Distributed Virtual Data Center for Enterprise and Service Provider Cloud

Author: Yves Louis – November 2011

I would like to acknowledge Max Ardica, Patrice Bellagamba and Victor Moreno for their significant contributions on the Data Center Interconnect reference architecture described in this paper. Without their great technical expertise this document would not exist!

Table of Contents

Goal of this Document

This paper describes the process and components required to connect the network, storage, and compute resources in distributed data centers into a virtual data center for cloud computing. It emphasizes the Disaster Recovery (DR) and Disaster Avoidance (DA) plans needed to support business continuity.

Audience

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

Introduction

Business resilience and DR are core capabilities of the data center IT infrastructure. The emergence of cloud computing has further highlighted the need for extremely robust network resilience strategies that address security, availability, and virtual machine (VM) mobility while maintaining the flexibility and agility of a cloud model.

One of the main concepts of cloud computing is providing almost unlimited resources for a given service, automatically and dynamically, in a fully-virtual environment. To provide those resources, the complete cloud computing architecture must be built with efficient tools and support business continuity throughout its compute, network, and storage resources.

The challenges of supporting business continuity in a cloud environment are not limited to a physical area or data center. The elasticity and flexibility of the network architecture must be addressed as well. Therefore, the compute, storage, and network components used for cloud computing may not reside in the same physical location. These resources could be spread over multiple locations and interconnected using a transparent transport mechanism that maintains security and end-to-end segmentation.

Four technical services are essential to supporting the high level of flexibility, resource availability, and transparent resource connectivity required for cloud computing:

- The Layer 3 network offers the traditional routed interconnection between remote sites and provides end-user access to cloud services.
- The extended LAN between two or more sites offers transparent transport and supports application and operating system mobility.
- Extended SAN services support data access and accurate data replication.
- IP Localization improves northbound and southbound traffic as well as server-to-server workflows.

Cisco® Data Center Interconnect (DCI) solutions address the business continuity, DR, and DA needs of enterprise and service provider cloud implementations. Cisco DCI reference solutions support the extension of network, storage and compute resources for multiple data centers in multiple locations.

DA and Disaster Prevention (DP) are terms that are often used in the context of cloud computing. DA and DP capabilities provide business continuity without service interruption

or performance impact by manually or automatically moving virtual machines (VMs) to different physical hosts. The destination hosts can be located in the same data center as the source host, or a different physical data center. VM movement is based on available hardware resources.

The concept of interconnecting data centers to provide a DR solution has existed for many years and has been deployed by many enterprises. Before we focus on the emerging DA requirements of cloud computing, let's review the well-known concept of DR and its evolution, as well as the network services associated with it.

Disaster Recovery

Traditional DR Solution for Business Continuity

Traditional data center interconnection architectures have supported DR backup solutions for many years.

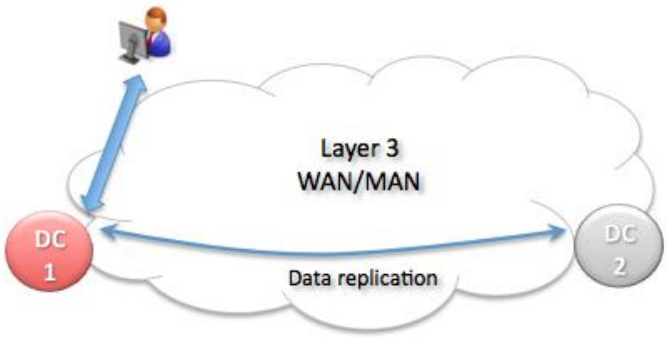
Disaster Recovery Modes

Disaster recovery can be implemented in Cold Standby, Warm Standby, and Hot Standby modes. Each option offers different benefits.

- **Cold Standby:** The initial DR solutions worked in Cold Standby mode, in which appropriately configured backup resources were located in a safe, remote location (Figure 1). Hardware and software components, network access, and data restoration were implemented manually as needed. This DR mode required restarting applications on the backup site, as well as enabling network redirection to the new data center.

  The Cold Standby model is easy to maintain and remains valid. However, it requires a substantial delay to evolve from a standby mode to full operational capability. The time to recover, also known as Recovery Time Objective (RTO), for this scenario can require up to several weeks. In addition, the Recovery Point Objective (RPO), which is the maximum data lost during the recovery process, is quite high. It is accepted that several hours of data might be lost in a Cold Standby scenario.

Figure 1 Cold Standby Mode



- **Warm Standby:** In Warm Standby mode, the applications at the secondary data center are usually ready to start. Resources and services can then be manually activated when the primary data center goes out of service and after traffic is being fully processed to the new location. This solution provides a better RTO and RPO than Cold-Standby mode, but does not offer the transparent operation and zero disruption required for business continuity.
- **Hot Standby:** In Hot Standby mode, the backup data center has some applications running actively and some traffic processing the service tasks. Data replication from the primary data center and the remote data center are done in a real time. Usually the RTO is a few hours and the RPO is zero, which means that the data mirrored in the backup site is exactly the same as in the original site. Zero RPO allows applications and services to restart safely. Immediate and automatic resource availability in the secondary data center improves overall application scalability and equipment utilization.

Data Replication

The different disaster recovery modes are deployed using Layer 3 interconnections between data centers through a highly-available routed WAN. The WAN offers direct access to the applications running in the remote site with synchronous or asynchronous data mirroring, depending on the service level agreement and enterprise business requirements.

The pressure to support business continuity has motivated many storage vendors to accelerate the RTO by offering more efficient data replication solutions that achieve the smallest possible RPO. Examples of highly efficient data replication solutions are host-based and disk-based mirroring.

For example, with Veritas Volume Replicator© host-based mirroring solution, the host is responsible for duplicating data to the remote site. In the EMC Symmetrix Remote Data Facility© (SRDF) and Hitachi (HDS) TrueCopy© disk-based mirroring solutions, the storage controller is responsible for duplicating the data[1].

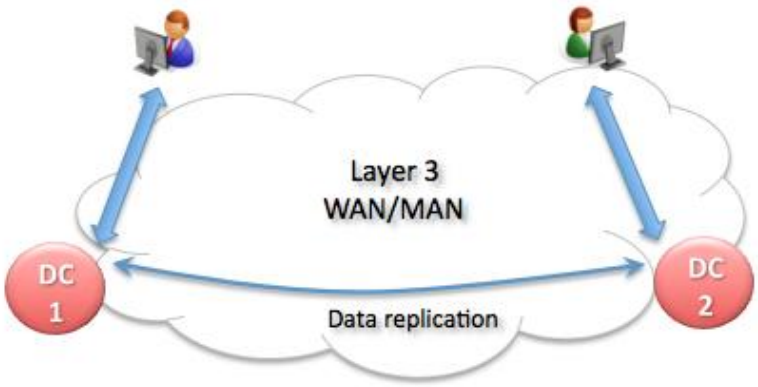
A disaster recovery solution should be selected based on how long the organization can wait for services to be restarted, and above all, how much data it can afford to lose after the failover happens. Should the business restart in a degraded mode, or must all services be fully-available immediately after the switch-over? Financial institutions usually require an RTO of less than one hour with an RPO equal to zero without a degraded mode. This is a fairly widespread practice so that no transactions are lost.

Active-Standby versus Active-Active

A Hot Standby data center can be used for application recovery or to relieve the primary data center from a heavy workload. Relieving data center resources from a heavy workload is usually referred to as Active-Active DR mode (Figure 2).

[1] The storage solution stretching configurations provided in this paper are not an exhaustive list of available solutions. The technology and solution architectures in this area are changing very fast, so we have listed some of the more popular technologies. Several new technologies that merit special attention include the IBM System Storage SAN Volume Controller (SVC)© and the new combination of Hitachi Data Systems (HDS)© with BlueArc© (a recent acquisition).

Figure 2 Active-Active DR Mode



One example of Active/Active DR mode involves an application that is active on a single physical server or VM while the network and compute stack are active in two locations. Some exceptions to this definition are specific software frameworks such as GRID computing, distributed database (i.e. Oracle RAC©) or some cases of server load balancing (SLB[2]). When the resources are spread over multiple locations running in Active-Active mode, some software functions are active in one location and on standby in the other location. Active applications can be located in either site. This approach distributes the workload into several data centers.

It is important to clarify these Active/Active modes by considering the different levels of components, which are all related for the final recovery process:

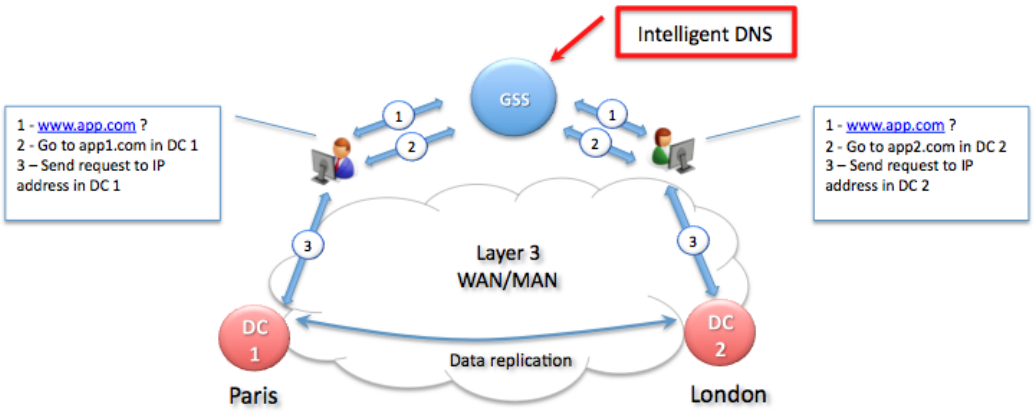
- The network is running on each site interconnecting all compute, network, and security services; and advertising the local subnets outside each data center. Applications active in the remote data centers are therefore accessible from the traditional routed network without changing or restarting any IP processes.
- All physical compute components are up and running with their respective bare metal operating system or hypervisor software stack.
- Storage is replicated on different locations and can be seen as Active/Active for different software frameworks. However, usually a write command for a specific storage volume is sent to one location at a time while the same data is mirrored to the remote location.

Let's have a deeper look at the service itself offered by multiple data centers in Active/Active mode (Figure 3). Assuming that we have application A on data center 1 (i.e. Paris) that offers an e-commerce web portal for a specific set of items. The same e-commerce portal offering the same items can also be available and active on a different location (i.e. London), but with a different IP identifier. For the end user, the service will be unique and the location transparent, but the request can be distributed by the network services based on the proximity criteria established between the end-user and the data center that hosts the same application. So, the same application in this case looks Active/Active, but the software that runs on each compute system is performed autonomously in the front-end tier. They are not

[2] For SLB purposes, traditionally a single active Virtual IP address (VIP) is presented to outside a data center and the server farm supporting this VIP is usually contained within the same Point of Delivery (PoD).

related except from a database point of view. Finally the whole session is maintained at the same servers and in the same location until the session is closed.

Figure 3 Active/Active Mode Services



Global Site Load Balancing Services

To accelerate the disaster recovery service and the dynamic distribution of the workload between the primary and secondary data centers, Cisco provides different network services to optimize the access and the distribution of the user traffic to the remote sites using a Global Site Load Balancing (GSLB) solution. This global GSLB solution for traditional Layer 3 interconnection between sites relies on three major technologies:

- **Intelligent Domain Name System:** An intelligent Domain Name System (DNS) known as the Global Site Selector (GSS) redirects the requests from end-users to the physical location where the application is active and fewer network resources are consumed. In addition, the GSS can be used to distribute traffic across multiple active sites, either in collaboration with the local services of a server load balancing (SLB) application. For example, a Cisco Application Control Engine (ACE) is deployed on each data center to inform the GSS of the health of the service it offers, or based on the load of the network (Figure 4), or in collaboration with the existing WAN edge routers (Figure 5) in the data center (e.g. redirection based on physical distances between the user and the application[3]), just to name the most common functions. Hence the user traffic is distributed accordingly across the routed WAN.

[3] The protocol used is Director Response Protocol (DRP). DRP Based Dynamic Network Proximity actively localizes client traffic by probing the client and routing it to the closest data center based on the lowest RTT measurement

Figure 4 Data Replication Based on Network Load

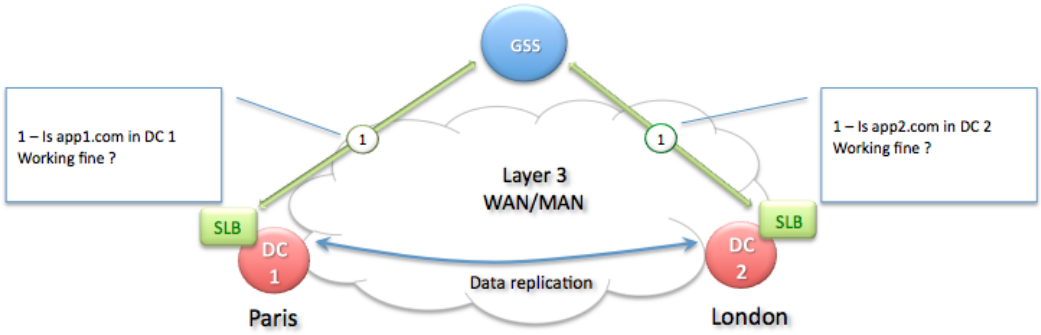
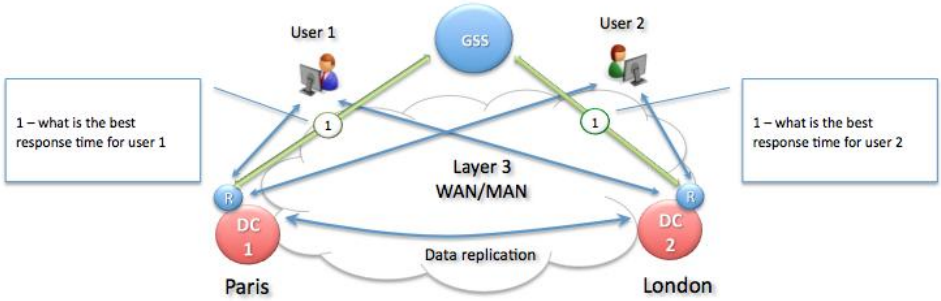


Figure 5 Data Replication in Collaboration with Routers



- **HTTP Traffic Redirection Between Sites:** In case of insufficient resources, the local SLB device will return an HTTP redirection message type (HTTP status code 3xx) to the end-user so that the web browser of the client can be automatically and transparently redirected to the elected backup data center where resources and information are available.

- **Route Health Injection:** Route Health Injection (RHI) provides a real-time, very granular distribution of user traffic across multiple sites based on application availability. This method is initiated by an SLB device that will inform the upward router about the presence or absence of selected applications based on extremely accurate information. This information is usually related to the status of the services that it supports. Therefore, the redirection of the user request to a remote site occurs in real time.

Cloud Computing and Disaster Avoidance

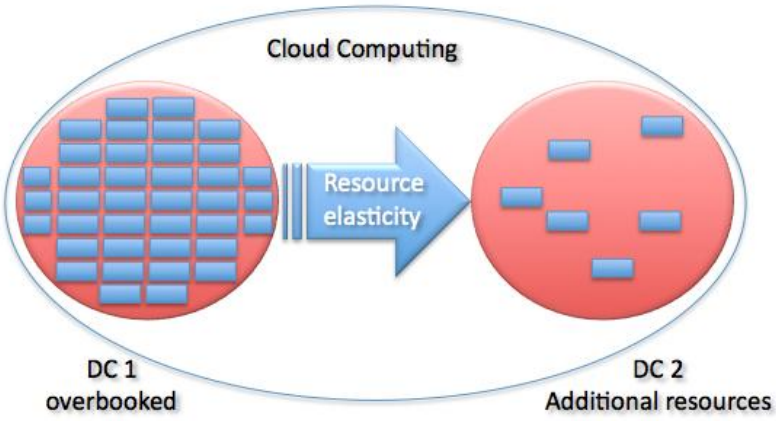
The Need for Distributed Cloud Networking

If Hot Standby disaster recovery solutions running in a traditional routed DCI network are still valid and are often part of the specifications of the enterprise, there are some applications and emerging services enabled with cloud computing that require an RTO and RPO of zero with a full transparent connectivity between the data centers.

These requirements have already been addressed, especially with the deployment of high-availability clusters stretched between multiple sites. However, the rapid evolution of business continuity for cloud computing requires us to offer more reliable and efficient DCI solutions to meet the flexibility and elasticity needs of the service cloud: dynamic automation and almost unlimited resources for the virtual server, network, and storage environment in a fully transparent fashion (Figure 6).

There are two important reasons why disaster recovery is critical to cloud computing. First, physical data centers have "layer zero" physical limits and constraints such as rack space, power, and cooling capacity. Second, there is not currently any application written for the distributed cloud that accounts for the type of network transports and distances needed to exchange data between different locations.

Figure 6 DCI Solution with Resource Elasticity



To address these concerns, the type of network interconnection between the remote data centers that handles the cloud infrastructure needs to be as resilient as possible and must be able to support any new connections where resources may be used by different services. It should also be optimized to provide direct, immediate, and seamless access to the active resources dispersed at different remote[4] sites.

The need for applications and services to communicate effectively and transparently across the WAN or metro network is critical for businesses and service providers using private, hybrid, or public clouds.

To support cloud services such as Infrastructure as a Service (IaaS), VDI/VXI, and UCaaS in a sturdy, resilient, and scalable network, it is also crucial to provide highly-available bandwidth

[4] Remote sites of less than few miles are not necessarily concerned with access optimization.

with dynamic connectivity regardless of VM movement. This is true for the network layer as well as for the storage layer.

The latency related to the distances between data centers is another essential element that must be taken into account for the deployment of the DCI. Each service and function has its own criteria and constraints, so it relies on the application requiring the lowest latency to be used as a reference to determine the maximum distance between physical resources.

To better understand the evolution of DCI solutions, it is important to highlight one major difference when comparing the active/standby behavior between members of a high availability (HA) cluster and the live migration of VMs spread over multiple locations. Both software frameworks require LAN and SAN extension between locations.

High Availability Cluster Requirement Versus Virtual Machine Mobility

When a failover occurs in an HA cluster, the software components have to be restarted on the standby node. After the storage has been replicated to the remote location using synchronous or asynchronous mode[5], the standby node can continue to handle the application safely.

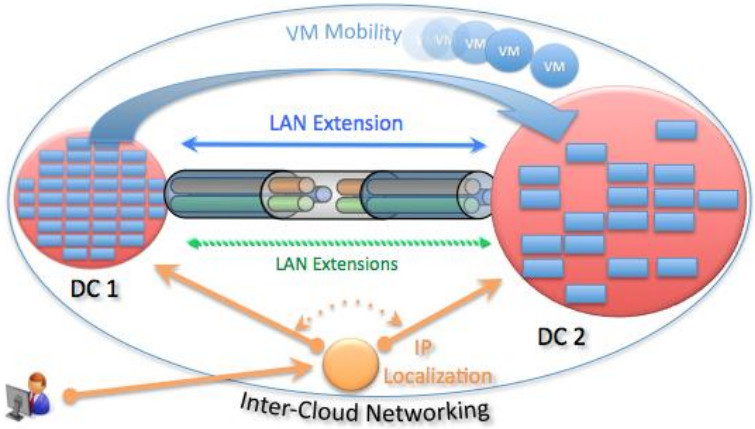
For cloud computing, it is usually necessary to keep the session stateful during and after the migration of a VM. When stateful session is necessary during the movement of the system, the distance between the two physical hosts is driven by the maximum latency supported, the synchronous replications for data mirroring, and the services associated with the storage replication. This means a maximum distance of 100 km between two hosts should not be exceeded when using an Active/Active storage mode requiring synchronous data replication engines. It is important to note that this distance drops to 40-50 km when using the default shared storage mode (see section on Storage Extension).

The elasticity of cloud computing is also driven by the requirement of the active sessions to be maintained with no interruption of service, therefore live migration services in real time are limited to metro distances du to the synchronous mirroring (zero RPO). Beyond metro distances[6] and using current storage tools, the whole cloud computing solution becomes a DR service. Therefore service functions such as Site Recover Manager (SRM)© are very efficient and built for that purpose, but in a stateless session mode.

[5] If some traditional components require synchronous links between members of a cluster such as a Quorum disk, it exists also some approaches to bypass the limitation of the distances imposed by the synchronous copy, like a majority node set function. With this last option, data mirroring can be achieved using an asynchronous mode and most of HA clusters can be stretched over unlimited distances.

[6] EMC recently introduced VPLEX Geo, which supports active/active data over distance to allow applications and data at two locations to be accessed, shared and dynamically moved thousands of km.

Figure 7 Four Components of DCI



Cisco addresses DCI requirements with a complete set of network and storage architecture solutions. The Cisco solution is based on the four components that are critical to providing transport elasticity, flexibility, transparency, and resiliency between two or more sites (Figure 7):

- Layer 2 extension
- SAN extension
- Routing interconnection between sites
- IP Localization for Ingress, Egress and Server to Server workflows

Layer 2 Extension

Layer 2 switching over the WAN or the metro network, whether it is a native Ethernet frame format or a Layer 2 over TCP/IP over any type of transport, should not add any latency to that imposed by the physical distance between sites (mostly dictated by the speed of light). Switching technologies used to extend the Layer 2 over Layer 3 (L2oL3) and obviously the native Layer 2 protocol must be computed by the hardware (ASIC) to achieve line rate transport. The type of LAN extension and the choice of distance are imposed by the maximum latency supported by HA cluster framework and the virtual mobility.

It is crucial to consider whether the number of sites to be connected is two or more than two. Technologies used to interconnect two data centers in a back-to-back or point-to-point fashion are often simpler to deploy and to maintain. These technologies usually differ from more complex multipoint solutions that provide interconnections for multiple sites.

The drawback to simple point-to-point technology is reduced scalability and flexibility. This is a serious disadvantage for cloud computing applications, in which supporting an increasing number of resources in geographically distributed remote sites is critical.

Therefore, enterprises and service providers should consider possible future expansion of cloud networking and the need to operate without disruption when considering Layer 2 Extension technologies. Solutions for interconnecting multiple sites should offer the same simplicity as interconnecting two sites, with transparent impact for the entire site. Dynamically adding or removing one or several resource sites in an autonomous fashion is often referenced as "Point to Cloud DCI". The network manager should be able to seamlessly

insert or remove a new data center or a segment[7] of a new data center in the cloud to provide additional compute resources without modifying the existing interconnections and regardless of the status of the other remote sites.

Whatever DCI technology solution is chosen to extend the Layer 2 VLANs between remote sites, the network transport over the WAN must also provide secure ingress and egress access into those data centers.

When extending Layer 2, a number of rules must be applied to improve the reliability and effectiveness of distributed cloud networking:

- The spanning tree domain should not be extended beyond a local data center, although all the links and switches that provide Layer 2 extension must be fully redundant.
- The broadcast traffic must be controlled and limited by the edge devices to avoid the risk of polluting remote sites and should not have any performance impact.
- All existing paths between the data centers must be forwarded and intra-cloud networking traffic should be optimized to better control bandwidth and latency.
- Some workflows are more sensitive than others. Therefore, when possible, diverse path should be enabled for some services such as heartbeat, used by clusters, or for specific applications such as management or monitoring.
- The Layer 3 services may not be able to natively locate the final physical position of a VM that has migrated from one host to another. This normal behavior of routed traffic may not be efficient when the Layer 2 network (Broadcast Domain) is extended over a long distance and hosts are spread over different locations. Therefore, the traffic to and from the default gateway must be controlled and restricted onto each local data center when appropriated. Similarly, the incoming traffic should be redirected dynamically on the physical site where virtualized applications have been activated.
- For long-distance inter-site communication, mechanisms to protect the links must be enabled and rapid convergence algorithms must be provided to keep the transport as transparent as possible.
- The VLANs to be extended must have been previously identified by the server and network team. Extending all possible VLANs may consume excessive hardware resources and increase the risk of failures. However, the DCI solution for LAN extension must provide the flexibility to dynamically remove or add any elected VLAN on demand without disrupting production traffic.
- Multicast traffic should be optimized, especially in a cloud architecture made up of geographically dispersed resources.

The diversity of services required in a cloud computing environment and the constraints related to the type of applications moving over the extended network require a set of diversified DCI solutions. Cisco offers three groups of technical solutions that meet these criteria:

[7] Public Clouds usually require support of multi-tenants throughout the same infrastructure. Therefore the elasticity of the service offered by the cloud must be very granular and enabled per tenant.

Native Extended Layer 2:

- **Point-to-Point Interconnections:** For point-to-point interconnections between two sites using a dedicated fiber or a protected dense wavelength-division multiplexing (DWDM[8]) mode, Cisco offers Multi-Chassis EtherChannel (MEC) solutions that allow multiple physical links of a Port Channel to be distributed over two different chassis. MEC is available through two approaches:
    - A single control plane managing the two chassis: This method is available on the Catalyst 6500 series with the function of Virtual Switching System (VSS).
    - An independent control plane: This option is available on Cisco Nexus 5000 and Cisco Nexus 7000 Series switches with the function of a virtual Port-Channel (vPC).

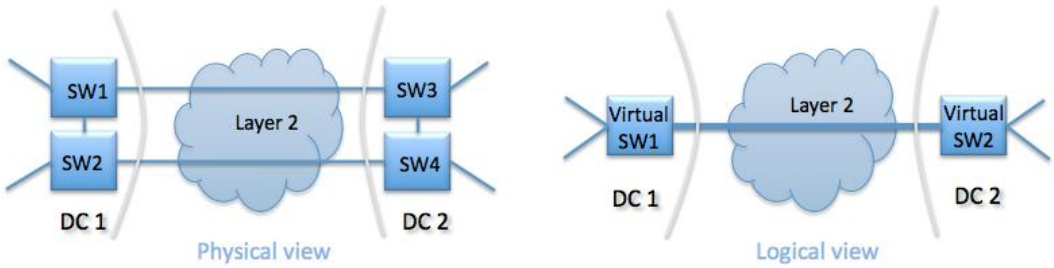
These options can provide active physical link and edge device redundancy to ensure the continuity of traffic between the remote sites, for type 1 and type 2 faults. Both approaches eliminate the use of the Spanning Tree protocol to control the loops. In addition, the MEC solution improves bandwidth utilization (Figure 8).

- **Multiple Site Interconnections:** For multi-site interconnections using optical links or using a DWDM service running in protected mode, FabricPath (TRILL) can quickly and seamlessly connect multiple remote sites in a fabric fashion, remove the extension of the Spanning Tree Protocol between remote data centers, and offer huge scalability compared to classical Ethernet. FabricPath is available on the Cisco Nexus 7000 Series Switches, with upcoming availability on Cisco Nexus 5500 Series[9] Switches. FabricPath can also be used in a point-to-point model, which supports tying additional data centers into the cloud without impacting the production network or affecting existing connections.
- **Security:** Traffic sent through the DCI Layer 2 extension can also be encrypted between a Cisco Nexus 7000 Series Switch deployed at the network edge using the Cisco feature called TrustSec (CTS). With CTS, encryption is performed by the hardware at line rate without impacting the performance or the latency of the traffic crossing the inter-site network. CTS offers a rich set of security services including the confidentiality of data transmitted over the WAN via a standard encryption mechanism (802.1AE).

[8] It is strongly recommended to interconnect DC using native Ethernet frame format using dedicated fiber. However if DWDM is the only option, for resiliency concerns it is imperative to rent the DWDM service running in protected mode.

[9] On Nexus 55xx series with Fairhaven Release - 2HCY11 NX-OS: 5.1(3)

Figure 8 MEC Solution



Extended Layer 2 over Layer 3 (L2 over L3):

EoMPLS

For point-to-point networks across very long distances, Ethernet over Multiprotocol Label Switching (EoMPLS) Pseudowire can be useful. The EoMPLS service is supported natively on Cisco Catalyst 6500 Series Switches with the Sup720 and Sup2T cards. In conjunction with the VSS function (clustered switches), the resiliency of the L2 VPN service can be easily improved for a DCI LAN extension. VSS provides a fully-redundant physical solution that enables a logical L2 over L3 link (Pseudowire) flawlessly and without the need to activate the Spanning Tree protocol between the remote sites. EoMPLS is also supported on Cisco ASR 1000 Series Routers. L2 over L3 extends the Layer 2 Pseudowire over unlimited distances.

With an additional SIP or ES+ card on the Catalyst 6500, the EoMPLS function can be encapsulated directly into a GRE tunnel. This gives the option to extend the Layer 2 VPN over a pure IP network. In this case, the technical knowledge and experience required for an MPLS environment is no longer imposed. In addition, the GRE tunnel may be encrypted using the standard point-to-point encapsulation method of IPSec.

For Multiple data center interconnections, Cisco offers two technologies to address the requirements of cloud computing:

- VPLS
- OTV

Virtual Private LAN Service

Virtual Private LAN Service (VPLS) is available via two approaches:

- **A-VPLS:** Advanced VPLS (A-VPLS) is designed for enterprise environments. A-VPLS is available on Cisco Catalyst 6500 Series Switches using a SIP-400 or ES+ WAN card[10]. This option takes advantage of the system virtualization capabilities of the Catalyst 6500 VSS so that all physical links and edge switches are redundant and active without extending the Spanning Tree protocol between sites. A-VPLS has been specifically designed to offer simplicity of implementation with all the features and performance of the MPLS transport protocol. This feature can also be implemented on a pure IP core network via a GRE tunnel.

[10] A-VPLS will be available in future natively supported on Sup2T without the need of additional WAN card. Currently the Catalyst 6500 with the Sup2T already supports the traditional VPLS deployments including in VSS mode.

- **H-VPLS:** H-VPLS is designed for service provider environments, in which very high capacity interconnections and segmentation are required. It can process information in very large private, public and hybrid cloud environments with large numbers of multi-tenants. H-VPLS is available with the Cisco ASR 9000 Series Routers .

Chassis Link Aggregation Group (MC-LAG)

MC-LAG enables downstream devices to dual-home one or more bundles of links using the Link Aggregation Control Protocol (LACP) 802.3ad in an active/standby redundancy mode, so the standby takes over immediately if the active link(s) fails. The dual-homed access device operates as if it is connected to a single virtual device. MC-LAG is usually enabled on the provider edge (PE) device.

Cisco Router 7600 Series Routers and the Cisco ASR 9000 Series Aggregation Services Routers support this feature. With MC-LAG, the two routers function as Point of Attachment (POA) nodes and run an Inter-Chassis Communication Protocol (ICCP) to synchronize state and to form a Redundancy Group (RG). Each device controls the state of its MC-LAG peer for a particular link-bundle; one POA is active for a bundle of links while the other POA is a standby. Multiple active link bundles per chassis are supported[11].

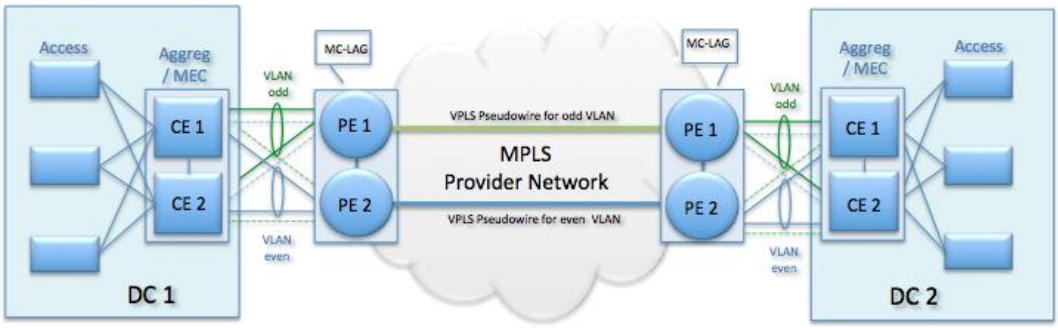
MC-LAG can work in conjunction with L2VPN such as VPLS, but other network transport services such as EoMPLS, L3VPN or QoS can be leveraged as well.

A use case in the context of DCI LAN extension is that at the edge of a provider's network, each customer edge (CE) device supporting LACP is dual-homed to two provider edge (PE) devices and distributes the load on a VLAN-based hashing mechanism onto multiple link bundles. Then the MC-LAG device bridges and extends the concerned VLANs over an MPLS core using a VPLS Pseudowire. The MEC function can be enabled on the aggregation layer to improve Layer 2 multipathing intra-data center, so all Layer 2 uplinks from the access layer to the aggregation layers are forwarded (Figure 9).

MC-LAG offers rapid recovery times in case of a link or node failure, while VPLS addresses the traditional fast convergence, fast reroute and path diversity features supported by MPLS.

[11] Figure 9: With the current release of the ASR9000 series, the VLAN load repartition would require multiple physical bundles while with the Cisco 7600 Series and future Cisco ASR 9000 release, VLAN balancing can be achieved over one unique bundle.

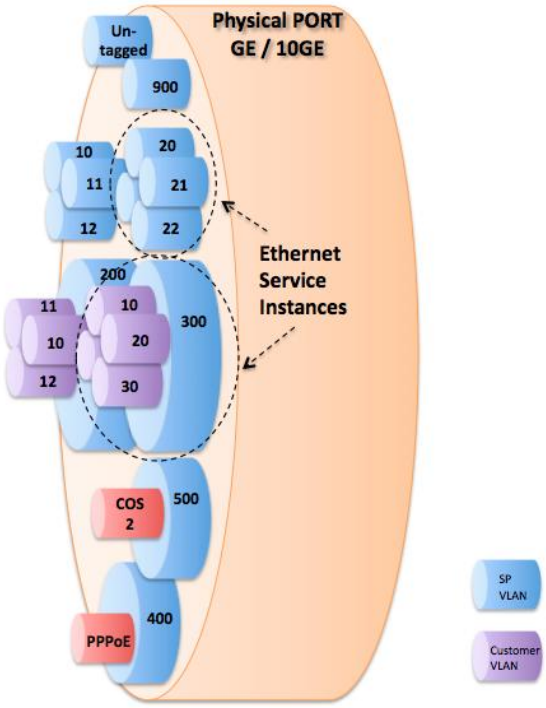
Figure 9 MEC and MC-LAG Function



Ethernet Virtual Connection

Ethernet Virtual Connection (EVC) is a Cisco carrier Ethernet equipment function dedicated to service providers and large enterprises. It provides a fine granularity to select and treat the inbound workflows known as service instances, under the same or different ports, based on flexible frame matching (Figure 10).

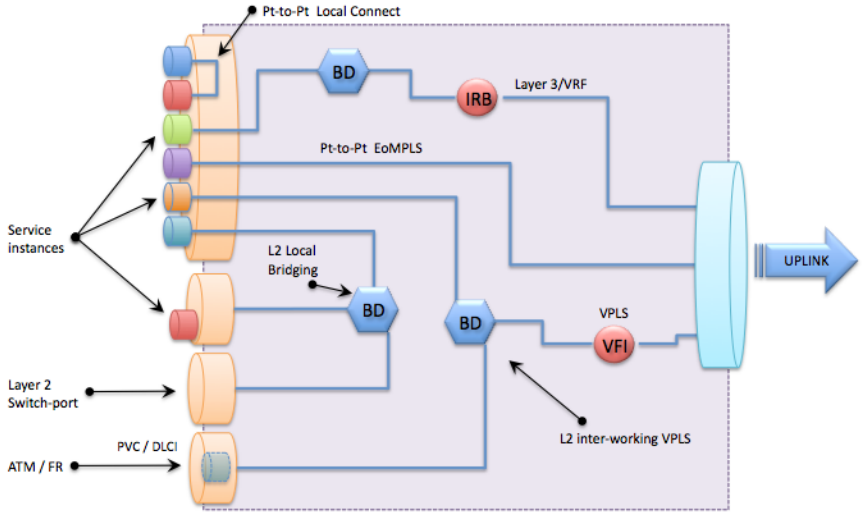
Figure 10 Ethernet Virtual Connection



One of benefits of EVC is the ability to address independent Ethernet encapsulation on the same physical interface with a mix of different services such as dot1q trunk, dot1q tunneling, EoMPLS Xconnect, VPLS attachment circuit, routing, and Integrated Routing and Bridging (IRB). Each service instance can match a unique identifier or a range of identifiers (i.e. a VLAN tag deployed mainly in the context of DCI).

Another important feature is the ability to aggregate multiple service instances into the same transport virtual instance. For example, EVC can multiplex multiple VLANs into the same Bridge Domain (BD) connected to a Virtual Forwarding Instance (VFI) of a VPLS (Figure 11).

Figure 11 Multiplexing Multiple VLANS



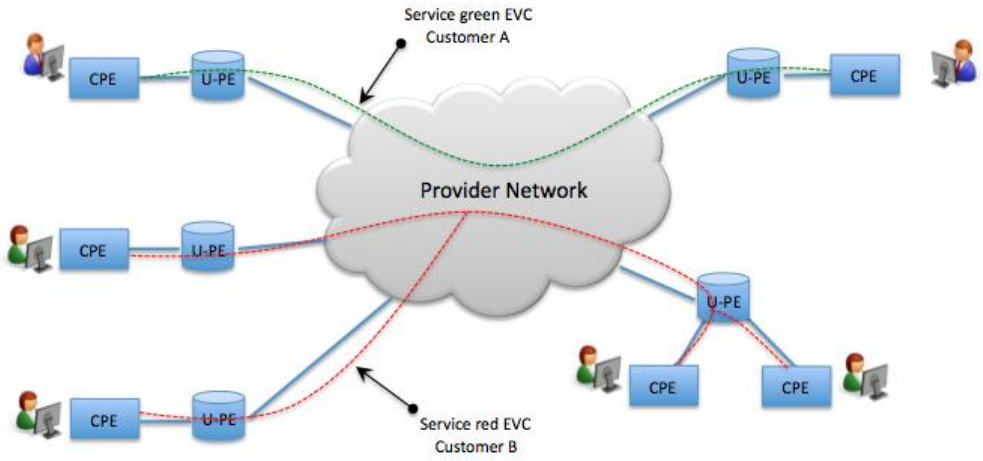
Another important element for a multi-tenancy environment is the ability for each service instance to change the VLAN tag using a new identifier, which allows dynamic VLAN ID translation.

In addition, the flexibility of the service mapping of EVC allows improving the scalability in terms of service instances, VLAN per Bridge Domain or per EoMPLS Xconnect or per VFI. Under the same or different physical interface, multiple service instances can be mapped into the same bridge-domain to provide L2 local bridging between physical interfaces and leverage the usage of VPLS to bridge L2 frames across the MPLS core.

If we leverage the concept of EVC in the public cloud context, a customer service is an EVC. This EVC is identified by the encapsulation of the customer VLANs within an Ethernet island or Bridge Domain, and is identified by a globally unique service ID. A customer service can be point-to-point or multipoint-to-multipoint.

Figure 12 shows two customer services: Service Green is point to point; Service Red is multipoint to multipoint.

Figure 12 Customer Service Options



Overlay Transport Virtualization (OTV)

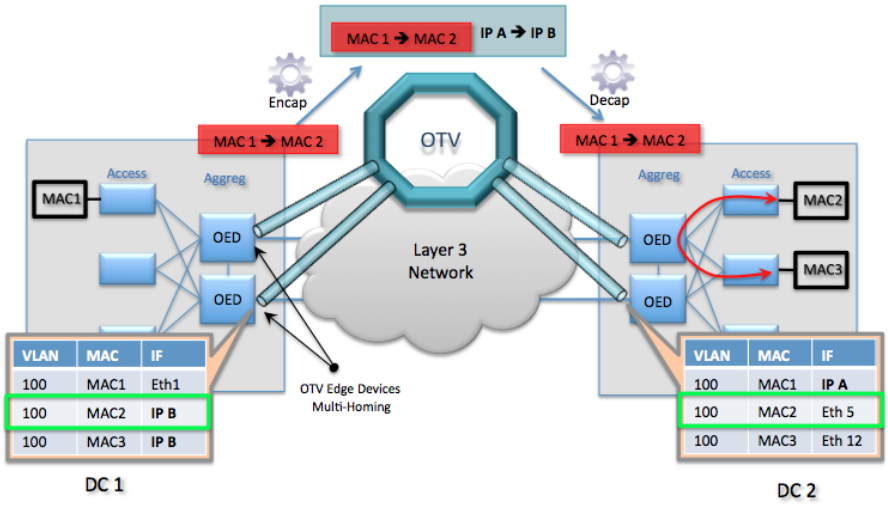
Cisco has recently introduced a new feature called OTV that extends Layer 2 traffic between multiple sites over a Layer 3 network. The edge devices that interconnect data centers are known as OTV edge devices.

OTV dynamically encapsulates Layer 2 packets into an IP header for the traffic sent to the remote data centers. Routing Layer 2 traffic on top of a Layer 3 network is known as "MAC routing" transport. MAC Routing leverages the use of a control protocol to propagate MAC address reachability information, this is in contrast with the traditional data plane learning done in technologies like VPLS.

In Figure 13, MAC 1 sends a Layer 2 frame to destination MAC 2. On the MAC table of the OTV edge device (DC1), MAC 1 is a local address (Eth1), while the destination MAC 2 belongs to a remote location reachable via the IP address B (remote OTV Edge device).

The local OTV-ED encapsulates the Layer 2 frame using an IP header with as IP destination "IP B". The remote OTV-ED removes the IP header and forwards the frames to its internal interface (Eth 5). Local Layer 2 traffic is treated like any classical Ethernet switch (i.e. MAC 2 ⇔ MAC 3 on DC2).

Figure 13 Overlay Transport Virtualization



A control plane protocol is used to exchange MAC reachability information between network devices, extending the VLANs between the remote sites while the learning process inside the data center is performed as in any traditional Layer 2 switch. This mechanism of advertisement destined to the remote OTV edge device differs fundamentally from classical Layer 2 switches, which traditionally leverage the data plane learning mechanism based on L2 source MAC address discovery: if the destination address is unknown after a MAC lookup on the MAC table, the traffic is flooded everywhere.

With OTV, the process for learning MAC addresses is performed by advertising the local MAC tables to all remote OTV edge devices. Consequently, if a destination MAC address is not known, the packet destined to the remote data center is dropped.

This technical innovation has the advantage of removing the risk of broadcasting unknown Unicast addresses from one site to another. This technique is based on a routing protocol, and provides a very stable and efficient mechanism of MAC address learning and Layer 2 extension while maintaining the failure domain inside each data center.

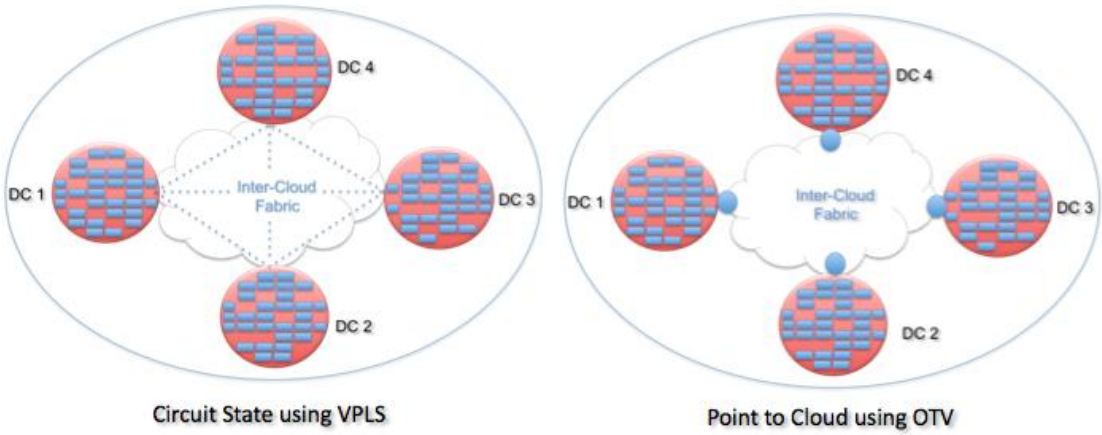
While OTV natively maintains the STP and the failure domain within each local data center, it provides the ability to deploy multiple OTV edge switches in the same data center in active mode. This function is known as Multi-Homing.

OTV works across any type of transport (Fiber, TCP/IP, MPLS) extended between the remote sites with the reliability and effectiveness of the Layer 3 protocol.

In addition to these primary functions, which are essential for the cloud networking, OTV offers several very important innovations:

- OTV connects two or more sites to form a single virtual data center (Distributed Virtual Data Center). No circuit states are required between the remote sites to establish the remote connection (Figure 14). Each site and each link are independent and maintain an active state. This is known as "Point to Cloud" service, which allows a data center to be securely attached or removed at any time without configuring the remote sites and without disturbing cloud services

Figure 14 Virtual Data Center



- OTV offers a native multicast traffic optimization function between all remote sites.
- OTV is currently available on Cisco Nexus 7000 Series Switches and the Cisco ASR 1000[12] Series Aggregation Services Routers.

Storage Extension

The distance between the physical resources and the effects of VM migration must be addressed to provide business continuity and DA when managing storage extension. The maximum distance is driven by the latency supported by the framework without impacting the performance.

VMs can be migrated manually for DA, or dynamically (e.g. VMware Dynamic Resource Scheduler) in a cloud environment. VM migration should occur transparently, without disrupting existing sessions.

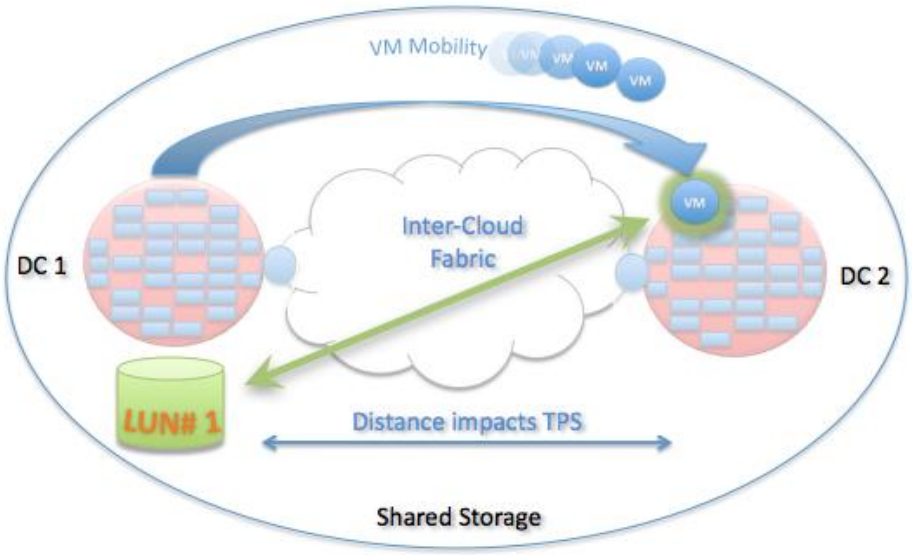
VMs use dedicated storage volumes (LUN ID) provisioned from a SAN or a NAS disk array. These storage volumes cannot be replaced dynamically without affecting the active application and stateful sessions.

In order for the VMs to move from one physical host to another in a stateful mode, they need to keep the same storage. This behavior is not a concern when the VM moves from one host to another within the same physical PoD or between PoDs inside the same physical data center, as the distances between hosts and storage disks are very short. The storage is therefore provisioned in shared mode with the same physical volumes accessible by any host.

Shared storage means that during and after the movement of a VM, the operating system remains attached to the same physical LUN ID when the migration occurs between two hosts.

[12] OTV is supported in ASR 1K IOS-XE 3.5.0S (Nov 2011)

Figure 15 Shared Storage



However, this shared mode of operating storage may have an adverse effect in the environment of DCI due to a long distance between hardware components. According to the rate of transactions per second (TPS) and depending on how much I/O the application itself consumes (e.g. database), beyond 50 km (1ms latency in synchronous mode), there is a risk of impacting the performance of the application. Assuming a VM has moved to a remote site, by default it continues to write and read data stored on its original physical volume (Figure 15).

Several storage services can be enabled to compensate for this behavior:

Cisco IOA: Cisco provides an I/O Acceleration (IOA) function on Cisco MDS 9000 Series Fabric Switches. IOA can halve the latency for synchronous write replication, thus doubling the distance between two sites for the same latency. In partnership with ECO partners NetApp and EMC, Cisco and VMware have tested and qualified two types of storage services to improve the sensitive remote I/O effect due to VM mobility between sites.

FlexCache from NetApp: This feature supports local storage cache of (i.e. secondary DC) data that has been previously read on the original disk. Any read command associated with the data already stored locally doesn't have to cross the long distance between the two sites, and thus this function has a negligible latency on read commands, although the original ID is still physically on the primary data center (shared storage). Therefore the current stateful sessions can retain their active state during and after VM migration without being disturbed. FlexCache operates in a NAS environment. The actual data is still written on the single location at the original site. Therefore this mode of storage remains shared.

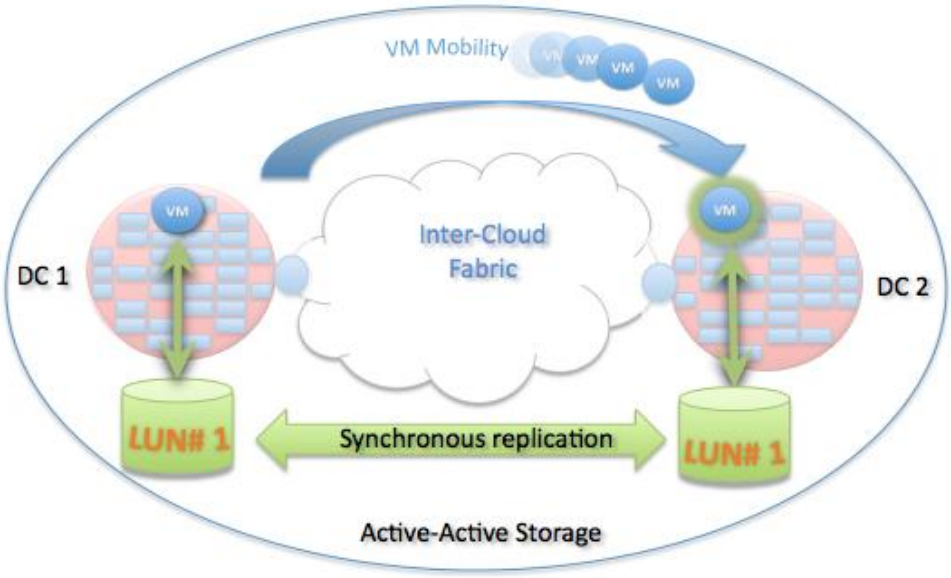
VPLEX Metro from EMC: This feature allows users to create a virtual volume that is distributed between two remote sites. Both volumes can synchronously present the same information on two different sites.. The volume is created and shared between two VPLEX clusters, connected via an extended Fiber Channel running in synchronous mode. The data is replicated and synchronized between the VPLEX devices using dedicated FC link.

The initiator (host) writes data on the same but virtual LUN ID available on both sites at the same time. This technology replicates the same settings of the SCSI parameters on both

storage targets (VPLEX), making the change of physical volume transparent to the hypervisor. The maximum distance between the two cluster members of the VPLEX metro should not exceed 100 km[13] due to the replication running in synchronous mode. Synchronous mode is required to maintain the transparency of this service.

This function works in a SAN environment and the storage mode is therefore known as Active/Active (Figure 16).

Figure 16 Active/Active Storage



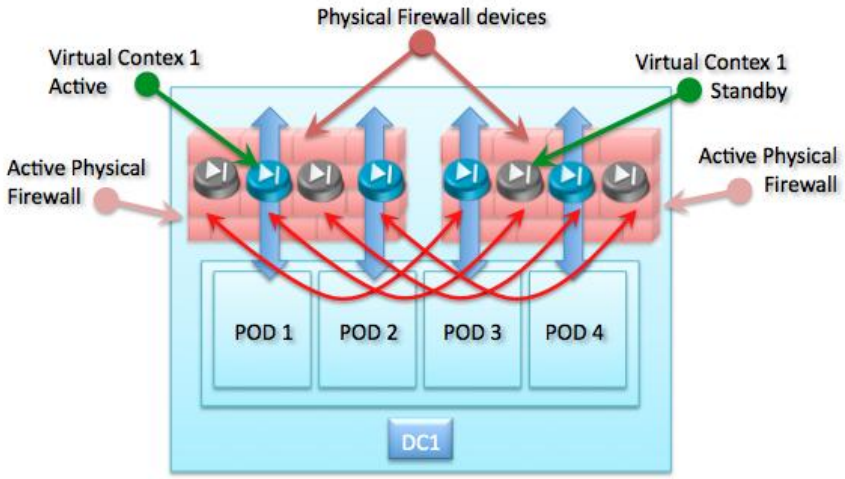
Security Services Placement

Modern firewalls, load balancers, and most stateful devices support the concept of virtual context, which is the ability to support multiple virtual firewalls or virtual load balancers. Up to 250 virtual contexts, fully autonomous and isolated from each other, can be enabled on a single physical appliance or service module.

To offer the high availability service required for business continuity, firewalls and load balancers work by pairing physical devices. Both devices remain active while the virtual contexts run in an active/standby fashion. In addition to providing redundancy, this mechanism distributes the active context between the two devices, improving the total throughput for active workflows (Figure 17)

[13] 100km is a safe maximum distance. Notice that according to last update from EMC, the distance between the 2 VPLEX clusters can go up to 5ms (250kms in synch mode). Therefore it is critical to validate that the VM does not migrate faster than the replication of the data (i.e. the VMotion VLAN using a different and shorter fiber than fiber used by the VPLEX service, if different!)

Figure 17 Virtual Context of Security



When interconnecting multiple data centers and deploying firewalls and other stateful devices such as load balancers, the distance between remote sites is an important consideration.

When data center sites are deployed in close proximity (such as within the few kilometers that is typical for large campus deployments), they can be considered a single, physically-stretched data center location. Under these premises, it would probably be acceptable to deploy the stateful devices in a stretched fashion, with one physical member of the HA pair in each data center site. For deployments where the distance between locations is farther, a pair of HA devices is typically deployed in each physical data center (Figure 18).

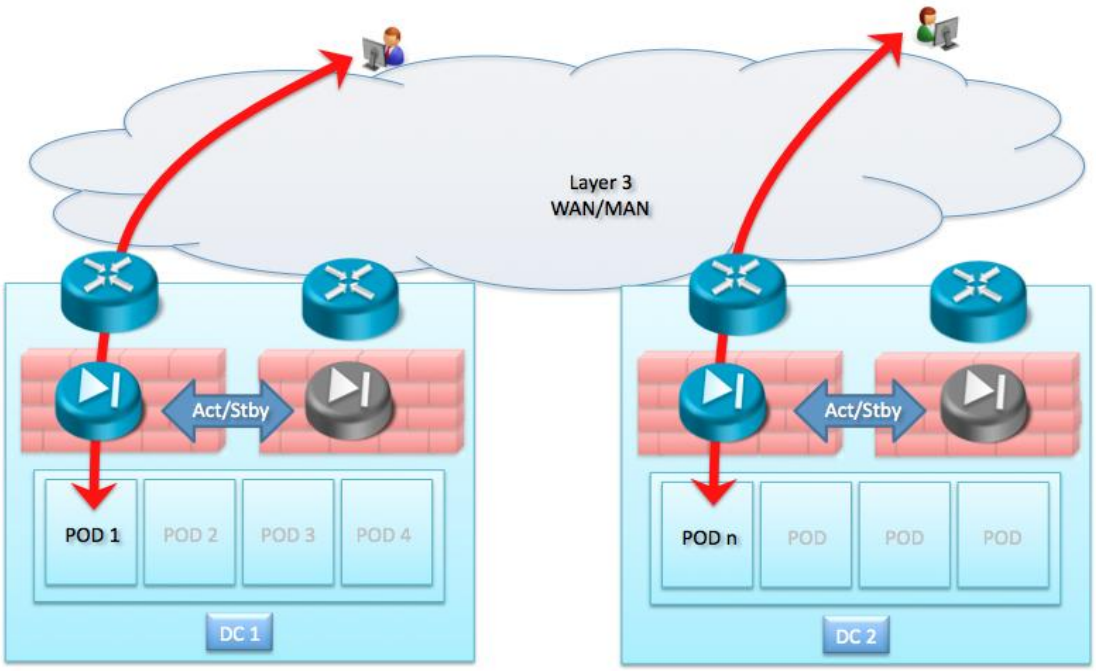
There are some important reasons to maintain each active/standby pair within the same data center. Redundancy is controlled by two devices[14], which means that dispersing the active/standby contexts in two different locations would limit the maximum number of data centers to two. On the other hand, keeping the active/standby pair of network services inside the same physical data center, allows replicating the same security policies in more than two data centers.

In addition, the link between the physical devices used for health check and process synchronization (replication of the active flows for stateful failover) must be extended in a very solid fashion. Due to its function of fault tolerance, it is also very sensitive to latency.

Last but not least, security and optimization functions usually require maintaining a stateful session. Therefore, for the same session, the traffic should be returned to the original virtual context that acknowledged the first flow, otherwise the flow will be dropped.

[14] In its next release, the Cisco ASA firewall will support a cluster of up to eight physical firewalls.

Figure 18 HA Device Pair



This behavior of symmetrical paths should be controlled and maintained, especially with the migration of VMs over a LAN extension as explained in the next topics.

Network Service Localization and Path Optimization

The ability to avoid disasters is improved by distributing physical compute and network resources between data centers that are geographically distributed over long distances. Geographic distribution provides higher elasticity and almost unlimited flexibility of the resources required to dynamically deploy VM loads.

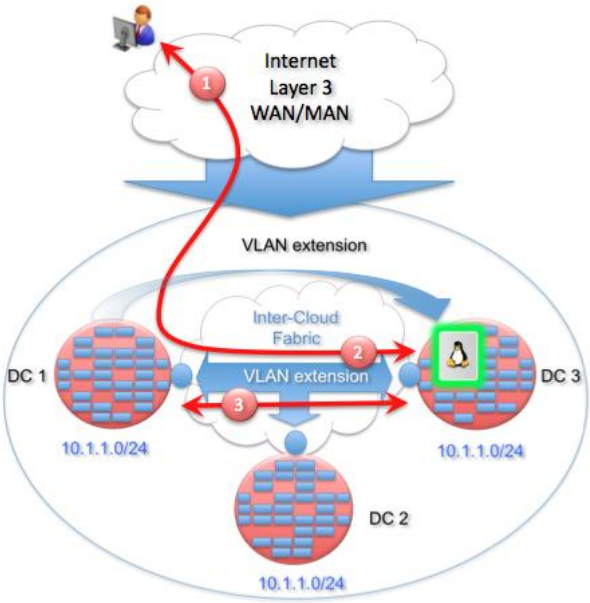
The network side can transparently support distributed applications by extending Layer 2 between multiple sites. Yet by definition, the Layer 3 traffic carried between users and active applications through the cloud does not have native knowledge of the physical IP device locations, other than the network prefix given through the most significant bit-group of the IP address. The IP subnet is a logical visible subdivision of the network that is usually limited to the local network. It therefore delimits its broadcast domain defined by the system mask. The IP subnet is usually established by the enterprise or service provider network team. In general, an IP subnet addresses a set of IP equipment that belongs to the same VLAN.

Traditionally, if an IP subnet or a VLAN is associated with a physical location inside a data center, with the concept of interconnecting cloud resources, the broadcast domain is stretched over the distances that separate the data centers (DCI theoretically can be established up to unlimited distances with L2 over L3 transport). Therefore, the concept of location induced natively by the IP subnet subdivision loses one of its original functions of localization.

Thus, depending on the distance between the remote sites, the native routing mechanism can have an impact on performance for three major types of communication (Figure 19):

1. Traffic from the user to the server
2. Traffic from the server to the user
3. Traffic from server to server (such as in a multi-tier application)

Figure 19 Three Communication Types

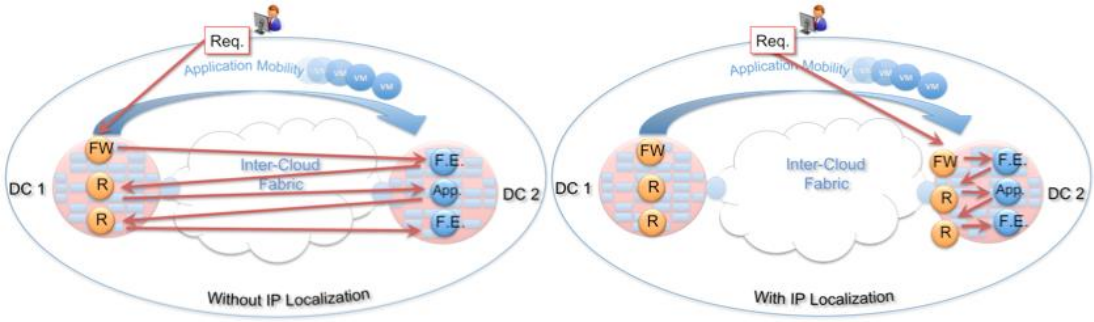


Server to Server Traffic

When a server migrates from one site to another, it must return the traffic to its default gateway because its IP address schema remains the same regardless of its physical location. Since there is one IP address (or virtual IP addresses (VIP)) for a given default gateway per subnet, this implies that after the migration of a logical server, the traffic must be returned to the original site where the active default gateway stands. In a complex multi-tier architecture, routers and firewalls are usually enabled to improve the communication and security between the tiers.

If, for example, a solution built with a 3-tier application (e.g. Web Server, Application and Database tiers) is moved from one data center to another, the traffic between each tier will have to return to the site where the gateways or firewalls are active. If we add to that the different network services required for optimization and data security (load balancer, SSL termination, IPS) enabled at different tiers, then up to ten round trips for a simple query may occur. Consequently, depending on the distance between the data centers, the latency for a request may be significantly affected (i.e. additional 10 to 20 ms for 100 km using dedicated fiber for a 10 round trips).

Figure 20 Reducing Query Latency



It is therefore crucial that the inter-application-tier or server-to-server traffic is better controlled to minimize the "ping-pong" effect (Figure 20).

Emerging solutions such as EMC VPLEX® Geo, which support active/active data over thousands of km with no service interruption, separate performance and distance considerations.

Cisco supports deployment options for enabling the same default gateway functionalities in different data center sites (FHRP localization). This functionality is completely transparent to the application layer as well as the network layer. By activating this IP localization service, after the migration of VMs it is possible to use a local default gateway configured with the same IP identification (same virtual MAC addresses and virtual IP) that were defined on the original site.

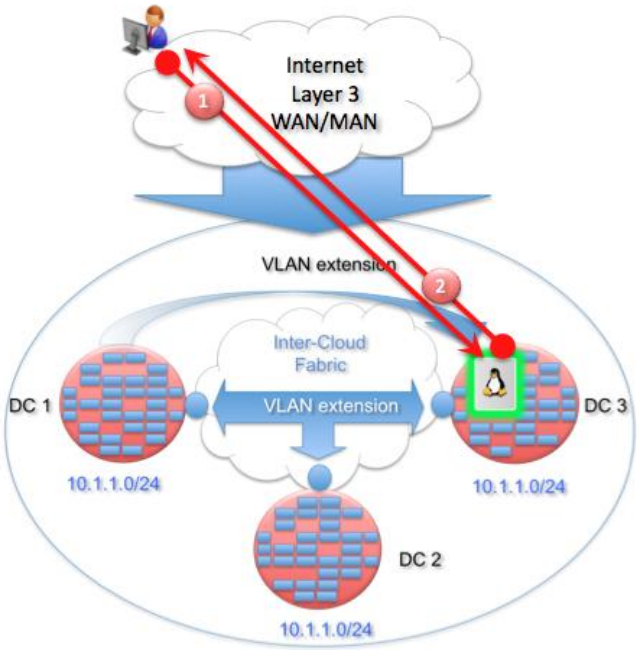
Server to Client traffic

The same function of IP localization can be applied to outbound traffic so that the responses from a server sent to an end-user can exit through its local WAN access without returning the session to the default gateway of origin.

However, it is imperative that when stateful services are deployed, the return traffic remains symmetrical with the incoming flows. This ensures the security of all current sessions without disrupting established sessions.

It is therefore important to involve the service of IP localization that exists for outgoing traffic with the other optimizations mechanisms available for the ingress traffic client to server (Figure 21).

Figure 21 IP Localization for Outgoing Traffic



Client to Server Traffic

When a user accesses an application running in a distant resource, the client must be able to use the optimal path and be dynamically redirected to the data center supporting the active application or VM. However, as explained previously, the routed Layer 3 network cannot determine the physical location of an IP device within the same subnet when it is stretched between different locations.

Without any ingress optimization, for long distances between remote sites, more bandwidth is consumed and some delays may be noticeable for the service offered by the cloud (Figure 19). For example, assuming a default application "A" available on the primary data center "X", migrates to a data center "Y", the requests from the remote user will be directed to the primary data center "X" and then the extended Layer 2 path to reach the active application that has moved to data center "Y", and vice versa for the return traffic.

It is therefore important to optimize the path for a remote user as soon as the application has migrated to the next location.

Cisco provides a number of IP localization services that, combined with other IP functions, support path optimization:

- Intelligent Domain Name Server
- Host Route Injection
- Locator/ID Separator Protocol, LISP

Intelligent Domain Name Server

The Global Site Selector (GSS) is an Intelligent Domain Name Server that distributes the user's requests to the remote sites where the applications are active. The GSS has already been described at the beginning of this article (Global Site Load Balancing Services) in a

traditional routed environment. The same equipment can be used for LAN extension in conjunction with other network services such as the SLB devices, as well as with centralized management tools used for migrating VMs[15].

GSS with KAL-AP:
In conjunction with load balancing equipment such as the Cisco Application Content Engine (ACE), the GSS periodically sends probes (Keep Alive Appliance Protocol, KAL-AP) to the load balancing device in order to determine the status of VMs and services distributed throughout the cloud.

When a VM migration is complete, the GSS locates the active application based on regular keep-alive probing, and immediately associates it with the public address of the hosting data center.

The existing sessions are maintained via the original data center to avoid interruption, while all the DNS requests from new clients for this application are updated with the public IP address used for the new site. This mechanism supports load distribution across multiple WANs (different active applications distributed between multiple locations).

In the meantime, this mechanism optimizes traffic by sending all new sessions directly to the location hosting the active service of the cloud without using the Layer 2 extension path.

By correlating the keep-alive function (KAL) from the GSS with other egress path optimization functions such as FHRP localization as described above, new incoming sessions established at the new data center will be optimized for direct return flow using their local default gateway.

To keep the existing sessions secure, the traffic for those current sessions must return to the original site via a mechanism of source-NAT activated at the upstream layer. This allows both scenarios to be used while ensuring symmetrical flows for all sessions.

vCenter and GSS:
A module script is added to the central management of VMs from VMware (vCenter) so that when the VM management migrates a VM on a remote host, manually or dynamically, it informs the upstream GSS about the new location of this particular VM in real time.

The GSS then immediately assigns a public address associated with the new site to the VM. Similarly with the KAL-AP probes described previously, the established sessions remain directed to the same site of origin to maintain the workflows.

Dynamic Routing Based on the Application State.

The Cisco load balancing service module, ACE, provides real-time information about the preferred route to access a specific application or service after it has moved to a new location. The ACE continually probes the state of the applications for which it is responsible.

When one of these services appears in its data center, it immediately sends a static route to reach the application to the adjacent router, which in turn notifies the routed wide area

[15] A mechanism of IP address translation is usually initiated to allow the use of different public IP addresses on each location, while the private IP addresses of the application are kept the same even after a migration process.

network (WAN) with a preferred metric. At the same time, the site of origin withdraws the IP route of the host (or application) that no longer exists from its local routers.

Unlike the GSS, information on Layer 3 routes concerns all existing sessions (current and new) and all sessions will be redirected almost in real time – according to the route table updates - to the new data center hosting the concerned application. However, the current active sessions will be kept safe during the migration because the IP address is unchanged for both the private and the public side. Nevertheless, local stateful devices such as firewalls and load balancers must initiate and validate a new session. Except for some specific protocols primarily related to maintenance purposes such as Telnet or FTP, usually for applications related to the cloud services such as IaaS and/or traditional http based software, this is not detrimental to the services supported.

In fact, an HTTP session that was established through stateful devices such as a firewall in the primary data center can be redirected to a secondary data center offering the same access and application security level and policy rules. After the migration of the concerned application on the new data center, the local stateful devices will accept and initiate a new session for that workflow according to the security policies.

Once granted, the session will be established transparently to the end user. Note that mechanisms based on cookies or SSL IDs or other identifiers used to maintain session persistence between the server supporting the application and the end-user, must be maintained.

vCenter, ACE and OTV – Dynamic Workload Scaling (DWS)

VCenter has the ability to manually or dynamically control system resource use and allocate workload based on the physical resources available throughout the cloud.

The Cisco ACE has the ability to distribute traffic load to multiple physical or virtual servers within a server farm, using different weights based on the performances of the systems or other criteria understood by the system managers.

OTV has the native ability to detect MAC address movement from one site to another via the extension of Layer 2 that it provides.

Combining vCenter, ACE, and OTV can provide a complete solution that can detect the movement of VMs from one site to another in real time (OTV) and modify the weight associated with each real server accordingly, via the Application Network Manager (ANM) of the ACE. By monitoring and defining alarm thresholds using vCenter©, this solution can alleviate the local server farm by sending requests with variable ratio to the servers located on the remote data center.

This provides a very fine granularity and dynamic distribution of the resources that can be preferable for a specific duration, and therefore optimizes the bandwidth between remote sites used by these flows.

Locator/ID Separation Protocol (LISP)

LISP VM-Mobility

Traditionally, an IP address uses a unique identifier assigned to a specific network entity such as physical system, virtual machine or firewall, etc. The routed WAN uses the identifier to also determine the network entity's location in the IP subnet. When VMs migrate from

one data center to another, the traditional IP address schema retains its original unique identifier and location, although the location has actually changed.

This is done because the Layer 2 VLAN between the physical and virtual machines supports the same IP subnet. The extended VLAN must share the same subnet so that the TCP/IP parameters of the VM remain the same from end to end, which is necessary to maintain active sessions for migrated applications.

To identify the location of a network device, LISP separates the identifier of the network device, server or application (known as the EndPoint Identifier) and its location (known as the Locator), once the separation is done, the LISP Mapping System will maintain an association between these two distinct address spaces (End-Point-Identifiers and Locators) The IP address of the identifier is preserved by having it encapsulated into a traditional IP frame for which the destination IP is the location (Locator) of the site where the server or application (EndPoint Identifier) has moved.

A traditional routed network provides reachability to the Locator while the IP address of the EndPoint Identifier can dynamically migrate to a new location without modification of the routing information pertinent to the locator space. Only the information in the LISP Mapping System is updated so the end-point identifier is now mapped to its new locator.

When a VM migrates from one data center to another, the movement is detected in real time and the association between the EndPoint Identifier and the Locator is immediately updated in the RLOC Mapping database. All traffic destined for the VM is dynamically, automatically, and transparently redirected to the new location.

For hybrid clouds, a service provider can move and house the data center of the enterprise without necessarily changing the full address space of network devices and servers.

LISP IP mobility

With the introduction of LISP, IP based solutions allowing a subnet to be dispersed across any location become a reality. As we move forward, many of the workload mobility requirements may be addressed with a combination of LISP mobility and LAN extensions as outlined in this paper, but there may also be a number of deployments in which the LISP mobility functionality alone may be sufficient to address the workload mobility requirements, eliminating the need for Layer 2 extensions. This works well today for specific scenarios such as disaster recovery and live moves. Just to list few use cases where LISP can be very efficient and help remove the need to extend the Layer 2 between sites:

- During process of migrating physical servers to a new data center some applications may be difficult to re-address at Layer 3, such as a Mainframe for example. Avoiding IP address renumbering may ease physical migration projects and reduces cost substantially.
- With hybrid cloud, Enterprise Private/internal cloud resources are moved to a Public/external cloud location. Avoiding IP address renumbering not only ease the migration of the applications from the Enterprise premise equipment, but also reduces the risks and accelerates the deployment of those services. In addition LISP guarantees optimal routing to the active application, regardless of its location, removing the hairpin effect and therefore improving the response time to access the service.

As the technology evolves more and more scenarios will be addressed. In the future, the network architect will have the choice between an L2 and an L3 solution in order to satisfy the DCI requirements that traditionally were focused exclusively on L2 solutions.

Cisco LISP VM-Mobility provides an automated solution to IP mobility with the following characteristics:

- Guaranteed optimal shortest path routing
- Support for any combination of IPv4 and IPv6 addressing
- Transparent to the EndPoints and to the IP core
- Fine granularity per EndPoint
- Autonomous system agnostic

Additional Services improving Cloud computing environment.

Cloud computing is driven by the inherent function of virtualization. In cloud environments, virtualization is required for storage and network transport as well as computing. Expanding virtualization to network and storage resources lets VMs communicate in a secure and flexible fashion while they migrate from one physical host to another. That function of network virtualization is supported by Cisco Nexus® 1000v Series Switches. These virtual switches offer port-profile portability as well as common switch features such as QoS, PVLAN, port-security, and monitoring.

In addition to the explosive technology adoption of cloud computing, the physical to virtual ratio in physical hosts is increasing exponentially as CPU and memory capacity grows.

Since each VM will introduce a new set of MAC addresses into the network and segmentation is achieved by using separate VLANs for each tenant, an exponential growth of MAC addresses and VLANs that share the same physical infrastructure is in progress. With this growth, standards-based protocols such as 802.1Q may not suffice as they are natively limited by 12 bits of VLAN identifier (4k VLANs). Cisco is working closely with VMware, Arista, Broadcom, Citrix and Red Hat (IETF WG) to address the massive scalability requirements for segments supporting mobile VMs through a new technology, referred to as Virtual Extensible LAN (VXLAN). This method of transport leverages the well understood principles of MAC-in-UDP encapsulation, introduced with OTV and LISP, to deliver a 24-bit identifier to provide the large segmentation required for the cloud environment (16M VLANs). With the introduction of this 24 bit identifier, VXLAN presents one instance of what is commonly referred to as a segment-id. Segment-ids must eventually be extended to different types of transport such as those provided by MPLS, LISP and OTV in order to deliver a true end-to-end segmentation solution for global and federated clouds. Thus, in addition to a large layer 2 domain deployment within the same data center with the L2 technologies explained in this paper in support of workload mobility, Enterprises and Service Providers may need to deploy repeatable pods in different subnets for the cloud to grow without any infrastructure changes and VXLAN provides a flexible mechanism for providing the segmentation required in those repeatable entities.

This paper does not provide a detailed discussion of Cisco products that already improve the scalability and security of virtualized networks within a physical data center, such as the Cisco Nexus1000v, VSG, and PVLAN. Pointers to these topics are provided at the end of this document. However, it is important to clarify the new transport function known as VXLAN to avoid any confusion with other existing protocols.

VXLAN addresses the scalability requirements of cloud computing segmentation. It is a new feature embedded into the Cisco Nexus1000v NX-OS and is intended to offer more logical network resources in a fully-virtualized cloud environment, while OTV, VPLS and LISP are responsible for interconnecting data centers for virtual applications as well as traditional HA frameworks. These functions of DCI are achieved by extending the Layer 2 domains between sites over a Layer 3 network (OTV) and by allowing IP mobility between data centers using dynamic routing updates through LISP, while VXLAN is an IaaS infrastructure solution focused on scalable segmentation. VXLAN, OTV, and LISP share similar frame formats, however they serve different networking purposes and therefore complement each other.

What is coming?

Ethernet VPN: E-VPN

To support service providers, Cisco is working with other network vendors to standardize a resilient and massively scalable solution using Ethernet VPN, which will extend Layer 2 traffic over MPLS.

Cisco introduced MAC routing to the L2VPN space in 2009. E-VPN takes the VPN principles introduced and matured by OTV and ports them into a Border Gateway Protocol (BGP) based standard proposal that leverages the MPLS data plane that SPs are used to operate upon. One could think of E-VPN as OTV over a native MPLS transport.

In addition to its strength and high scalability, E-VPN improves redundancy and multicast optimization in MPLS with all-active attachment circuits for multi-homing deployment, which are usually missing in traditional VPLS-based LAN extension solutions and were introduced with MAC routing by OTV.

The mechanism of MAC routing is mainly used to populate the MAC table that exists on each virtual forwarding instance (VFI), but also provides all the necessary information to achieve active multi-homing, load balancing and simplified operations. The process of MAC routing is used for remote advertisement, in which the VFI populated with the information of local and remote MAC addresses connects to the VPLS Pseudowire and forwards the Layer 2 traffic accordingly to its MAC table information.

While the process of MAC routing is used to transport Layer 2 traffic outside the customer site, traditional MAC learning still occurs on the internal interfaces for local switching using classical Ethernet (LACP). E-VPN is expected to support TRILL in a future release.

Detailed information can be retrieved from the IEFT web site repository: E-VPN

Network Virtualization

Cisco ASR 9000 Series Aggregation Services Routers will support a new model of cluster called Network Virtualization (NV). NV will allow a single control plane to virtualize the different routers and members of a cluster, while the data plane is extended on two chassis. The data plane and the distributed fabric processors will be active at the same time in both chassis, doubling the total throughput of device forwarding capacity. A next release will support up to eight chassis per cluster.

State synchronization and health control will be performed through the RSP ports on the supervisor using regular Ethernet frames with dot1Q tags to support low latency between the two members and fast recovery times for any type of failure.

Additional Cisco ASR Series products, such as the Cisco ASR 9000v Series Aggregation Services Routers, can be attached to the cluster so that the centralized control plane manages and controls all the components that form the cluster.

Conclusion

Achieving the high level of flexibility, resource availability, and transparency necessary for distributed cloud services requires four components:

- **Routing Network:** The routing network offers the traditional interconnection between remote sites and gives end-users access to the services supported by the cloud. This component is improved using GSLB-based services such as DNS, HTTP redirection, dynamic host routes, and LISP.

- **LAN Extension:** The technical solution for extending the LAN between two or more sites using dedicated fibers or Layer 2 over Layer 3 transport mechanisms for long distances.

- **Storage Services:** The storage services used to extend access between SAN resources. SANs are highly sensitive to latency and therefore impose the maximum distances supported for the service cloud. It is preferable to use an Active/Active replication model to reduce the latency to its minimum value.

- **Path Optimization Solution:** The path optimization solution improves the server-to-server traffic as well as the ingress and the egress workflows.

Unlike the classical data center interconnection solutions required for geo-clusters that can be stretched over unlimited distances, DA and live migration for the service cloud require that active sessions remain stateful. As a result, maintaining full transparency and service continuity with negligible delay requires that the extension of the LAN and the SAN be contained within metro distances.

Enterprises and service providers may still have strong requirements to extend the LAN and SAN over very long distances, such as the need for operation cost containment or DP in stateless mode. These needs can be addressed if interrupting (even for a short period of time) and restarting sessions after workloads are migrated is acceptable to the system managers. Those requirements can be achieved using tools such as Site Recovery Manager from VMware© or an active-active storage solution such as EMC VPLEX Geo© for more generic system migration.

Some definitions & Acronyms

Note that several generic definitions come from Wikipedia.org.

Recovery time objective (RTO) is the duration of time and a service level within which a business process must be restored after a disaster (or disruption) in order to avoid unacceptable consequences associated with a break in business continuity.

Recovery point objective (RPO) describes the acceptable amount of data loss measured in time.

SRDF (Symmertrix Remote Data Facility) is a family of EMC products that facilitates the data replication from one Symmetrix storage array to another through a Storage Area Network or IP network. More in line EMC SRDF product family.

EMC VPLEX (**VPLEX**) is a virtual storage solution introduced by EMC. VPLEX implements a distributed "virtualization" layer within and across geographically disparate Fibre Channel storage area networks and Data Centers. More in line with "VPLEX Architecture Deployment" as well as the new VPLEX Geo.

NetApp FlexCache software creates a caching layer in the storage infrastructure that automatically adapts to changing usage patterns, eliminating performance bottlenecks. In addition, FlexCache automatically replicates and serves hot data sets anywhere in the infrastructure using local caching volumes. More in line with NetApp FlexCache.

Hitachi TrueCopy, formerly known as **H**itachi **O**pen **R**emote **C**opy (HORC) or Hitachi Remote Copy (HRC) or Hitachi Asynchronous Remote Copy (HARC), is a remote mirroring feature from Hitachi storage arrays available for both open systems and IBM z/OS. Truecopy is an implementation of IBM's PPRC protocol.

Synchronous replication causes each write to the primary volume to be performed to the secondary as well, and the I/O is considered complete only when updates to both primary and secondary have completed. Synch replication guarantees zero packet loss, however overall performance decreases considerably.

Asynchronous replication considers the write complete as soon as it is acknowledged by local storage. Remote storage is updated, but generally incurs with a small delay. Asynchronous replication offers greater performance than synchronous replication, but if local storage is lost, the remote storage is not guaranteed to have the current copy of data. Recent data may be lost.

Oracle Real Application Clusters (RAC) provides software for clustering and high availability in Oracle database environments. More in line Introduction to Oracle Real Application Clusters.

VMware® vCenter Site Recovery Manager (SRM) is an extension to VMware vCenter that enables integration with array-based replication, discovery and management of replicated

data stores, and automated migration of inventory from one vCenter to another. More in line SRM.

Grid computing is a term referring to the combination of computer resources from multiple administrative domains to reach a common goal. More in line with Wikipedia.

A **Point of Delivery**, or **POD**, is "a module of network, compute, storage, and application components that work together to deliver networking services. The POD usually belongs to the data center build block. The POD is a repeatable pattern, and its components maximize the modularity, scalability, and manageability of data centers."

A **subnetwork**, or **subnet**, is a logically visible subdivision of an IP Network. The practice of dividing a network into subnetworks is called **subnetting**.

All computers that belong to a subnet are addressed with a common, identical, most-significant bit-group in their IP address. This results in the logical division of an IP address into two fields, a network or routing prefix and the rest field. The rest field is a specific identifier for the computer or the network interface.

The **Domain Name System** (**DNS**) is a hierarchical distributed naming system for computers, services, or any resource connected to the Internet or a private network. It associates various information with domain names assigned to each of the participating entities. Most importantly, it translates domain names meaningful to humans into the numerical identifiers associated with networking equipment for the purpose of locating and addressing these devices worldwide. More in line with Wikipedia.

Server Load balancing (SLB) offers three main functions:

- It is a computer networking methodology to distribute workload across multiple servers to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload. Using multiple components with load balancing, instead of a single component, may increase reliability through redundancy.
- In addition the SLB device provides application health-check to distribute the workload in a very intelligent fashion.
- By doing deep stateful inspection, the SLB service improves the security at the application level.

In order to accomplish these functions at line rate, the load balancing service is usually provided by dedicated hardware device. More in line with Cisco ACE.

Layer 2 refers to the Data-Link layer. The Data Link Layer is the protocol layer, which transfers data between nodes on the same local area network (LAN) segment. For our purposes the data link protocols refers to Ethernet for local area networks (multi-node). Layer 2 is by definition a non-routable protocol and is confined in the same POD. However it can be encapsulated into a Layer 2 VPN tunnel over a Layer 3 network to be extended between multiple data centers. More in line Ethernet.

Layer 3 or Network Layer is responsible for routing packets delivery including routing through intermediate routers. Traditionally interconnection between multiple sites and access into a data center is achieved using the routed layer 3 network. More in line Layer 3.

WANs are used to connect LANs and other types of networks such as data center or campus together, so that users and computers in one location can communicate with users and computers in other locations.

More in line with Cisco Docwiki Internetworking Technology Handbook

Infrastructure as a Service (IaaS). The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).
Providing Infrastructure Services requires a scalable and elastic network that can be segmented per customer supported by the Data center interconnect (DCI).

Virtual desktop infrastructure (VDI) allows IT departments to centrally host and manage user desktops on VMs residing in the data center. The users can access their server-hosted virtual desktop from anywhere there is connectivity and from a broad range of end-user devices using one of several remote display protocols. VDI decouples the user's desktop computing environment from the physical hardware (PC, laptop computer, etc.). They are hosted on a data center server VM and delivered across the network using a remote display protocol. The end device no longer stores the user's applications or data, which are instead housed in centralized storage residing in the data center.

Cisco **Virtualization Experience Infrastructure (VXI)** enhances desktop virtualization with the integration into the Cisco Unified Communication as well as network optimization services such as WAAS while the network access is maintained secured. More in line with Cisco VXI.

Unified Communication-as-a-Service (UCaaS) is based on a hosted or on-demand service platform that allows the end-user to integrate all of the enterprise communication tools without the necessary capital expenditure.

A **storage area network** (**SAN**) is a dedicated storage network that provides access to consolidated, block level storage. SANs primarily are used to make storage devices (such as disk arrays and tape libraries accessible to servers so that the devices appear as locally attached to the operating system. A SAN typically has its own network of storage devices that are generally not accessible through the regular network by regular devices.

Network-attached storage (NAS), in contrast to SAN, uses file-based protocols such as NFS or SMB/CIFS where the storage is remote, and computers request a portion of an abstract file rather than a disk block.

Wavelength-division multiplexing (**WDM**) is a technology which multiplexes a number of optical carrier signals onto a single optical fiber by using different wavelengths (colors) of laser light. This technique enables bidirectional communications over one strand of fiber, as well as multiplication of capacity. Dense wavelength division multiplexing (DWDM), refers originally to optical signals multiplexed within the 1550 nm band.

Multiprotocol Label Switching (MPLS) is a highly scalable, protocol agnostic, data-carrying mechanism. In an MPLS network, data packets are assigned labels. Packet-forwarding decisions are made solely on the contents of this label, without the need to examine the

packet itself. This allows one to create end-to-end circuits across any type of transport medium, using any protocol. The primary benefit is to eliminate dependence on a particular Data Link Layer technology, such as Ethernet, and eliminate the need for multiple Layer 2 networks to satisfy different types of traffic. More in line with Wikipedia MPLS.

Ethernet over MPLS (EoMPLS) technology leverages an existing MPLS backbone network to deliver Transparent LAN Services (i.e. TLS) based on Ethernet connectivity to the customer site. The concept of Transparent LAN Services is the ability to connect two Ethernet networks, which are geographically separate, and have the two networks appear as a single logical Ethernet or VLAN domain. VLAN transport capability allow service providers or an Enterprise to extend VLAN networks in different locations at wire speed.

Virtual Private LAN Service (**VPLS**) is a way to provide Ethernet-based multipoint-to-multipoint communication over IP/MPLS networks. It allows geographically dispersed sites to share an Ethernet broadcast domain by connecting sites through pseudo-wires.

VPLS is a virtual private network (VPN) technology. VPLS allows any-to-any (multipoint) connectivity. In a VPLS, the local area network (LAN) at each site is extended to the edge of the provider network. The provider network then emulates a switch or bridge to connect all of the customer LANs to create a single bridged LAN.

The differences between VPLS and H-VPLS are related to the scaling attributes of each solution. By contrast, H-VPLS partitions the network into several edge domains that are interconnected using an MPLS core. More in line with Cisco VPLS.

Secure Socket Layer (SSL) is a cryptographic protocol that provide a secured and encrypted communication over Internet. More in line SSL.

First Hop Redundancy Protocol (FHRP) IP routing redundancy is designed to allow for transparent fail-over at the first-hop IP router.

HSRP, **GLBP** and **VRRP** enable two or more devices to work together in a group, sharing a single IP address, the virtual IP address. The virtual IP address is configured in each end user's workstation as a default gateway address and is cached in the host's Address Resolution Protocol (ARP) cache. More in line with Cisco FHRP.

For More Information

- Design Zone for Data Centers (includes DCI, security, Application, VMDC..)
  - Design Zone for Data Center
  - High Scale Data Center Interconnect
- Data Center Designs: Business Continuance and Disaster Recovery:
  - Business Continuance and Disaster Recovery
- Data Center Interconnect (including vPC, OTV, VPLS):
  - http://www.cisco.com/go/dci
- Workload Mobility (includes DWS, EMC, NetApp):
  - Workload Mobility
- Data Center Application Networking Services:
  - Data Center Application Networking Services
- Global Site Selector (GSS) :
  - Global Site Selector
- FabricPath/TRILL
  - Scaling Data Centers with FabricPath and the Cisco FabricPath Switching System
  - Cisco FabricPath at a Glance (pdf)
- Locator/ID Separation Protocol (LISP):
  - LISP main page
  - LISP4.cisco.com
- Cloud Computing:
  - Cisco Cloud
- Nexus 1000v
  - Nexus 1000v
- VXLAN
  - VXLAN Cisco Main page
  - IETF vxlan-00 draft