

Overlay Transport Virtualization for Geographically Dispersed Virtual Data Centers: Improve Application Availability and Portability

What You Will Learn

Geographically dispersed data centers provide added application resiliency and workload allocation flexibility. To this end, the network must provide Layer 2, Layer 3 and storage connectivity between data centers. Connectivity must be provided without compromising the autonomy of data centers or the stability of the overall network.

This document summarizes the attributes of the Cisco® Overlay Transport Virtualization (OTV) technology and the way it provides Layer 2 connectivity across data centers that:

- Is nondisruptive (transparent to the core and sites)
- Is transport agnostic
- Is multihomed and multipathed
- Preserves failure isolation between data centers

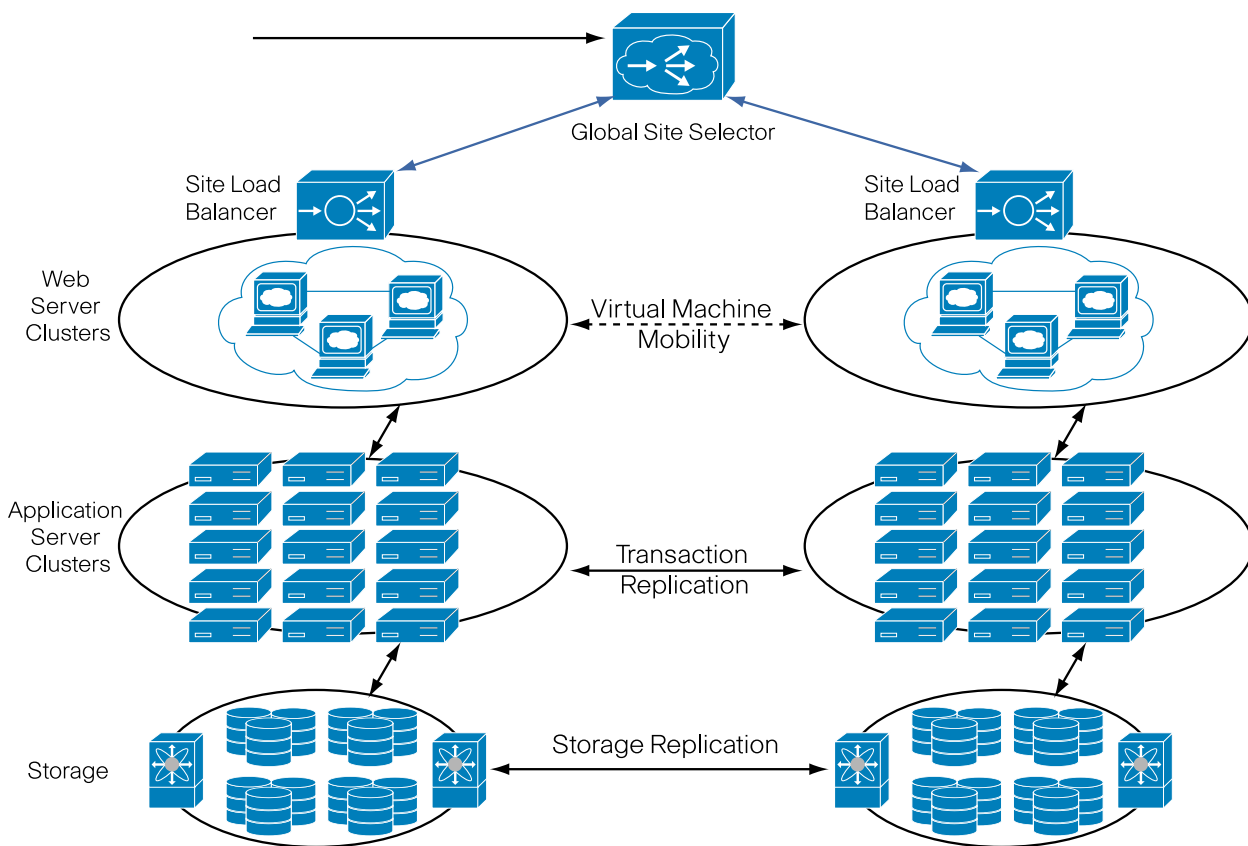
Technology decision makers, IT managers, and network architects will find this document useful.

Challenge

Businesses face the challenge of providing very high availability for applications while keeping operating expenses low. Applications must be available any time and anywhere with optimal response times.

The deployment of geographically dispersed data centers allows the IT designer to put in place effective disaster-avoidance and disaster-recovery mechanisms that increase the availability of the applications. Geographic dispersion also enables optimization of application response through improved facility placement and allows flexible mobility of workloads across data centers to avoid demand hotspots and fully utilize available capacity.

To enable all the benefits of geographically dispersed data centers, the network must extend Layer 2 connectivity across the diverse locations. As shown in Figure 1, LAN extensions may be required at different layers in the data center to enable the resiliency and clustering mechanisms offered by the different applications at the web, application, and database layers. The figure also shows the Layer 3 and storage connectivity requirements; this document focuses on the Layer 2 connectivity requirements.

Figure 1. Geographically Dispersed Application Clusters

Existing mechanisms for the extension of Layer 2 connectivity are less than optimal in addressing connectivity and independence requirements and present many challenges and limitations that OTV effectively overcomes. Some of the challenges include:

- **Fate sharing:** The extension of Layer 2 domains across multiple data centers can cause the data centers to share failures that would normally have been isolated when interconnecting data centers over an IP network. These failures propagate freely over the open Layer 2 flood domain. A solution that provides Layer 2 connectivity yet restricts the reach of the flood domain is needed to contain failures and preserve the resiliency achieved through the use of multiple data centers.
- **Complex operations:** Layer 2 VPNs can provide extended Layer 2 connectivity across data centers, but will usually involve a mix of complex protocols, distributed provisioning, and an operationally intensive hierarchical scaling model. A simple overlay protocol with built-in capabilities and point-to-cloud provisioning is crucial to reducing the cost of providing this connectivity.
- **Bandwidth utilization:** When extending Layer 2 domains across data centers, the use of available bandwidth between data centers must be optimized to obtain the highest connectivity at the lowest cost. Balancing the load across all available paths while providing resilient connectivity between the data center and the transport network requires added intelligence above and beyond that available in traditional Ethernet switching and Layer 2 VPNs.
- **Transport independence:** The nature of the transport between data centers varies depending on the location of the data centers and the availability and cost of services in the different areas. A cost-effective solution for the interconnection of data centers must be transport agnostic and give the network designer the flexibility to choose any transport between data centers based on business and operational preferences.

Business Benefits

OTV provides an operationally optimized solution for the extension of Layer 2 connectivity across any transport. OTV is therefore critical to the effective deployment of distributed data centers to support application availability and flexible workload mobility.

OTV enhances the data center high-availability model and brings the data center closer to 99.999 percent high availability. The Layer 2 connectivity provided by OTV coupled with the scoping of failure domains enables nondisruptive cluster-based disaster-recovery schemes as well as stateful workload mobility instrumental in providing transparent disaster avoidance. The capability to move workloads transparently also allows the insertion of maintenance and change management processes that do not require any downtime. The average cost of downtime in the data center has been estimated at US\$42,000 per hour of downtime. The average downtime per year has been estimated at 87 hours per year, which is equivalent to 99.0 percent availability and an average yearly cost of US\$3.65 million. Reducing downtime by providing effective mechanisms for disaster handling as well as transparent change management can bring data center availability from 99.0 to 99.999 percent. This means that the yearly downtime of the data center can be decreased from 87 hours to 5 minutes, providing savings of US\$3.65 million per year to the average operation.

Depending on the nature of the business, the cost of downtime may be much higher. For instance, some well-known online merchant services handle an average of US\$2000 per second in merchant transactions, which translates to a cost of downtime of US\$7.2 million per hour. Clearly, anything less than 99.999 percent availability is not acceptable for such businesses, which will incur a loss of approximately US\$600,000 per year at 99.999 percent availability compared to US\$626.4 million per year at 99.0 percent availability.

Whether the cost of downtime to the business is within the average or at the extreme end described in the preceding example, the return on investment (ROI) for high-availability optimizations in the data center is very attractive. For instance, consider an average data center with two sites and for which the network infrastructure has been upgraded to provide OTV services at a cost of US\$1 million. The ROI based on the standard ROI formula would be 2.65. In the extreme case, the same US \$1 million would yield a ROI of 625.4.

$$\text{ROI} = \frac{(\text{Gain from Investment} - \text{Cost of Investment})}{\text{Cost of Investment}}$$

Although the ROI based on downtime data may be attractive enough to justify the investment in OTV, there are other business benefits to the transparent extension of LANs across geographic locations that make the ROI look even better. One such benefit is the capability to optimally balance the load across available resources over multiple data centers. This level of flexibility allows the provisioning of smaller consolidated data centers and the optimization of power and cooling resource utilization by avoiding the formation of processing or temperature hotspots. An intelligent reallocation of the workload can translate into considerable savings in power and cooling costs.

Other benefits are harder to quantify: for instance, the productivity (and sometimes competitive) gains achieved by reducing the response time of critical applications. The geographic dispersion of data centers can reduce application response time simply through the physical proximity of the servicing data center.

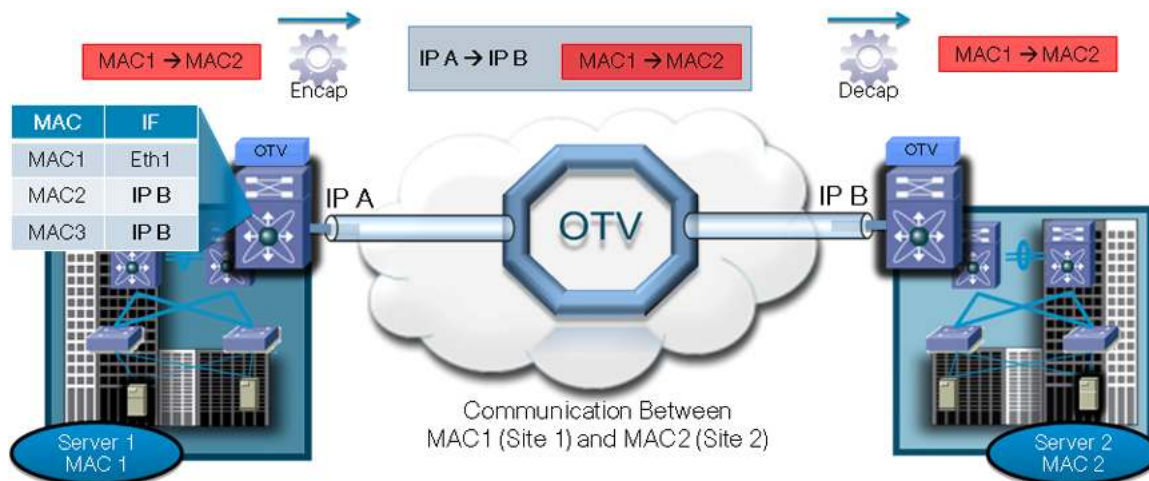
Solution

OTV is a “MAC address in IP” technique for supporting Layer 2 VPNs to extend LANs over any transport. The transport can be Layer 2 based, Layer 3 based, IP switched, label switched, or anything else as long as it can carry IP packets. By using the principles of MAC routing, 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.

The core principles on which OTV operates are the use of a control protocol to advertise MAC address reachability information (instead of using data plane learning) and packet switching of IP encapsulated Layer 2 traffic (instead of using circuit switching) for data forwarding. These features are a significant departure from the core mechanics of traditional Layer 2 VPNs. In traditional Layer 2 VPNs, a static mesh of circuits is maintained among all devices in the VPN to enable flooding of traffic and source-based learning of MAC addresses. This full mesh of circuits is an unrestricted flood domain on which all traffic is forwarded. Maintaining this full mesh of circuits severely limits the scalability of existing Layer 2 VPN approaches. At the same time, the lack of a control plane limits the extensibility of current Layer 2 VPN solutions to properly address the requirements for extending LANs across data centers.

OTV uses a control protocol to map MAC address destinations to IP next hops that are reachable through the network core. OTV can be thought of as MAC routing in which the destination is a MAC address, the next hop is an IP address, and traffic is encapsulated in IP so it can simply be carried to its MAC routing next hop over the core IP network. Thus a flow between source and destination host MAC addresses is translated in the overlay into an IP flow between the source and destination IP addresses of the relevant edge devices. This process is called encapsulation rather than tunneling as the encapsulation is imposed dynamically and tunnels are not maintained. Since traffic is IP forwarded, OTV is as efficient as the core IP network and will deliver optimal traffic load balancing, multicast traffic replication, and fast failover just like the core would. Figure 2 illustrates this dynamic encapsulation mechanism.

Figure 2. How OTV Works



The mappings of the MAC address to the IP next hop in the forwarding table illustrated in Figure 2 are advertised by a control protocol, thus eliminating the need for flooding of unknown unicast traffic across the overlay. The control protocol is extensible and includes useful MAC-address-specific information such as VLAN, site ID, and associated IP address (for IP hosts). This rich information, most of which is not available when you rely on data flood learning, is instrumental in building into OTV the necessary intelligence to implement multihoming, load balance, prevent loops, localize First Hop Resiliency Protocol (FHRP) capability, and even localize Address Resolution Protocol (ARP) traffic without creating additional operational overhead for each function.

Thus, OTV can be used to provide connectivity based on MAC-address destinations while preserving most of the characteristics of a Layer 3 interconnection. Some of the main benefits achieved with OTV include:

- No effect on existing network design
 - Transport agnostic: OTV is IP encapsulated and can therefore use any core capable of forwarding IP traffic. OTV therefore does not pose any requirements on the core transport.

- Transparent to the sites: OTV extensions do not affect the design or protocols of the Layer 2 sites that OTV interconnects. It is as transparent as a router connection to the Layer 2 domain and therefore does not affect the local spanning tree or topology.
- Failure isolation and site independence
 - Failure boundary preservation and site independence preservation: OTV does not rely on traffic flooding to propagate reachability information for MAC addresses. Instead, a control protocol is used to distribute such information. Thus, flooding of unknown traffic is suppressed on the OTV overlay; ARP traffic is forwarded only in a controlled manner, and broadcasts can be forwarded based on specific policies. Spanning tree Bridge Protocol Data Units (BPDUs) are not forwarded at all on the overlay. The result is failure containment comparable to that achieved using a Layer 3 boundary at the Layer 2 domain edge. Sites remain independent of each other, and failures do not propagate beyond the OTV edge device.
- Optimized operations
 - Single-touch site additions and removals: OTV is an overlay solution that needs to be deployed only at specific edge devices. Edge devices join a signaling group in a point-to-cloud fashion. Therefore, the addition or removal of an edge device does not affect or involve configuration of other edge devices on the overlay.
 - Familiar and succinct command-line interface (CLI): Configuring an OTV overlay is as simple as configuring an interface on a switch and requires very few steps (as few as three commands on each device).
 - Single protocol with no add-ons: Before OTV, multihoming, loop prevention, load balancing, multipathing, etc. required the addition of protocols to solutions: for instance, virtual private LAN service (VPLS). With OTV, all capabilities are included in a single control protocol and single configuration.
- Optimal bandwidth utilization resiliency and scalability
 - Multipathing (cross-sectional bandwidth and end-to-end Layer 2 multipathing): When multihoming sites, OTV provides the capability to actively use multiple paths over multiple edge devices. This feature is crucial to keeping all edge devices active and thus optimizing the use of available bandwidth. The capability to use multipathing is also critical to supporting end-to-end multipathing when the Layer 2 sites are using Layer 2 multipathing locally. OTV is the only Layer 2 extension protocol that can provide transparent multipathing integration when virtual PortChannel (vPC) or Transparent Interconnection of Lots of Links (TRILL) technology is used at the Layer 2 sites being interconnected over an IP cloud. Since OTV is IP encapsulated, OTV also uses any multipathing available in the underlying IP core.
 - Multipoint connectivity: OTV provides multipoint connectivity in an easy-to-manage point-to-cloud model. OTV uses the IP Multicast capabilities of the core to provide optimal multicast traffic replication to multiple sites and avoid head-end replication that leads to suboptimal bandwidth utilization.
 - Transparent multihoming with built-in loop prevention: Multihoming does not require additional configuration because it is built into OTV along with all the necessary loop detection and suppression mechanisms. The loop prevention mechanisms in OTV prevent loops from forming on the overlay, and also prevent loops from being induced by sites when these are multihomed to the overlay. In the past, it was necessary to use complex extensions of spanning tree over Layer 2 VPNs or reduce the multihoming of the VPN to an active-standby model. With OTV, all edge devices are active while loops are prevented inherently in the protocol.
 - Fast failover: Since OTV uses a control protocol based on an Interior Gateway Protocol (IGP), OTV benefits from the fast failover characteristics of equal-cost multipath (ECMP). Additionally, OTV can use bidirectional forwarding detection (BFD) extensions to provide failover between edge devices in less than 1 second.

- Scalability
 - Optimized and distributed state: OTV does not create nailed up tunnels; the only state maintained is that of a MAC-address routing table. As in any packet-switched model, no hard state is maintained, and therefore scalability is much better than that achieved in circuit-switched models. State is distributed and can be programmed in the hardware conditionally to allow larger numbers of MAC addresses to be handled by the overlay.
 - Flexibility: The level of scalability possible with OTV is designed to allow the network architect to deploy OTV capabilities at the data center aggregation layer if necessary. Deployment of circuit-switched solutions at the aggregation layer can be prevented by the scalability limitations of the solutions because the number of aggregation devices is usually much greater than the number of endpoints that a circuit-switched solution can support with its full mesh of circuits. Having the flexibility to deploy the required LAN extensions at any layer can significantly improve the operational model of the data center.
 - Reduced effect of ARP traffic: Scalability is improved by the capability of the OTV control plane to localize ARP traffic and therefore reduce the effect of ARP traffic on all hosts present on the extended LAN.
- Transparent migration path
 - Incremental deployment: Since OTV is agnostic to the core and transparent to the sites, it can be incrementally deployed over any existing topology without altering its network design. When vPC is being used as a data center interconnect (DCI) and there is a desire to migrate to OTV, OTV can simply be deployed over the existing vPC interconnect to bring the benefits of failure containment and improved scale. Alternatively, OTV can use the main IP core between data centers as its transport, and the vPC interconnect can be taken out of service later. In the latter scenario, VLANs can be migrated in phases from the vPC interconnect to the OTV overlay.

Why Cisco?

With OTV, Cisco is again bringing innovation and leading the industry with the introduction of next-generation technology that shapes the standards of the future. OTV is the result of years of experience at Cisco in interconnecting data centers and providing Layer 2 and 3 technologies. Cisco has developed a solution that is custom built to meet the data center challenges and is aligned with the broader set of data center innovations that will be changing data center networking in the next few years. OTV is aligned with technologies like Layer 2 multipathing and TRILL to provide elastic and resilient Layer 2 domains and a coherent and well-designed feature set that supports the data center of the future. Rather than retrofitting technologies designed for carrier applications, OTV focuses on the main operational challenges and required efficiencies in the data center to provide a lean and comprehensive solution to the LAN extension challenge experienced in the data center.

For More Information

For more information about Cisco OTV, visit <http://www.cisco.com/go/nexus7000>



Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

CCDE, CCENT, CCSI, Cisco Eos, Cisco HealthPresence, Cisco IronPort, the Cisco logo, Cisco Nurse Connect, Cisco Pulse, Cisco SensorBase, Cisco StackPower, Cisco StadiumVision, Cisco TelePresence, Cisco Unified Computing System, Cisco WebEx, DCE, Flip Channels, Flip for Good, Flip Mino, Flipshare (Design), Flip Ultra, Flip Video, Flip Video (Design), Instant Broadband, and Welcome to the Human Network are trademarks; Changing the Way We Work, Live, Play, and Learn, Cisco Capital, Cisco Capital (Design), Cisco:Financed (Stylized), Cisco Store, Flip Gift Card, and One Million Acts of Green are service marks; and Access Registrar, Aironet, AllTouch, AsyncOS, Bringing the Meeting To You, Catalyst, CCDA, CCDR, CCIE, CCIP, CCNA, CCNP, CCSP, CCVP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Lumin, Cisco Nexus, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Collaboration Without Limitation, Continuum, EtherFast, EtherSwitch, Event Center, Explorer, Follow Me Browsing, GainMaker, iLYNX, IOS, iPhone, IronPort, the IronPort logo, Laser Link, LightStream, Linksys, MeetingPlace, MeetingPlace Chime Sound, MGX, Networkers, Networking Academy, PCNow, PIX, PowerKEY, PowerPanels, PowerTV, PowerTV (Design), PowerVu, Prisma, ProConnect, ROSA, SenderBase, SMARTnet, Spectrum Expert, StackWise, WebEx, and the WebEx logo are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0910R)