Cisco Dynamic Workload Scaling Solution

What You Will Learn

Cisco® Application Control Engine (ACE), along with Cisco Nexus® 7000 Series Switches and VMware vCenter, provides a complete solution for dynamic workload scaling in distributed data center environments.

Today's data centers are evolving from a hierarchical architecture model to a flatter Layer 2 geographically distributed model. One of the main factors behind these changes is the rapid adoption of server virtualization technology in data centers. In a virtualized data center environment, virtual machines can move anywhere in the geographically distributed data center, or they can move to cloud-based data centers hosted by cloud service providers. Geographically dispersed data centers provide added application resiliency and workload allocation flexibility. To support these features, the network must provide Layer 2 connectivity between data centers. Connectivity must be provided without compromising the autonomy of data centers or the stability of the overall network to enable efficient use of the remote computing resources when local resources are scarce.

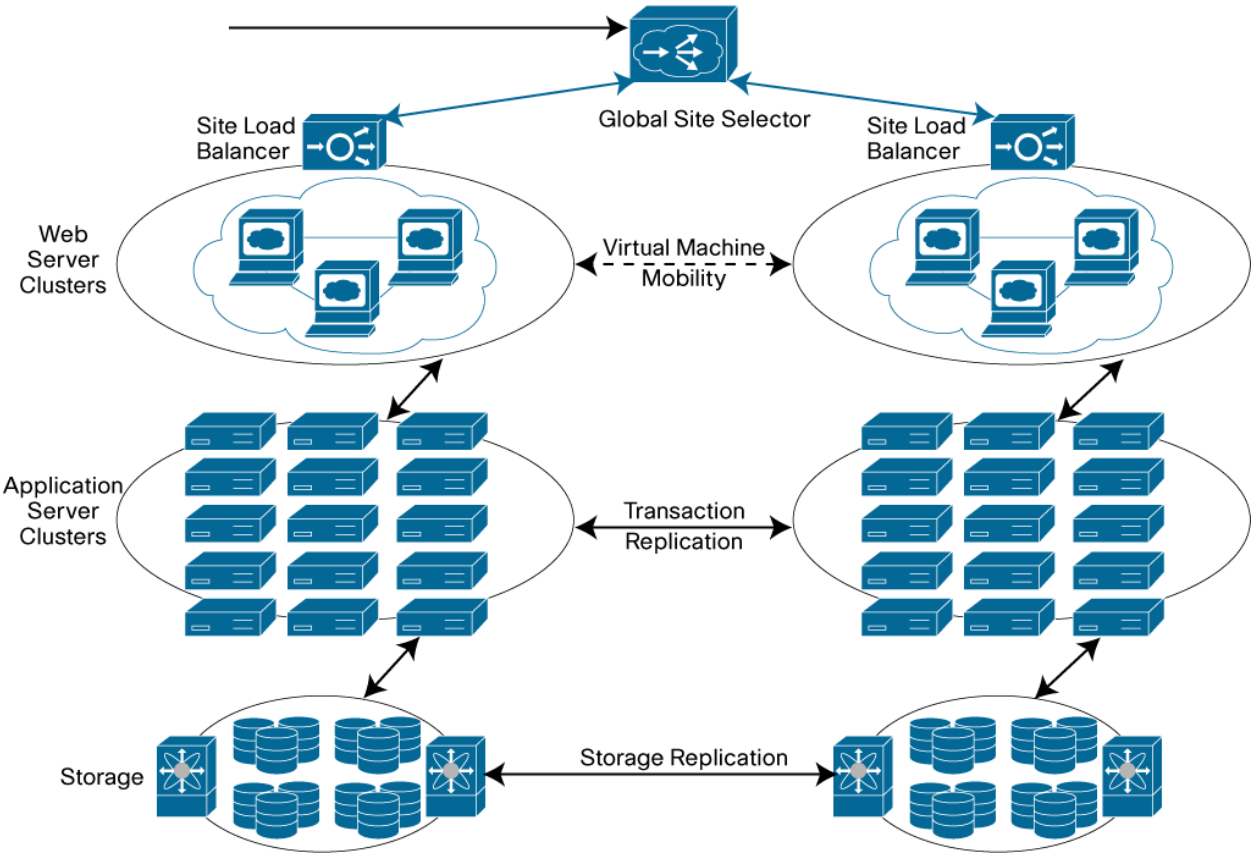
Challenge

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 mechanisms that increase the availability of applications. Geographic dispersion 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 to enable workload mobility between data centers. 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 Cisco Overlay Transport Virtualization (OTV) technology effectively overcomes.

Figure 1. Geographically Dispersed Application Clusters



Cisco Solution

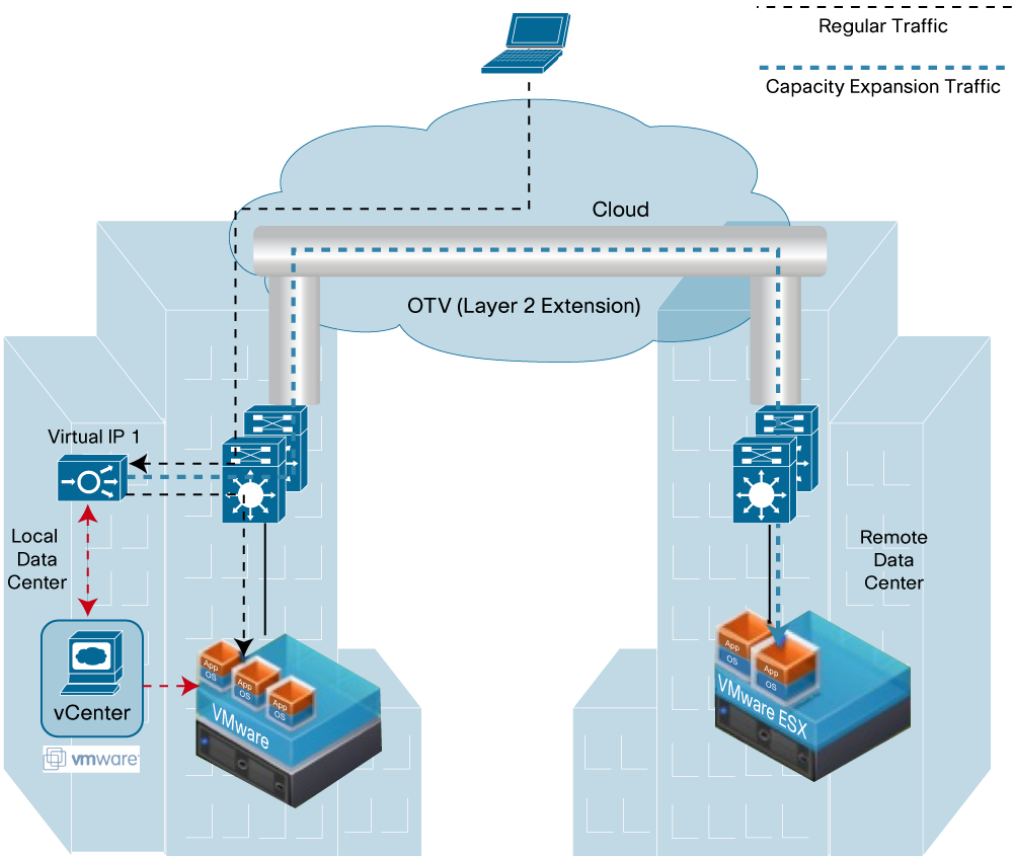
Consider a use case in which an IT department wants to add web servers to provide high availability for applications when the traffic load exceeds a threshold, but the organization does not have more capacity in its on-premises data center. The organization also has several other remote data centers that have enough computing resources that can be used when the on-premises data center is overloaded. Cisco OTV is a new data center interconnect (DCI) technology that can connect multiple geographically distributed data centers over an existing IP network. Using OTV, the IT department in this example can extend its on-premises data center's Layer 2 network to other data centers.

Now assume that the IT department has interconnected its data centers using OTV. The IT administrators of the organization now can clone several virtual machines from a template and move those virtual machines using OTV to other data centers. The cloned virtual machines in the remote data centers are an extension of the on-premises infrastructure and are exposed only to users through a load balancer that provides a virtual IP address. From the viewpoint of Cisco ACE, virtual machines from both on-premises and remote data centers are just a pool of resources that can be used to provide the capacity to scale an application; however, load balancing to remote virtual machines can add latency because of the roundtrip to remote data centers, so the IT department wants to increase the local computing resources before bursting traffic to remote data centers.

Cisco ACE, along with the Cisco Nexus 7000 Series and VMware vCenter, provides a complete solution for dynamic workload scaling for data centers. In this solution, Cisco ACE actively monitors the CPU and memory information of the local virtual machines and computes the average load of the local data center. As shown in Figure 2, during normal operations when the average load is below a preconfigured threshold, Cisco ACE load balances the incoming traffic to only local virtual machines. However, during peak hours, local virtual machines may be overloaded, and additional capacity may be required to service the incoming requests. When the average load of the local data center

crosses a configured threshold, Cisco ACE adds the remote virtual machines to its load-balancing rotation pool, adding more computing resources to service the increased load.

Figure 2. Cisco Dynamic Workload Scaling Solution

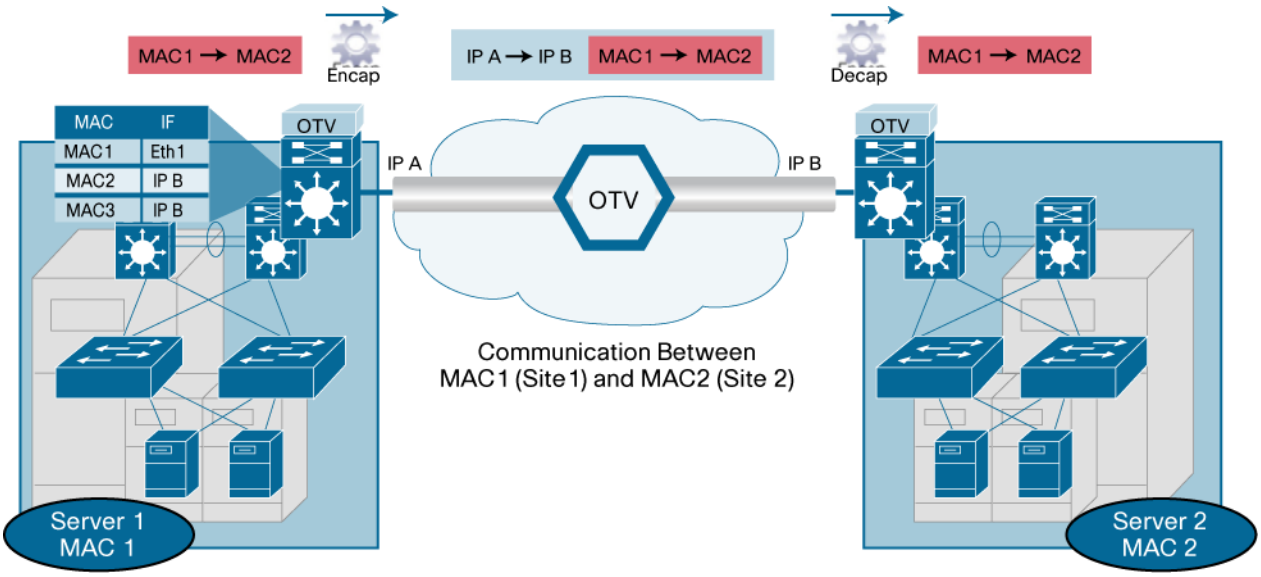


Cisco Solution Design

The Cisco Dynamic Workload Scaling solution integrates Cisco load-balancing technology with VMware virtualization and Cisco OTV technology. Virtualization technology is gaining momentum in enterprise data centers. Enterprises are adopting this technology to optimize the use of computing resources, save costs, and gain operational benefits. Cisco OTV provides an optimized solution for Layer 2 connectivity extension across any transport. Cisco OTV is therefore critical to the effective deployment of distributed data centers to support application availability and flexible workload mobility with virtualization technology. Cisco's load-balancing technology is in the center of the overall solution and ties all the pieces together.

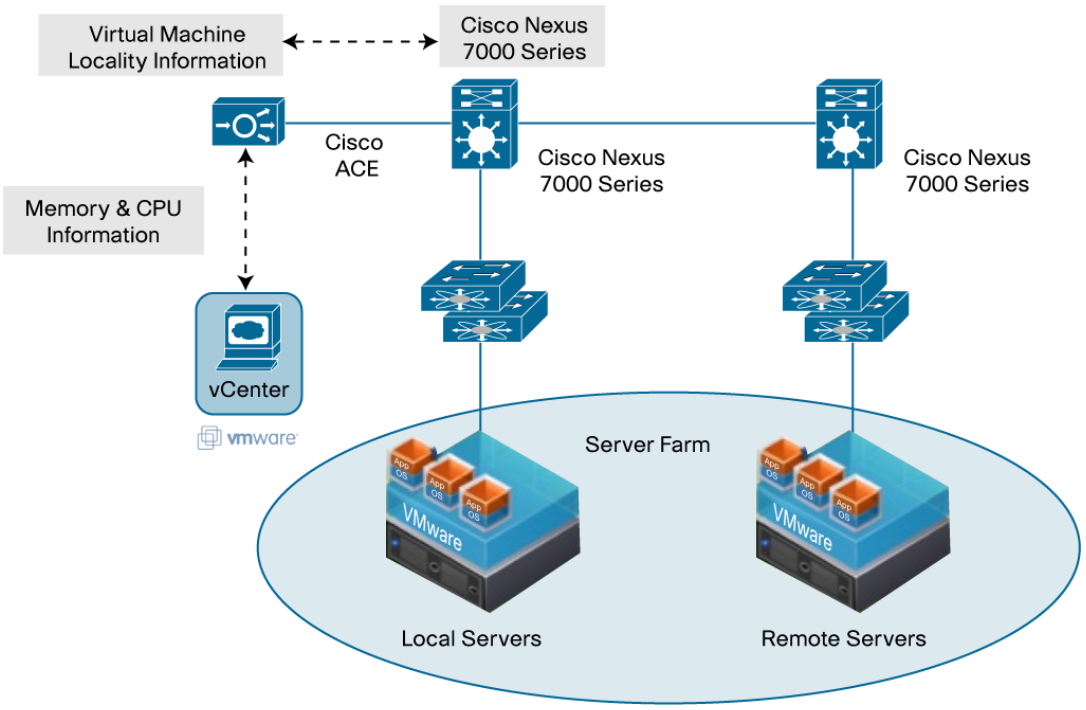
Cisco OTV is a "MAC address in IP" technique for supporting Layer 2 VPNs to extend LANs over any transport. The transport can be Layer 2 based, Layer 3 based, IP switched, label switched, or anything else as long as it can carry IP packets. By using the principles of MAC routing, OTV provides an overlay 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. 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 because the encapsulation is imposed dynamically and tunnels are not maintained. Figure 3 illustrates this dynamic encapsulation mechanism.

Figure 3. How OTV Works



The Cisco ACE load balancer provides the intelligence in the Cisco Dynamic Workload Scaling solution. The load balancer needs to be aware of the locality of the virtual machines to effectively distribute the load between local and remote virtual machines. Enterprises may want to first use the local computing resources before using the computing resources in a remote data center. Therefore, the load balancer needs to be aware of the current resource utilization of the local virtual machines to determine whether to distribute the traffic locally or send it to the remote resources. Cisco ACE has features that integrate Cisco Nexus 7000 Series Switches and VMware vCenter natively on the load balancer. Cisco ACE queries the Cisco Nexus 7000 Series Switch to obtain the virtual machine locality information. Cisco ACE also queries VMware vCenter to obtain CPU and memory utilization data for local virtual machines. With the virtual machine locality and computing resource utilization information, Cisco ACE is equipped to make intelligent load-balancing decisions based on the user configuration. Figure 4 illustrates the mechanism.

Figure 4. How Cisco ACE Obtains Virtual Machine Locality and Resource Utilization Information



What Cisco Offers

With the Cisco Dynamic Workload Scaling solution, Cisco is again bringing innovation and leading the industry with the introduction of next-generation technology that shapes the future. Cisco OTV technology is the result of years of experience in interconnecting data centers and providing Layer 2 and 3 technologies. The Cisco Dynamic Workload Scaling solution is an end-to-end solution 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 coming years

For More Information

- Cisco ACE and Cisco ACE Global Site Selector (GSS) products: http://www.cisco.com/go/ace
- Layer 2 extension between remote data centers:
  - http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/white_paper_c11_493718.html
  - http://www.cisco.com/en/US/netsol/ns975/index.html
- • Cisco OTV technology: http://www.cisco.com/en/US/prod/switches/ps9441/nexus7000_promo.html
- • Cisco Catalyst 6500 Series Switches: http://www.cisco.com/go/6500
- • Cisco Nexus 7000 Series Switches: 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.**

Printed in USA C22-643589-00 1/11