaces it's important to make sure that those interfaces do not become the PIM DRs by changing the interfaces PIM DR priority.

Figure 1-42 shows a different flavor of the same design previously discussed. The only difference is the use of two devices dedicated to perform OTV functions. These can be physical devices or Virtual Device Contexts (VDCs) carved out of the Nexus 7000 deployed in the core.

Figure 1-42 OTV Deployed on Dedicated Devices

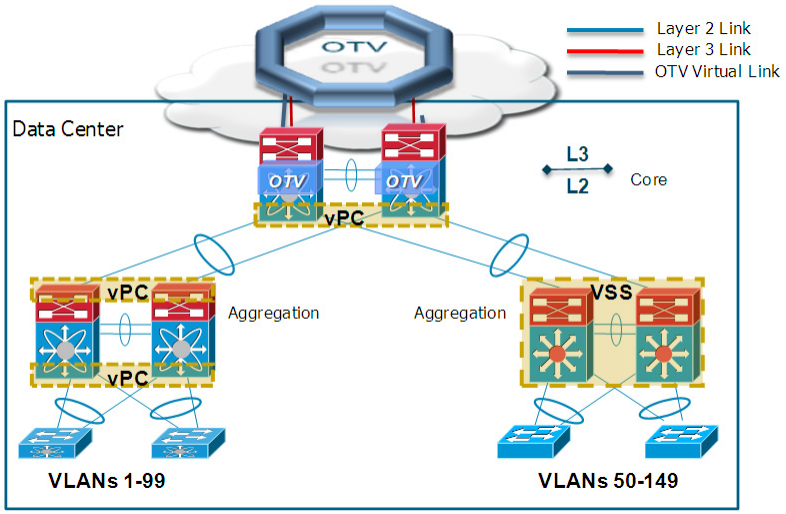


From a logical perspective, the design in Figure 1-42 completely separates the infrastructure used to provide Layer 2 connectivity (intra and inter sites) from the one providing Layer 3 connectivity services. All other design considerations hold true in this design.

Layer 2-Layer 3 Boundary at the Core

Figure 1-43 shows a different deployment option, where the OTV edge devices are still positioned in the DC Core, but this network layer now represents the boundary between Layer 2 and Layer 3 also.

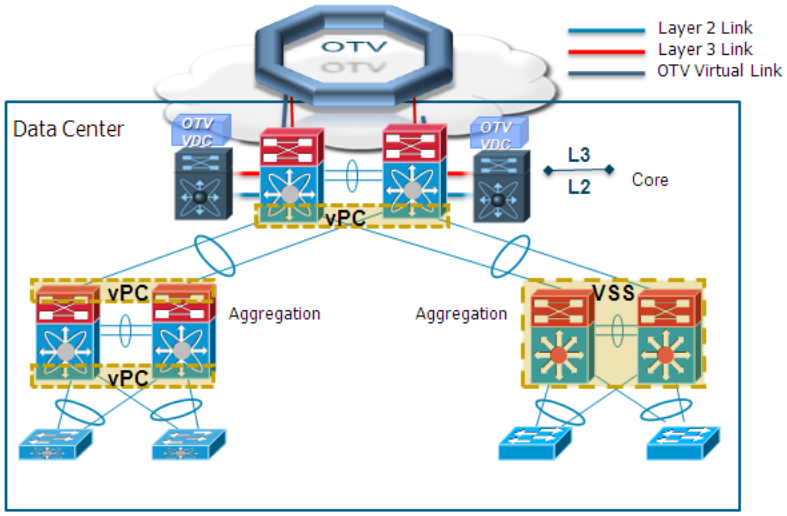
Figure 1-43 OTV in the Core (Layer 2-Layer 3 Boundary at the Core)



Layer 2 trunks are used between each aggregation block and the transport, and similarly, as in the previous deployment scenario, logical port channels are recommended to interconnect each POD to the transport layer devices. Routing for all VLANs now occurs at the transport layer. All other design considerations (STP root positioning, storm-control requirement and so on) from the previous model apply.

Since in this model SVI routing and OTV LAN extension are performed on the same devices (the core Nexus 7000 switches), it is required to introduce the use of a dedicated VDC to perform the OTV functions because of the considerations made in the "OTV and SVIs Coexistence" section on page 1-30. As a consequence, the design would actually become the one shown in Figure 1-44.

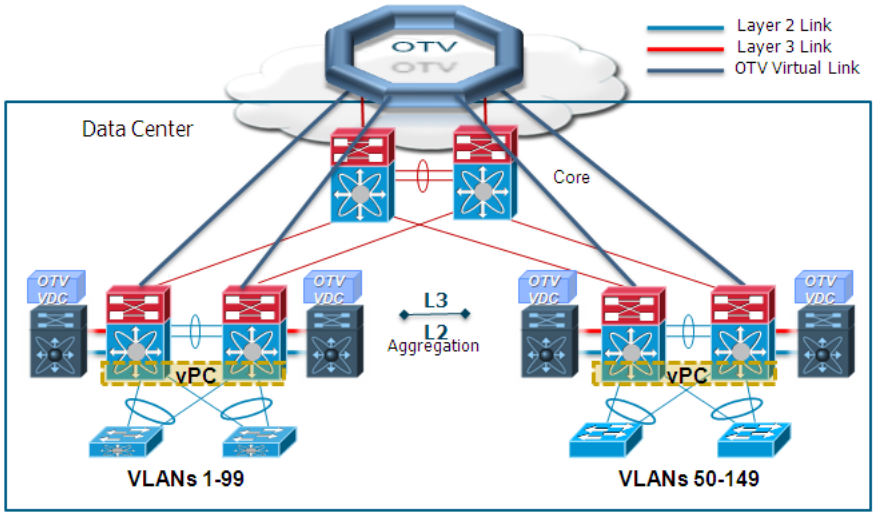
Figure 1-44 OTV VDCs at the Core Layer



OTV at the DC Aggregation

A popular design model proposes the deployment of OTV functionality at the aggregation layer, where the demarcation line between Layer 2 and Layer 3 is located (Figure 1-45).

Figure 1-45 OTV at the Aggregation Layer



Once again, the use of OTV VDCs is introduced in this deployment model, but differently from the ones discussed above there is no need now to modify the design of the data center network. Layer 3 links are still connecting each aggregation block to the DC core and VLANs remain contained inside each POD.

Some important design considerations:

- Each aggregation block represents a separate STP domain. This is the result of both having routed links interconnecting each POD to the core and having native OTV STP isolation functionality. This means that a separate root and backup root can be defined on a per-POD basis. Also, a control plane issue happening in a given aggregation block would not impact the other PODs.
- Storm-control configuration is now simplified, because of failure isolation functionality (for broadcast and unknown unicast) natively available with OTV.
- OTV is leveraged in this design to provide Layer 2 connectivity both to remote data center sites and between the different aggregation blocks deployed in the same physical location.

An OTV capable device (Nexus 7000 only at this time) needs to be deployed at the aggregation layer.

Because of these considerations, this model is usually positioned in green field data center designs, where Nexus 7000 is commonly deployed at the aggregation layer or in a collapsed core/aggregation layer design. Best practices and configuration guidelines focused on this specific deployment option can be found in the "OTV at the DC Aggregation" section on page 1-38.

OTV in Point-to-Point Deployments

The final model is positioned for point-to-point OTV deployment where the two data center sites are connected with dark fiber links (or leveraging protected DWDM circuits).

Figure 1-46 Point-to-Point OTV Deployments with Collapsed DC Aggregation/Core

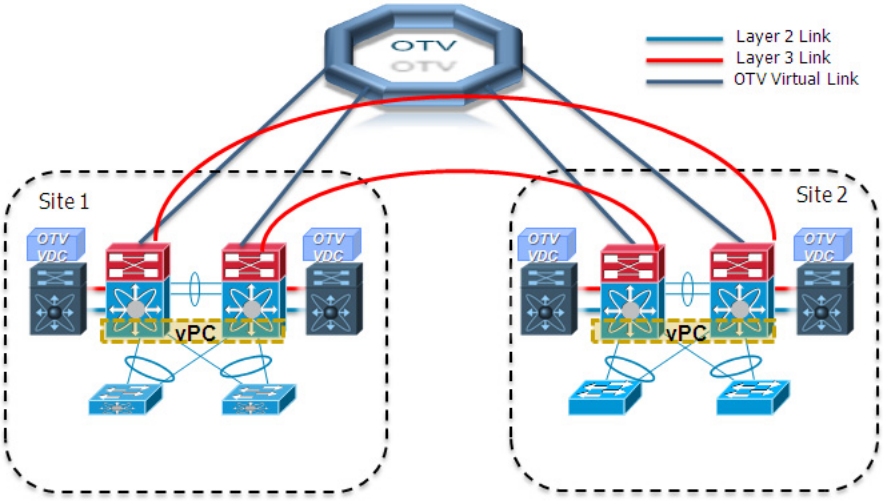
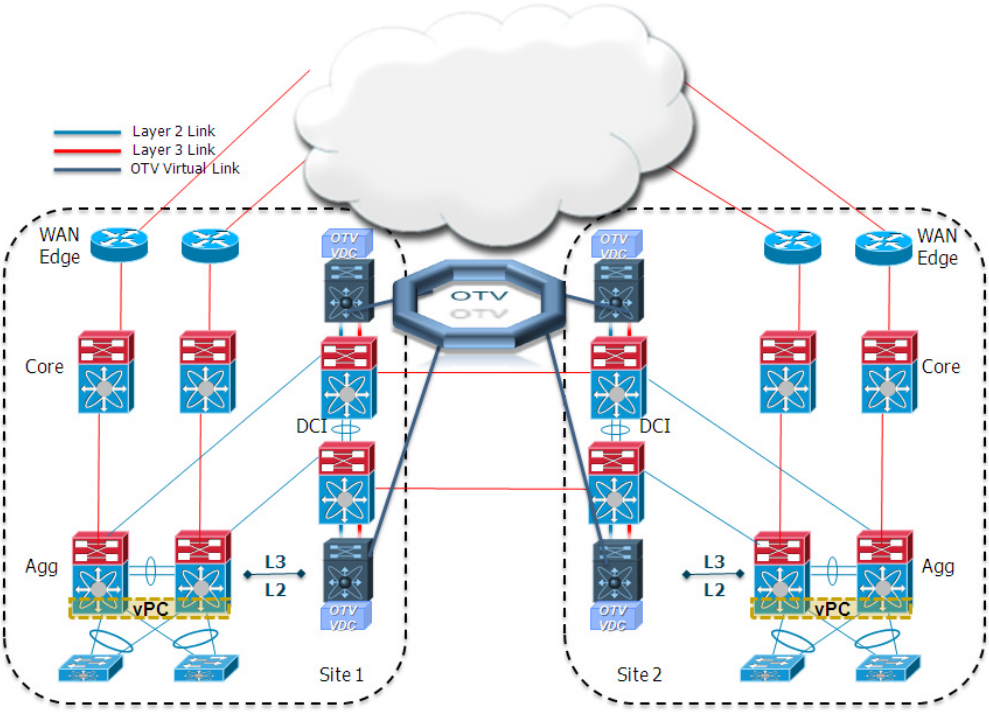


Figure 1-46 highlights a network with collapsed aggregation/transport data center layers and shows once again the deployment of the OTV VDCs to allow the coexistence on the same physical Nexus 7000 devices of the SVI routing functionality.

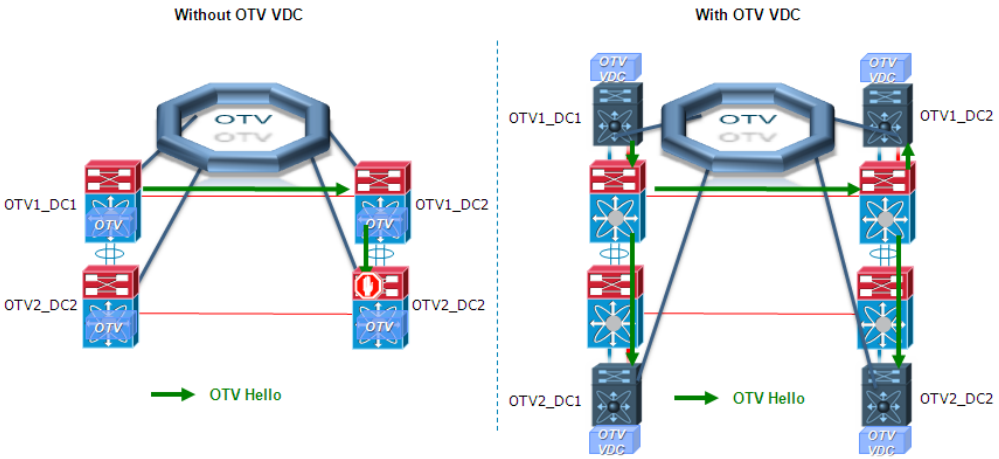
The use of OTV VDCs is also required in point-to-point deployments where a pair of separate physical devices is deployed in a dedicated DCI layer to perform the OTV functionality (Figure 1-47).

Figure 1-47 Point-to-Point OTV Deployments with Dedicated DCI Layer



The use of VDCs must be introduced in this case, even if the SVIs are not defined on the DCI layer devices. The basic requirement for OTV to be functional is in fact that every OTV control packet originated by the Join interface of an OTV edge device must be received on the Join interface of all the other edge devices being part of the same OTV overlay. This requirement can only be met in a dark fiber scenario if the OTV functionality is moved on a dedicated VDC, as highlighted in Figure 1-48.

Figure 1-48 Delivery of OTV Hello Packets on Join Interface



The left shows what happens when the OTV edge device 1 in the left DC generates an OTV control plane packet. The packet is sourced from the Join Interface and it is correctly received on the Join Interface of the edge device 1 in the right DC. That device then forwards the frame toward the edge device 2 on the portchannel transit link between them. When the packet is received, edge device 2 drops it because it was not received on the Join Interface. The same process happens with control packets generated by the other edge devices. The end result is that the 4 OTV edge devices fail to become fully adjacent between them.

The right side of Figure 1-48 highlights how the deployment of the OTV VDCs ensures that the control protocol packets will always be received on the Join interfaces of all the deployed OTV edge devices. Notice, for this to happen, it is required that routing and PIM is enabled on the port-channel link connecting the two Nexus 7000 devices (Layer 3 peering can be established between SVIs on a dedicate VLAN).

The point-to-point deployment of LAN extensions between remote data center sites over dedicated dark fiber connections has been already discussed in the DCI System Release 1.0 Design Guide available at the link below:

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns949/ns304/ns975/data_center_interconnect_design_guide.pdf

The solution discussed in that guide proposed the use of Multi-Chassis EtherChannel technologies (VSS and vPC) to logically bundle the dark fiber links and extend VLAN connectivity across a back-to-back EtherChannel connection. The deployment of OTV in a similar topology brings the following design advantages:

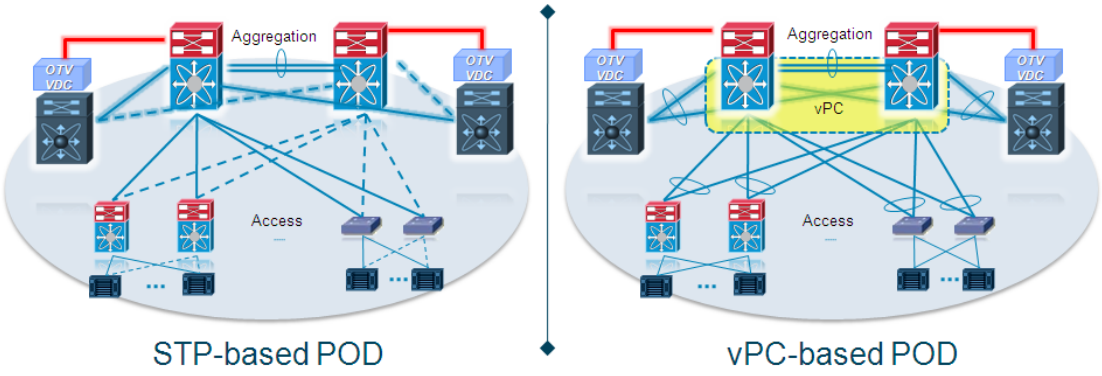
- Provision of Layer 2 and Layer 3 connectivity leveraging the same dark fiber connections: As discussed in the design guide, a dedicated set of fibers is required to provide Layer 2 and Layer 3 connectivity when leveraging vPC at least in one site.
- Native STP isolation: There is no need to explicitly configure BPDU filtering to prevent the creation of a larger STP domain extending between the two sites.

- Improved Layer 2 data plan isolation: The required storm-control configuration is simplified in the OTV deployment scenario because of the native suppression of unknown unicast frames and for the broadcast containment capabilities of the protocol.
- Simplified provisioning of the FHRP isolation functionality to have an active default gateway available at each site.

Deploying OTV at the DC Aggregation

In the deployment model where OTV is enabled at the Aggregation layer (at the boundary between Layer 2 and Layer 3 network domains), the OTV VDC can be envisioned just as another block attached to the Aggregation layer and in this context the OTV internal interfaces seamlessly participate in the STP or vPC topologies that are already in place within the existing data center site. Figure 1-49 shows the two possible models: on the left the STP-based POD deployment, on the right the vPC based one.

Figure 1-49 OTV VDC Site Transparency



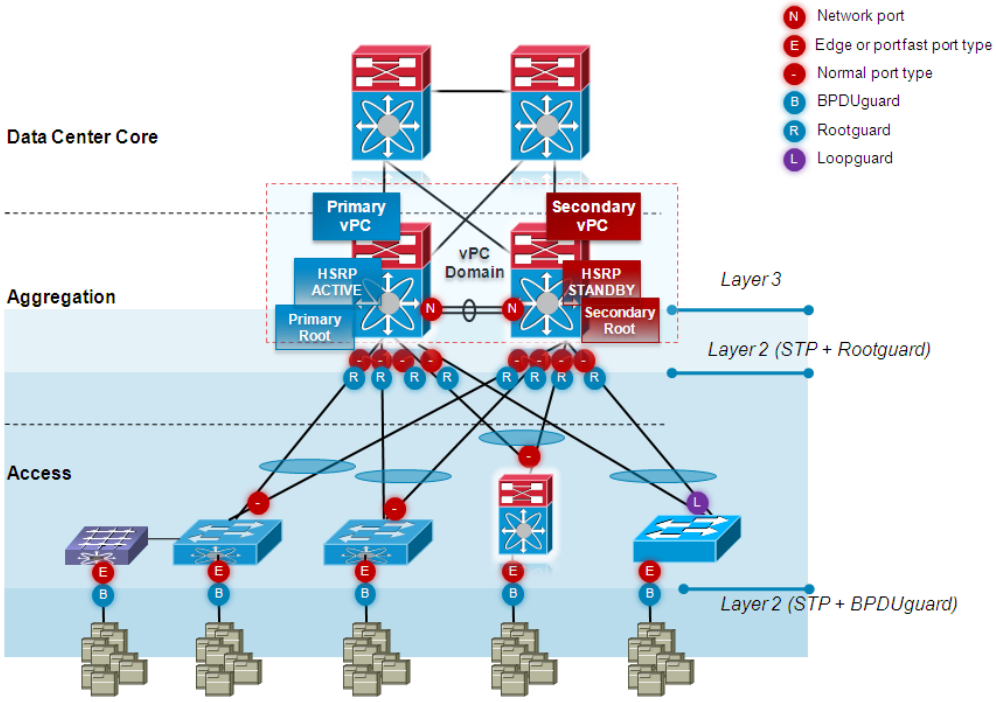
Spanning Tree Considerations

Before configuring OTV you should review and implement Cisco recommended STP best practices at each site. OTV is independent from STP but it greatly benefits from a stable and robust Layer 2 topology. An in-depth description of all STP features and guards can be found in the Cisco Data Center Design Zone at

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html

As a quick reference Figure 1-50 shows where the STP features are deployed within different areas of the data center.

Figure 1-50 STP Data Center Best Practices



The downstream interfaces leading to the access switches (and to the OTV VDC internal interfaces) are configured for Rootguard on the aggregation devices. Access relies on the STP information from the aggregation, but not the other way around. This is a particularly important best practice for OTV as it protects from backdoor Layer 2 connections to other OTV sites.

As previously mentioned, the OTV VDC can be envisioned just as another block attached to the Aggregation layer. This means that the OTV internal interfaces seamlessly participate in the STP topology deployed in the data center. Figure 1-51 and Figure 1-52 shows a couple of deployment options that mainly differ on the way traffic is sent from the access layer toward the OTV VDC devices.

Figure 1-51 Single STP Root for All the VLANs

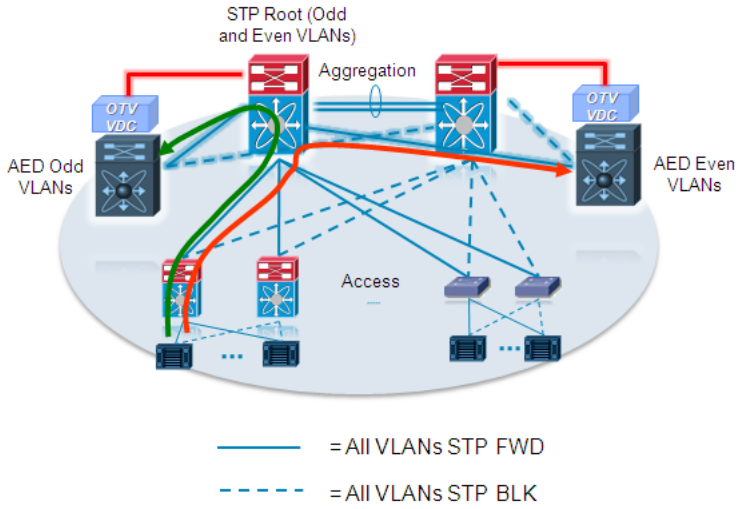
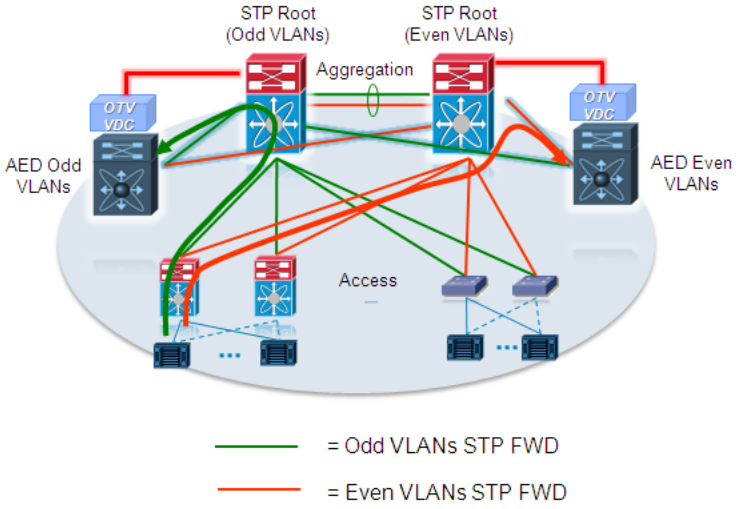


Figure 1-52 Different STP Root for Odd and Even VLANs



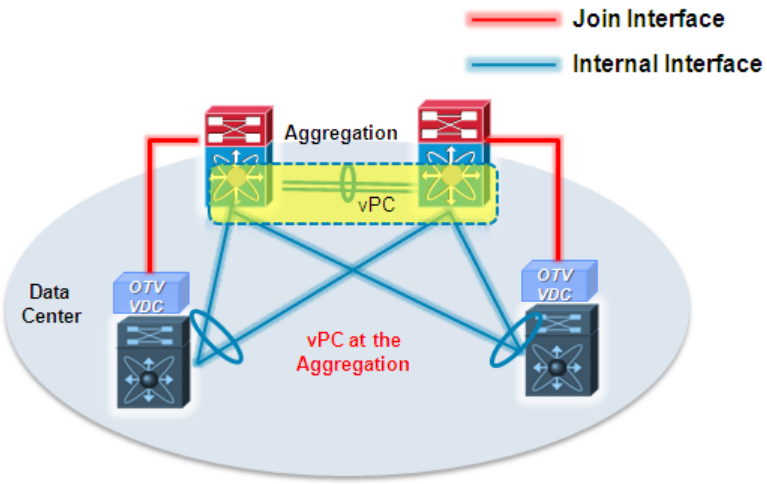
The model in Figure 1-52 leverages the OTV multi-homing property of splitting the VLANs between edge devices (one becomes Authoritative for the even VLANs, the other for the odd ones).

The choice of what design to deploy is a tradeoff between the operational simplicity of the first option, versus a better traffic load balance (from the access to the aggregation layer devices) of the second one.

vPC Considerations

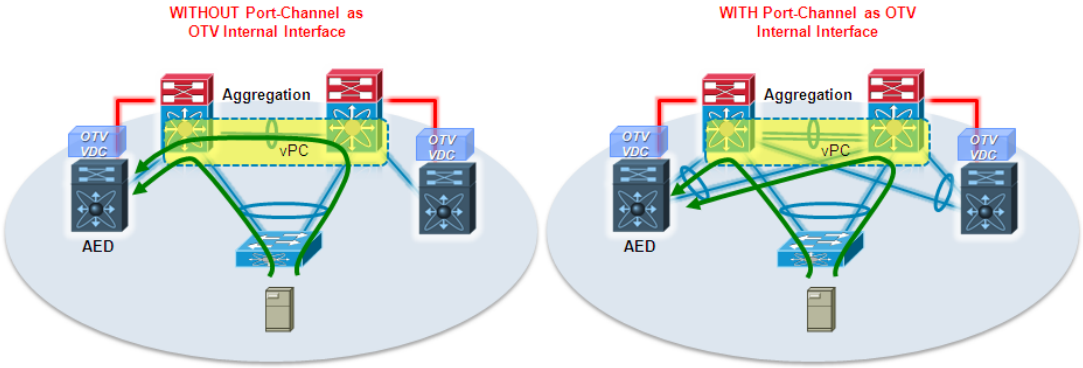
When deploying OTV in a vPC-based site the network administrator has two options (Figure 1-53).

Figure 1-53 vPC Based Design



In this model, the two OTV VDCs leverage a locally defined port-channel to connect their internal interfaces to the vPC peers formed at the aggregation layer. This option provides a good level of resiliency with the minimum amount of ports. Also, from a traffic flow point of view, the use of vPC at the aggregation allows to optimize the Layer 2 flows originated from server connected to the access layer devices that need to be sent to the OTV AED.

Figure 1-54 Traffic Flow Considerations



As shown in Figure 1-54, there is always going to be an optimal shortest path available for this type of traffic, avoiding the use of the vPC peer-link between the aggregation devices.

Guidelines related to OTV when deployed with vPC:

- The vPC domain IDs across all the OTV connected sites must be different.
- In 5.0(3) NX-OS release when deploying OTV with vPC, the OTV ARP Cache must be disabled with the no otv suppress-arp-nd command on the overlay interface, as this could cause connectivity issues. This issue has been fixed in the following 5.1(1) software release.

The network administrator should also deploy all Cisco best practices recommended for vPC which are outside the scope of this document.

OTV Configuration

This OTV configuration addresses OTV deployment at the aggregation layer where the boundary between Layer 2 and Layer 3 is positioned. As discussed in the "OTV and SVIs Coexistence" section on page 1-30, defining a dedicated VDC to perform OTV functions is required. Creating an OTV VDC configuration is outside the scope of this document. Refer to the NX-OS Configuration Guide for VDC setup.

http://www.cisco.com/en/US/docs/switches/datacenter/sw/5_x/nx-os/virtual_device_context/configuration/guide/vdc_nx-os_cfg.html

The following logical procedure highlights the following configuration steps:

Step 1 Minimal configuration required to enable OTV functionality. In a multi-homed site, OTV should first be brought up as single-homed by enabling OTV on a single OTV edge device in each site. This phased approach ensures that sites and transport networks are properly configured before enabling multi-homing.

Step 2 Multi-homing configuration: if the data centers are OTV multi-homed, it is a recommended best practice to bring the Overlay up on the redundant edge device only after the OTV connection has been tested in as single-homed. This section would detail what are the steps required in order to provide fault tolerance with redundant OTV edge devices deployed in each data center site and belonging to the same overlay.

Step 3 FHRP isolation configuration to provide optimal outbound traffic path, as previously discussed in the "FHRP Isolation" section.

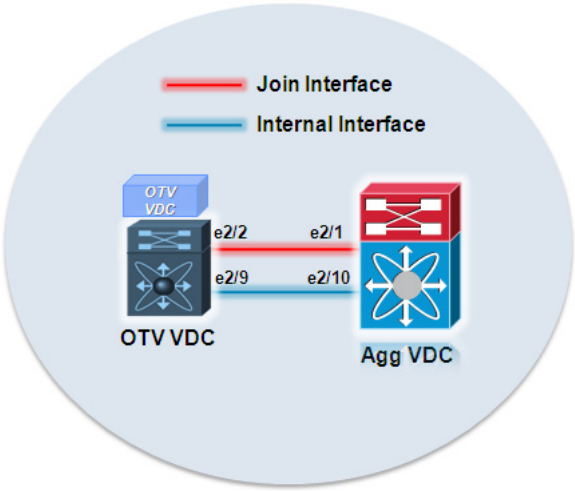
Note Complete devices configuration can be found in Appendix A.

Configuring OTV on a Single Edge Device (Multicast Mode)

Figure 1-55 serves as a reference topology for this part of the configuration.

Note These configuration considerations can also be used for other OTV deployments.

Figure 1-55 OTV Configuration Topology of Reference



Step 1 Configure OTV VDC VLANs.

a. Configure VLANs to be extended via OTV on the OTV VDC.

Recommendation: Enable only the data VLANs to be extended (VLAN 5 to VLAN 10 in this example) and the OTV site-vlan (VLAN 15 in this example).

```
OTV-VDC-A(config)# vlan 5-10, 15
```

Step 2 Configure Join Interface.

Before configuring the join-interface let's consider the following guidelines:

- Only one join interface can be specified per overlay. You can decide to use one of the following methods:
  - Configure a single join interface, which is shared across multiple overlays.
  - Configure a different join interface for each overlay, which increases the OTV reliability and load balancing capabilities.
- For a higher resiliency, it is possible to leverage a Layer 3 port channel, but it is not mandatory. There are no requirements for 1 Gigabit-Ethernet versus 10 Gigabit-Ethernet or dedicated versus shared mode on the Join Interface.

- In the 5.0(3) Cisco NX-OS release supporting OTV, the join interface must belong to the default VRF. This restrictions has been lifted up from release 5.1(1). Also, only Layer 3 physical interfaces (and subinterfaces) or Layer 3 port channel interfaces (and subinterfaces) can be configured as Join interfaces in this initial OTV software release.

a. Configure the join-interface (in this example a Layer 3 physical interface):

```
OTV-VDC-A(config)# interface ethernet 2/2
OTV-VDC-A(config-if)# description [ OTV Join-Interface ]
OTV-VDC-A(config-if)# ip address 172.26.255.98/30
OTV-VDC-A(config-if)# ip igmp version 3
OTV-VDC-A(config-if)# no shutdown
```

Recommendation: as in the 5.0(3) NX-OS release a defect prevents the use of IP addresses with the last octect in the range 224–239. This issue has been fixed in the following 5.1(1) release.

Note the configuration of IGMPv3 on the join-interface. This allows OTV to join the SSM groups of the transport/core network needed to extend the sites' multicast traffic.

b. Configure the other side of the link on the aggregation VDC.

```
AGG-VDC-A(config)# interface ethernet 2/1
AGG-VDC-A(config-if)# description [ Connected to OTV Join-Interface ]
AGG-VDC-A(config-if)# ip address 172.26.255.99/30
AGG-VDC-A(config-if)# ip pim sparse-mode
AGG-VDC-A(config-if)# ip igmp version 3
AGG-VDC-A(config-if)# no shutdown
```

Note the configuration of IGMPv3 as well as PIM sparse mode on the aggregation interface. The assumption is that PIM is already configured on the aggregation VDC.

The multicast configuration needed by OTV in the core/transport is very basic. The Cisco multicast best practices to be followed in transport do not vary with the deployment of OTV and are out of scope for this paper. The following paper describes the IP multicast best practices for Enterprise Customers:

http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6552/ps6592/whitepaper_c11-474791.html

Following is an example of PIM configuration at the aggregation VDC:

```
AGG-VDC-A# show running pim
feature pim
interface ethernet 2/1
 ip pim sparse-mode
interface ethernet 1/10
 ip pim sparse-mode
interface ethernet 1/20
 ip pim sparse-mode

ip pim rp-address 172.26.255.101
ip pim ssm range 232.0.0.0/8
```

On the Nexus 7000, Source Specific Multicast (SSM) is enabled by default once PIM is configured. The SSM default range is 232.0.0.0/8.

c. Verify the join interface connectivity through the transport by pinging the other sites' join interface IP addresses.

The connectivity test must be successful, because OTV, as a requisite, needs IP reachability between each site across the transport.

You can also have the Join Interface be part of your IGP protocol or leverage static routing configuration (this latter option will be shown at Step 4-h).

Step 3 Configure Internal Interfaces.

a. Configure the OTV internal interfaces that carry the VLANs to be extended together with the site VLAN. The configuration is identical on both sides of the link, in this case e2/10 on OTV-VDC-A and e2/12 on AGG-VDC-2.

```
interface ethernet 2/10
 description [ OTV Internal Interface ]
 switchport
 switchport mode trunk
 switchport trunk allowed vlan 5-10, 15
 no shutdown
```

From a Spanning-Tree perspective, the OTV VDC looks like as an access layer switch connected to the aggregation, so the same configuration suggested in Figure 1-50 should be adopted here as well.

Step 4 Configure Overlay Interface for multicast enabled transport infrastructure. If a specific deployments needs to leverage a unicast-only core, skip to "Configuring OTV in Unicast-Only Mode" section on page 1-49.

a. Enable OTV. By default, the OTV feature is disabled on the device. You must explicitly enable the OTV feature to access the configuration.

```
OTV-VDC-A(config)# feature otv
```

b. From release 5.2(1) only: configure the OTV site-identifier. This value should be identical for all the OTV edge devices belonging to the same DC site.

```
OTV-VDC-A(config)# otv site-identifier 0x1
```

c. Configure the otv site-vlan. OTV sends hello messages on the site VLAN to determine if there are other edge devices on the local site. The site VLAN is shared across all the overlays within the site.

Recommendations:

- Use a dedicated VLAN as OTV site VLAN.
- Do not extend the OTV site VLAN.
- Ensure that the site VLAN is active on the OTV internal interfaces and on the port channel link connecting to the other aggregation layer device. It is critical to enable the site VLAN on multiple internal interfaces, because at least one of these interfaces needs to be always up in order for the OTV Edge Device to be able to forward OTV traffic.
- The Site-VLAN must be configured before entering the **no shutdown** command for any overlay interface and must not be modified while any overlay is up within the site.
- Using the same site VLAN at each site is not mandatory, but it could help during debugging and provide protection in case of accidental site merging.
- Finally, the site VLAN should always be defined, even in scenarios where a single OTV Edge Device is defined in a given site. Missing the site VLAN definition would not allow the OTV Edge Device to forward OTV encapsulated traffic.

In this example the site VLAN is VLAN 15.

```
OTV-VDC-A(config)# otv site-vlan 15
```

d. Create and configure the Overlay interface.

It is possible to configure multiple overlay interfaces, but there must be no overlapping of VLANs between the different configured overlays. Another consideration to take into account is that the overlay interface number must match on all the OTV edge devices that should be part of the same overlay.

The first command under this logical interface specifies the interface to be used as join interface that was configured in a previous step above.

```
OTV-VDC-A(config)# interface overlay 1
OTV-VDC-A(config-if)# otv join-interface ethernet 2/2
```

e. As part of the overlay interface configuration, it is necessary to specify the OTV control-group.

As explained in Control Plane Considerations, page 1-4, this is a single PIM-SM or PIM-bidir group used to form adjacencies and exchange MAC reachability information.

```
OTV-VDC-A(config-if)# otv control-group 239.1.1.1
```

f. Define the OTV data-group used to map the sites' multicast groups to a range of SSM addresses in the transport network to carry the sites' multicast traffic.

```
OTV-VDC-A(config-if)# otv data-group 232.1.1.0/26
```

A common question is related to the dimensioning of the data-group. Theoretically, a single multicast group could be defined as an OTV data group. As always, the right number of groups to be used depends on a tradeoff between the amount of multicast state to be maintained in the core and the optimization of Layer 2 multicast traffic delivery. If a single data group was used in the core to carry all the (S,G) site multicast streams, remote sites would receive all the streams as soon as a receiver joined a specific group. On the other side, if a dedicated data group was used for each (S,G) site group, each site would receive multicast traffic only for the specific groups joined by local receivers.

g. Specify the range of VLANs to be extended across the Overlay and bring up the interface:

```
OTV-VDC-A(config-if)# otv extend-vlan 5-10
```

h. Create a static default route to point out the Join interface. This is required to allow communication from the OTV VDC to the Layer 3 network domain.

```
OTV-VDC-A(config)# ip route 0.0.0.0/0 172.26.255.99
```

i. Bring up the overlay interface.

If the data centers are OTV multi-homed, it is a recommended best practice to bring the Overlay up in single-homed configuration first, by enabling OTV on a single edge device at each site. After the OTV connection has been tested in as single-homed, then enable the functionality on the other edge devices of each site.

```
OTV-VDC-A(config-if)# no shutdown
```

j. Verify the OTV edge device joined the Overlay, discovered the other edge devices in the remote sites, and established OTV control plane adjacencies.

```
OTV-VDC-A# sh otv overlay 1
OTV Overlay Information
Overlay interface Overlay1
 VPN name            : Overlay1
 VPN state           : UP
 Extended vlans      : 5-10 (Total:6)
 Control group       : 239.1.1.1
 Data group range(s) : 232.1.1.0/26
 Join interface(s)   : Eth2/2 (172.26.255.98)
 Site vlan           : 15 (up)

OTV-VDC-A# show otv adjacency
Overlay Adjacency database
Overlay-Interface Overlay1 :
Hostname       System-ID      Dest Addr            Up Time   Adj-State
OTV-VDC-C      0022.5579.7c42 172.26.255.90        1w0d      UP
```

```
OTV-VDC-D       0022.5579.36c2 172.26.255.94         1w0d       UP
```

k. Display all MAC addresses learned locally, and via the OTV connection.

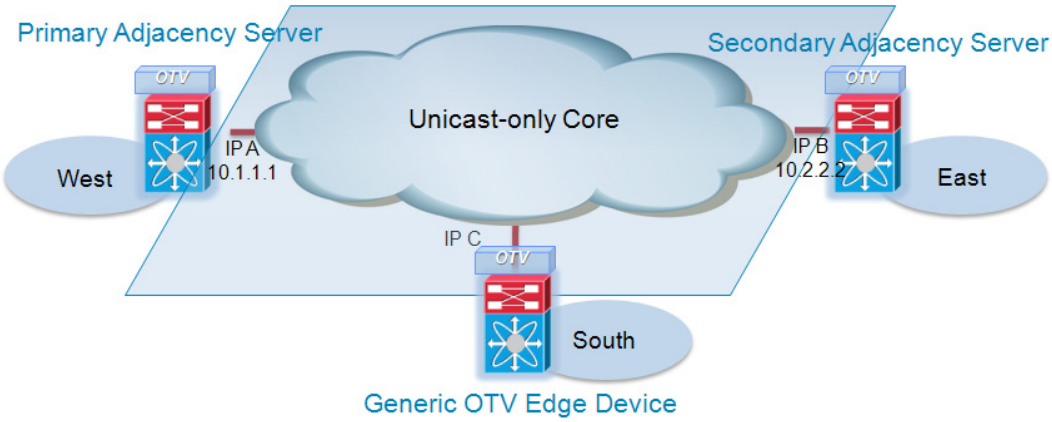
```
OTV-VDC-A# show otv route
OTV Unicast MAC Routing Table For Overlay1
VLAN MAC-Address     Metric  Uptime    Owner      Next-hop(s)
---- --------------  ------  --------  ---------  -----------
 10  0000.0c07.ac64  1       00:00:55  site       ethernet 2/10
 10  001b.54c2.3dc1  1       00:00:55  Overlay    OTV-VDC-C
```

Configuring OTV in Unicast-Only Mode

Two pieces of configuration are required to deploy OTV across a unicast-only transport infrastructure: first, it is required to define the role of Adjacency Server (usually enabled on a generic OTV edge device), whereas the other piece of configuration is required in each OTV edge device not acting as an Adjacency Server (i.e acting as a client). All client OTV edge devices are configured with the address of the Adjacency Server. All other adjacency addresses are discovered dynamically. Thereby, when a new site is added, only the OTV edge devices for the new site need to be configured with the Adjacency Server addresses. No other sites need additional configuration.

The recommendation is usually to deploy a redundant pair of Adjacency Servers in separate DC sites, as shown in Figure 1-56.

Figure 1-56 Configuration for Adjacency Server and Generic OTV Device



The following exhibits show sample configuration for Adjacency server (Primary and Secondary) and the generic OTV Device (OTV client). Notice that an Adjacency Server is at the same time also an OTV Client (we can think of the Adjacency Server functionality as a special process that happens to be running on a generic OTV edge device).

a. Primary Adjacency Server Configuration

```
feature otv
otv site-identifier 0x1
otv site-vlan 15
interface Overlay1
  otv join-interface e2/2
  otv adjacency-server unicast-only
  otv extend-vlan 5-10
```

b. Secondary Adjacency Server Configuration

```
feature otv
otv site-identifier 0x2
otv site-vlan 15
interface Overlay1
  otv join-interface e1/2
  otv adjacency-server unicast-only
  otv use-adjacency-server 10.1.1.1 unicast-only
  otv extend-vlan 5-10
```

c. Generic OTV Edge Device Configuration

```
feature otv
otv site-identifier 0x3
otv site-vlan 15
interface Overlay1
  otv join-interface e1/1
  otv use-adjacency-server 10.1.1.1 10.2.2.2 unicast-only
  otv extend-vlan 5-10
```

As shown above, the configuration on the Primary Adjacency Server is very simple and limited to enable AS functionality (**otv adjacency-server** command). The same command is also required on the Secondary Adjacency Server device, but also needs to point to the Primary AS (leveraging the **otv use-adjacency-server** command). Finally, the generic OTV Edge Device must be configured to use both the Primary and Secondary Adjacency Servers. The sequence of adjacency server address in the configuration determine primary or secondary adjacency server role. As previously mentioned, this order is relevant since an OTV edge device will always use the neighbor-list provided by the Primary Adjacency Server, unless it detects that specific device is not available anymore (control plane Hellos are always exchanged as keepalives between each OTV device and the Adjacency Servers).

The following exhibits show the CLI output that helps you identify the role of an OTV device. The output is self-explanatory.

Primary Adjacency Server

```
Primary_AS# show otv overlay 1
OTV Overlay Information
Site Identifier 0000.0000.0001
Overlay interface Overlay1
 VPN name            : Overlay1
 VPN state           : UP
 Extended vlans      : 5-10 (Total:6)
 Join interface(s)   : E2/2 (10.1.1.1)
 Site vlan           : 15 (up)
 AED-Capable         : Yes
 Capability          : Unicast-Only
 Is Adjacency Server : Yes
 Adjacency Server(s) : [None] / [None]
```

Secondary Adjacency Server

```
Secondary_AS# show otv overlay 1
OTV Overlay Information
Site Identifier 0000.0000.0002
Overlay interface Overlay1
 VPN name            : Overlay1
 VPN state           : UP
 Extended vlans      : 5-10 (Total:6)
 Join interface(s)   : E1/2 (10.2.2.2)
 Site vlan           : 15 (up)
 AED-Capable         : Yes
 Capability          : Unicast-Only
 Is Adjacency Server : Yes
 Adjacency Server(s) : 10.1.1.1 / [None]
```

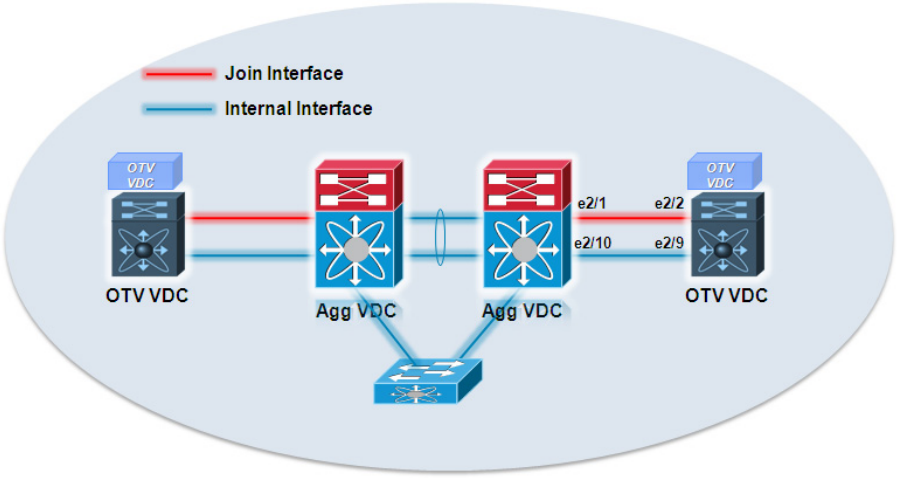
Generic OTV Edge Device

```
Generic_OTV# show otv overlay 1
OTV Overlay Information
Site Identifier 0000.0000.0003
Overlay interface Overlay1
 VPN name            : Overlay1
 VPN state           : UP
 Extended vlans      : 5-10 (Total:6)
 Join interface(s)   : E1/1 (10.3.3.3)
 Site vlan           : 15 (up)
 AED-Capable         : Yes
 Capability          : Unicast-Only
 Is Adjacency Server : No
 Adjacency Server(s) : 10.1.1.1 / 10.2.2.2
```

Configuring OTV Multi-Homing

Once the first OTV edge device configuration is completed, it is recommended to enable OTV on a second local OTV edge device in order to provide a more resilient LAN extension solution (Figure 1-57).

Figure 1-57 OTV Multi-Homing Configuration



The configuration steps 1-4 listed in the previous section apply identically for the configuration of the second OTV edge device. However, before activating the internal interface ("no shut" on e2/9 in the example above) it is recommended to enable the Join interface ("no shut" on e2/2) and the OTV Overlay interface. Once it is verified that the edge device discovered the other OTV edge devices and established control plane adjacencies via the Overlay interface, it is possible to enable the internal interface. This would cause the two local OTV edge devices to discover each other via the site VLAN and become authoritative for a subset of VLANs (one device will be the AED for the odd VLANs, the other one for the even VLANs). The establishment of internal control plane adjacencies can be verified using the **show otv site** CLI command, as highlighted below.

```
East-a# sh otv site
Site Adjacency Information (Site-VLAN: 15) (* - this device)
Overlay1 Site-Local Adjacencies (Count: 2)

  Hostname                         System-ID         Ordinal
  -------------------------------  ----------------  -------
  OTV-VDC-B                        001b.54c2.3dc2          0
* OTV-VDC-A                        0022.5579.36c2          1
```

The fact that the new edge device will become Authoritative after already establishing OTV adjacencies on the overlay interface with the remote edge devices ensure that traffic disruption for these VLANs will be minimized.

If running NX-OS release 5.2(1) or later, the concept of dual site adjacencies is introduced, which ensure that a device become a candidate for the AED role only when it is fully ready to perform the LAN extension functionality. Because of this multi-homing hardening capability, it is less important to closely follow the procedure describe above when deploying a multi-homed solution.

The following output highlights the establishment of dual internal adjacencies:

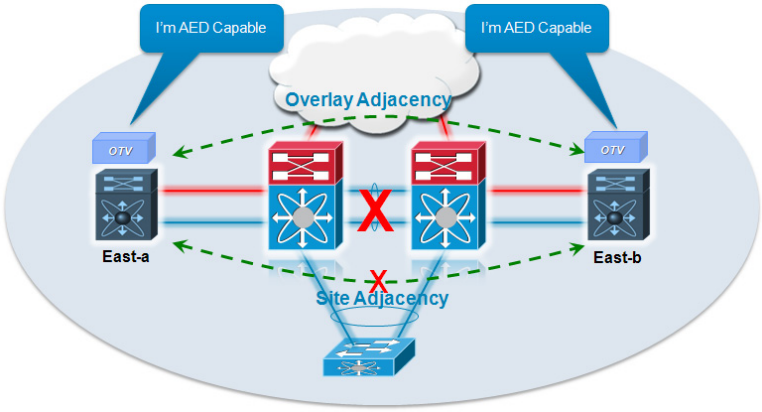
```
East-a# sh otv site

Dual Adjacency State Description
    Full     - Both site and overlay adjacency up
    Partial  - Either site/overlay adjacency down
    Down     - Both adjacencies are down (Neighbor is down/unreachable)
    (!)      - Site-ID mismatch detected

Local Edge Device Information:
    Hostname East-a
    System-ID 0022.5579.7c42
    Site-Identifier 0000.0000.0002
    Site-VLAN 15 State is Up

Site Information for Overlay1:
Local device is AED-Capable
Neighbor Edge Devices in Site: 1
Hostname                System-ID        Adjacency-       Adjacency-   AED-
                                         State            Uptime       Capable
--------------------------------------------------------------------------------
East-b                  0022.5579.b842   Full              6w4d         Yes
```

Since both OTV Edge devices are up in the East site, they will divide VLANs between them as shown below. Notice the "*" sign in front of VLANs which indicates that the particular device in question is AED for those VLANs:

```
East-a# show otv vlan
OTV Extended VLANs and Edge Device State Information (* - AED)
VLAN   Auth. Edge Device                    Vlan State       Overlay
----   -----------------------------------  ----------       -------
 5*    East-a                               active           Overlay1
 6     East-b                               inactive(Non AED)Overlay1
 7*    East-a                               active           Overlay1
 8     East-b                               inactive(Non AED)Overlay1
 9*    East-a                               active           Overlay1
 10    East-b                               inactive(Non AED)Overlay1
```

Now consider a case when proper communication between two OTV edge devices does not work over the site VLAN. Figure 1-58 examplifies a case where the connection between aggregation layer devices fails. This means that both OTV devices cannot discover each other over the site VLAN anymore and hence the Site Adjacency goes down. Without the support for the Overlay Adjacency both OTV device in this scenario would become AED for all VLANs potentially causing a loop.

Figure 1-58 Site Adjacency is Down but Overlay Adjacency is Up



Notice the Adjacency state below. Partial state refers to a scenario where either one of the adjacency is up. However, it is worth noticing that both OTV Edge Devices are still advertising themselves as AED capable, since they still have connectivity to the L2 and L3 network domains. As long as neighbors are AED Capable and have at least one adjacency up, they will continue to be part of the AED election and divide VLANs between them for active-active load balancing.

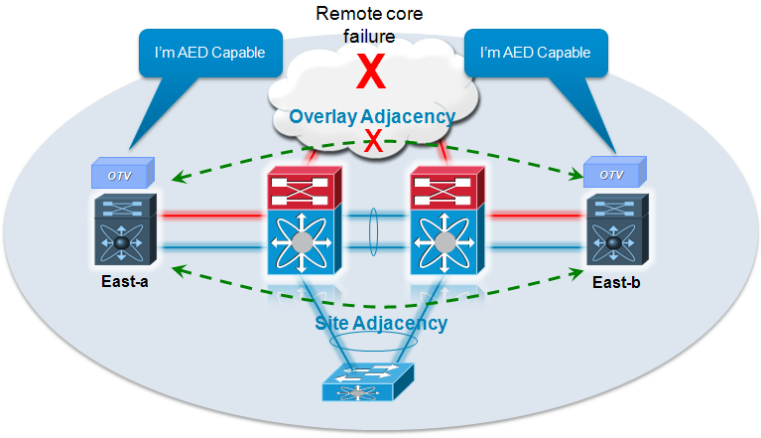
```
East-a# sh otv site
Dual Adjacency State Description
    Full      - Both site and overlay adjacency up
    Partial   - Either site/overlay adjacency down
    Down      - Both adjacencies are down (Neighbor is down/unreachable)
    (!)       - Site-ID mismatch detected

Local Edge Device Information:
    Hostname East-a
    System-ID 0022.5579.7c42
    Site-Identifier 0000.0000.0002
    Site-VLAN 15 State is Up

Site Information for Overlay1:

Local device is AED-Capable
Neighbor Edge Devices in Site: 1

Hostname                System-ID        Adjacency-       Adjacency-   AED-
                                         State            Uptime       Capable
------------------------------------------------------------------------------
East-b                  0022.5579.b842   Partial  (!)      6w4d         Yes
```

As depicted in the output below, East-a and East-b continue to be AED for odd and even VLANs respectively.

```
East-a# show otv vlan
OTV Extended VLANs and Edge Device State Information (* - AED)
VLAN   Auth. Edge Device                    Vlan State       Overlay
----   -----------------------------------  ----------       -------
 5*    East-a                               active           Overlay1
 6     East-b                               inactive(Non AED)Overlay1
 7*    East-a                               active           Overlay1
 8     East-b                               inactive(Non AED)Overlay1
 9*    East-a                               active           Overlay1
 10    East-b                               inactive(Non AED)Overlay1
```

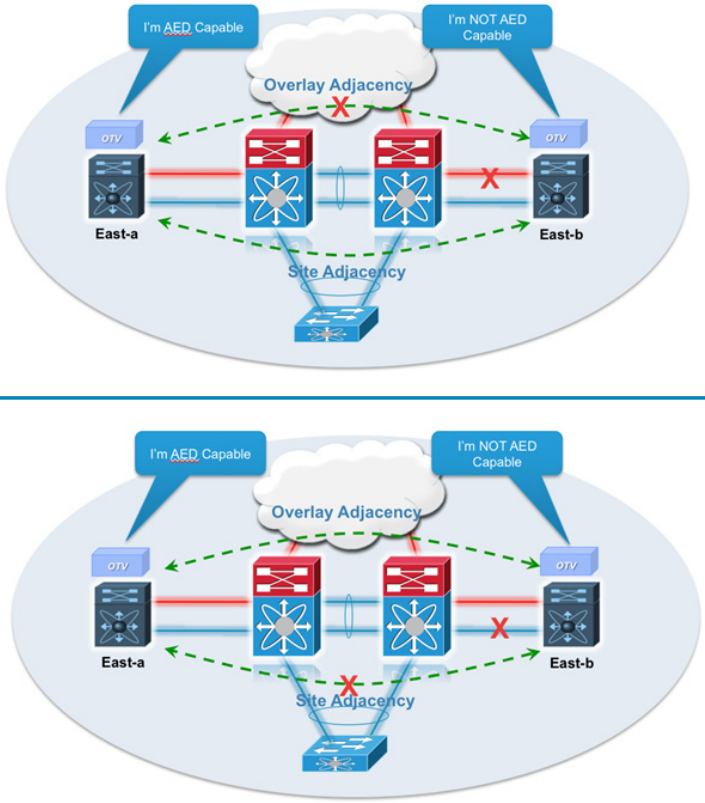
A similar behavior is experienced when the Overlay Adjacency fails because of a remote failure in the L3 core, while the Site Adjacency remains available. Figure 1-59 shows how both OTV devices remain AED capable also in this scenario, providing LAN extension services to their set of odd or even VLANs.

Figure 1-59 Site Adjacency is Down but Overlay Adjacency is Up



Let's finally consider the case where one of the OTV devices undergoes an actual failure (for example its Join or Internal Interface, as shown Figure 1-60), which renders him incapable of forwarding OTV traffic. In this type of scenario the OTV device would send a notification to its neighbor about its local failure and inform it's forwarding readiness (or lack thereof). Following the AED notification, the peer OTV device would trigger an AED election process and exclude the OTV device that is not AED capable anymore.

Figure 1-60 The Join or Internal Interface of the OTV Device Fails, Rendering It Not AED Capable



In the exhibit below, East-b is not AED capable anymore. Please note that even though adjacency state between East-a and East-b is "Partial" as in the previous case (since the Site Adjacency is still established), East-b is not considered for AED role anymore since it is not AED capable.

```
East-a# sh otv site
Dual Adjacency State Description
    Full     - Both site and overlay adjacency up
    Partial  - Either site/overlay adjacency down
    Down     - Both adjacencies are down (Neighbor is down/unreachable)
    (!)      - Site-ID mismatch detected

Local Edge Device Information:
    Hostname East-a
    System-ID 0022.5579.7c42
    Site-Identifier 0000.0000.0002
    Site-VLAN 15 State is Up

Site Information for Overlay0:

Local device is AED-Capable
Neighbor Edge Devices in Site: 1

Hostname                System-ID        Adjacency-       Adjacency-   AED-
                                         State            Uptime       Capable
------------------------------------------------------------------------------
East-b                  0022.5579.b842   Partial             6w4d         No
```

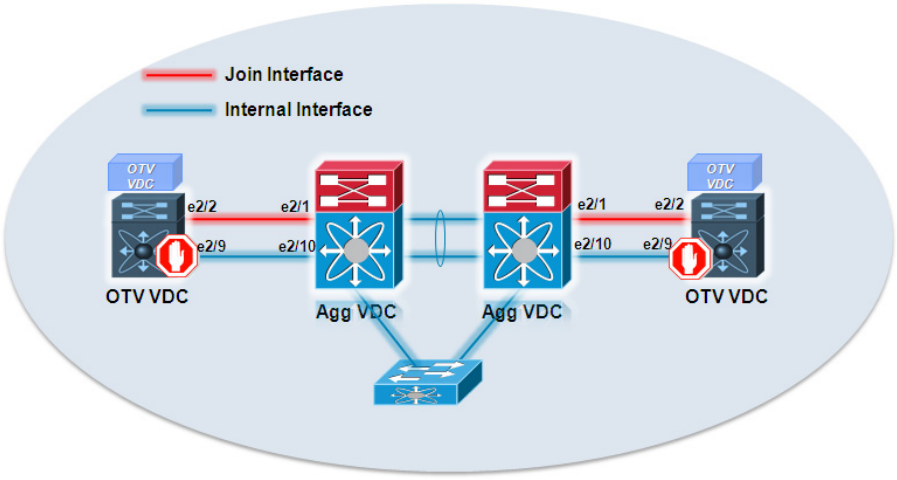
The output below highlights how East-a has now become AED for all VLANs as East-b is not AED capable anymore.

```
East-a# show otv vlan
OTV Extended VLANs and Edge Device State Information (* - AED)
VLAN   Auth. Edge Device                    Vlan State       Overlay
----   -----------------------------------  ----------       -------
 5*    East-a                               active           Overlay1
 6*    East-a                               active           Overlay1
 7*    East-a                               active           Overlay1
 8*    East-a                               active           Overlay1
 9*    East-a                               active           Overlay1
 10*   East-a                               active           Overlay1
```

Configuring FHRP Isolation

As mentioned in the "FHRP Isolation" section on page 1-28, filtering FHRP messages across the OTV Overlay allows to provide the same active default gateway in each data center site . This concept is highlighted in Figure 1-61.

Figure 1-61 FHRP Isolation



In the future, OTV will offer a simple command to enable these filtering capabilities. In the meantime, it is possible to deploy the configuration highlighted above to achieve the same purpose.

To enable FHRP filtering, the following two steps are required (the configuration samples refer to the diagram in Figure 1-57):

Step 1 Configure a VLAN ACL (VACL) on the OTV VDC.

The VLAN ACL is required to identify the traffic that needs to be filtered. The configuration below applies to the HSRP version 1 protocol, which is enabled by default on NX-OS and in bold are highlighted the specific commands required to filter HSRP version 2 packets (similar configuration can be created for VRRP). Notice how it is also required to apply a specific filter to ensure suppression of the Gratuitous ARP (GARP) messages that may be received across the OTV Overlay from the remote sites. This can be achieved leveraging the **ip arp inspection filter** command.

```
ip access-list ALL_IPs
  10 permit ip any any
!
mac access-list ALL_MACs
  10 permit any any
!
ip access-list HSRP_IP
```

```
  10 permit udp any 224.0.0.2/32 eq 1985
  20 permit udp any 224.0.0.102/32 eq 1985
!
mac access-list HSRP_VMAC
  10 permit 0000.0c07.ac00 0000.0000.00ff any
  20 permit 0000.0c9f.f000 0000.0000.0fff any
!
arp access-list HSRP_VMAC_ARP
  10 deny ip any mac 0000.0c07.ac00 ffff.ffff.ff00
  20 deny ip any mac 0000.0c9f.f000 ffff.ffff.f000
  30 permit ip any mac any
vlan access-map HSRP_Localization 10
        match mac address HSRP_VMAC
        match ip address HSRP_IP
        action drop

vlan access-map HSRP_Localization 20
        match mac address ALL_MACs
        match ip address ALL_IPs
        action forward
!
feature dhcp
ip arp inspection filter HSRP_VMAC_ARP <OTV_Extended_VLANs>
vlan filter HSRP_Localization vlan-list <OTV_Extended_VLANs>
```

After applying the configuration above to the set of VLANs that are trunked from the Agg VDC to the OTV VDC, all HSRP messages will be dropped once received by the OTV VDC.

Step 2 Apply a route-map to the OTV control protocol (IS-IS).

Even though HSRP traffic is filtered via the VACL defined in the step above, the vMAC used to source the HSRP packets is still learned by the OTV VDC. Therefore, OTV advertises this MAC address information to the other sites via an IS-IS update. While this in itself is not causing harm, it would cause the remote OTV the edge devices to see constant MAC moves happening for the vMAC (from the internal interface to the overlay interface and vice versa). To prevent these MAC moves from being advertised and allow for a cleaner design, the following OTV route-map has to be configured (once again, this configuration applies to HSRP version 1 and 2).

```
mac-list OTV_HSRP_VMAC_deny seq 10 deny 0000.0c07.ac00 ffff.ffff.ff00
mac-list OTV_HSRP_VMAC_deny seq 11 deny 0000.0c9f.f000 ffff.ffff.f000
mac-list OTV_HSRP_VMAC_deny seq 20 permit 0000.0000.0000 0000.0000.0000
!
route-map OTV_HSRP_filter permit 10
  match mac-list OTV_HSRP_VMAC_deny
!
otv-isis default
  vpn Overlay0
    redistribute filter route-map OTV_HSRP_filter
```

Summary

This document introduced a Cisco innovative LAN extension technology called Overlay Transport Virtualization (OTV). OTV is a new feature of the Nexus OS operating system that allows Ethernet traffic from a local area network (LAN) to be tunneled over an IP network to create a "logical data center" spanning several data centers in different locations, solving many of challenges that make it difficult to shift large workloads between facilities, potentially opening new frontiers in disaster recovery, data center consolidation, and energy management.

The primary OTV characteristics discussed are the following:

- Capability of extending Layer 2 LANs over any network by leveraging IP-encapsulated MAC routing.
- Simplification of configuration and operation by enabling seamless deployment over existing network without redesign, requiring minimal configuration commands and providing single-touch site configuration for adding new data centers.
- Increasing resiliency by preserving existing Layer 3 failure boundaries, providing automated multi-homing, and including built-in loop prevention.
- Maximizing available bandwidth by using equal-cost multipath and optimal multicast replication (in deployments where the transport infrastructure is multicast enabled).