DATA CENTER CONVERGED SOLUTIONS DESIGN GUIDE

APPLICATION NOTE

AT THE SPEED OF IDEAS™

Alcatel·Lucent
Enterprise

TABLE OF CONTENTS

ABSTRACT

Data centers are not just next-generation network upgrades. They signify a transformation from an information technology (IT)-centric infrastructure to a service-centric infrastructure. But to meet that objective, all components, including servers, storage, network and applications are virtualized to cost-effectively deliver computational elasticity, and data and application serviceability.

Based on business objectives and the type of cloud applications deployed, the key goals for any data center architecture are:

- Deterministic latency
- Redundancy/high availability
- Manageability
- Scalability

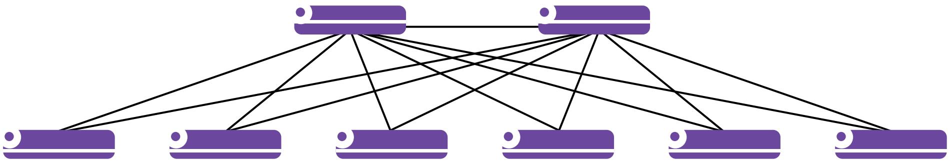
Alcatel-Lucent Enterprise offers a broad range of solutions, each addressing the fundamental requirements of data center networks, for both today and tomorrow. This document provides guidelines to design solutions to meet the needs of any organization.

1 TRADITIONAL DATA CENTER ARCHITECTURE MODELS

1.1 Spine-Leaf

This is a two-tiered networking design. The main building blocks are leaves and spines. Spines forward traffic along optimal paths between nodes at Layer 2 or Layer 3, while leaves control the flow of traffic between locally connected servers. Cross-sectional interconnect bandwidth can be improved though link aggregation groups (LAGs), or by employing Layer 3 equal-cost multipath routing (ECMP) for multipath. There is single-hop latency for server-to-server communication within the leaves. Additional latency is a factor when traffic needs to bridge the spine for communication, with a maximum of three hops for any-to-any communication.

Figure 1. Spine-leaf network design



An oversubscription ratio of 1:1 would be ideal, but that is unachievable due to budget, space and power constraints. In spine-leaf architecture, the oversubscription ratio is determined as follows:

[Number of downlink ports * downlink speed] / [Number of uplink ports * uplink speed]

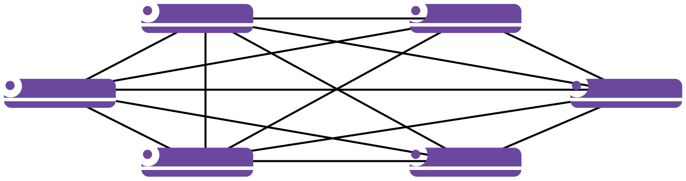
Typical networks are designed with oversubscription levels ranging from 2:1 to 6:1.

Scalability is limited by the number of ports available at the spine layer.

1.2 Mesh

A mesh network is flat and employs one of two connection topologies: full or partial mesh. In full mesh topologies, each node is connected directly to every other node. In partial mesh topologies, some nodes are connected to all the others, but others are connected only to nodes they exchange most data with. A mesh network scores high on reliability: if one node can no longer operate then the rest can still communicate with each other, directly or through one or more intermediate nodes. The main drawback is expense: a large number of cables and connections are required to form a full mesh as the number of nodes increases.

Figure 2. Mesh network design.



There is single-hop latency for server-to-server communication within a node. Additional latency is a factor when traffic needs to traverse the mesh, a maximum of two hops for any-to-any communication in full mesh architecture.

In full mesh architecture, the oversubscription ratio is determined as follows:

[Number of server ports * downlink speed] / [Number of uplink/mesh ports * uplink speed]

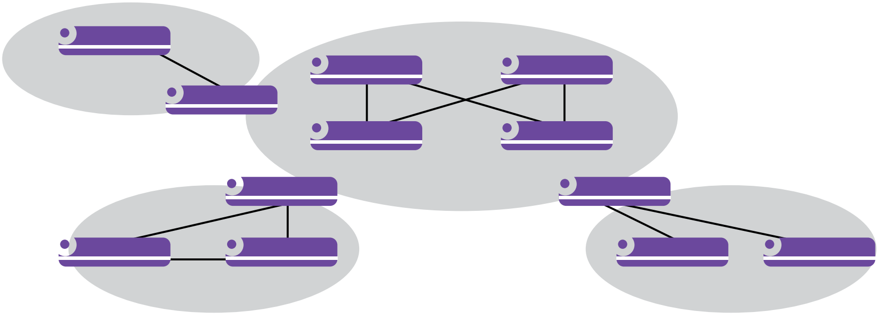
Typically, full mesh designs have low oversubscription factors, as each node is directly connected to its neighbor.

Scalability is limited by the need to build either a full or partial mesh. As the number of nodes increases, the ratio of server-facing ports to ports forming a mesh is reduced.

1.3 Layer 3-Routed

The network architecture of most data centers is a multilayered core-distribution-access setup. Servers at the access layer have redundant links for reachability. For maximum efficiency in existing networks, ECMP is used to achieve full proficiency. As the network expands, flooding occurs due to the burgeoning network, which makes the control exchange between the routers difficult and causes congestion. This issue can be resolved by segmenting the network into logical domains, which then limits the routing information exchanged in the entire area. This configuration makes the network scalable, but also highly segmented, requiring extensive management.

Figure 3. Layer 3-routed network design



Routed networks are not optimal for low latency. For any-to-any communication, the traffic may have to traverse many nodes residing in different domains. Latency is not deterministic.

It is difficult to estimate the oversubscription ratio in such networks. General principles cannot be employed, as they are in either spine-leaf or mesh architectures. The percentage of server traffic limited within the domain/area and what percentage of server traffic is allowed to cross domains are important factors in defining and constructing oversubscription requirements.

Scalability is unlimited, as additional domain/areas can readily be connected to existing networks. High levels of IT intervention are needed for either additional provisioning or to change existing setups.

Table 1. Pros and cons of spine/leaf, mesh and Layer 3-routed network design

	SPINE/LEAF	MESH	LAYER 3-ROUTED
Pros	• Layer 2/Layer 3 common fabric implementation • Simpler design • Fewer interconnects • Easy to scale within boundary	• Layer 2/Layer 3 differentiated fabric • Modular design • Highly scalable • No transit hops, resulting in lower latency and lower over-subscription ratios	• End-to-end routed fabric • Easy to secure at IP layer • Fewer interconnects • Easy to scale
Cons	• Scalability limited to number of ports in the spine layer • Additional layer of transit hop may impact latency and over-subscription	• More links used for interconnects	• Highly oversubscribed architecture • Number of transit hops is not deterministic, impacting latency • Complex to design and maintain

2 ALCATEL-LUCENT ENTERPRISE DATA CENTER ARCHITECTURES

Traditional multi-tier architectures were designed to facilitate clear demarcations among different IT zones (i.e., access, core and the data center), with each tier adding to latency, manageability and cost. Many of today's applications have very low delay tolerance, so the number of tiers has to be reduced to minimize system latency. A flat fabric simplifies management, reduces cost and allows resilient, low-latency networks to be designed. Flat architecture concepts supported in the Alcatel-Lucent Operating System (AOS) are described below. These solution architectures provide high availability, deterministic low latency and can scale up or down with demand. Each solution is tightly integrated with the OmniVista™ 2500 Virtual Machine Manager (VMM), providing a unified platform for virtual machine visibility and provisioning with virtual network profile (vNP) across the network, allowing seamless vMotion.

Table 2. Flat architecture concepts supported in the AOS

SERVICE/CAPABILITY	MCLAG	VIRTUAL CHASSIS	VIRTUAL CHASSIS AND SPB
Single pane management	No	Yes	Yes
Dual Homing	Yes	Yes	Yes
Signaling transfer point independence	Yes	Yes	Yes
Link/Node Resiliency	Yes	Yes	Yes
Mesh	Yes[1]	Yes[2]	Yes
Priority-based flow control	No	Yes	Yes
Data center bridging exchange	No	Yes	Yes
Edge virtual bridging	No	Yes	Yes
Virtual Ethernet port aggregator	Yes	Yes	Yes
Virtual network profile	Yes	Yes	Yes
Virtual machine manager	Yes	Yes	Yes
Layer 2 virtualization	Yes	Yes	Yes
Layer 3 virtualization	Yes	Yes	Yes[3]

[1] Only two MCLAG groups can be connected back to back in mesh. This will only provide Layer 2 virtualization.
[2] Only two VC groups can be connected back to back in mesh. This will provide both Layer 2/Layer 3 virtualizations.
[3] Under development; functionality will be available in a subsequent release.

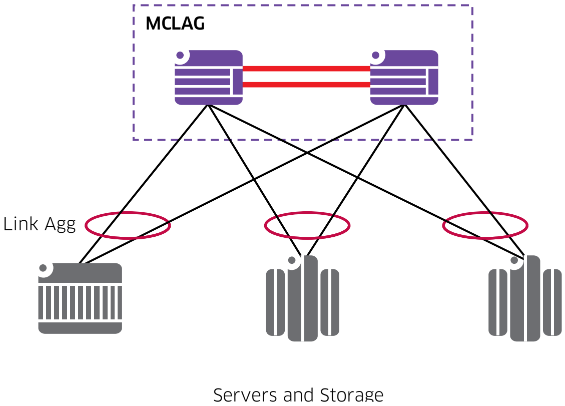
Table 3. OmniSwitch Product Port Capacity

PLATFORM	OMNISWITCH 6900	OMNISWITCH 10K
10G	64	256
40G	6[a]	64[b]
10G VFL	Yes	Yes
40G VFL	Yes	Yes

[a] 40 10G ports are fixed. Two OS6-QNI-U3 pluggable modules will provide six 40G ports.
[b] OS10K-QNI-U8 module provides eight 40G interfaces. Only two VC groups can be connected back to back in mesh.

2.1 Multi-Chassis Link Aggregation Group (MCLAG)

Figure 4. MCLAG network design



This configuration facilitates the dual-homing of servers/storage and access devices with links distributed across the MCLAG switches. There is no logical loop between the edge devices and multi-chassis peer switches, even though a physical loop exists. Single interface servers, storage and edge devices can be connected to any MCLAG switch. Alternatively, dual-home connections can be established with only one MCLAG switch, if node resiliency is unimportant.

Only two switches with unique chassis IDs and the same group IDs can form MCLAG peering. The administrator must ensure all MCLAG configurations are consistent across the two peers or the MCLAG will remain non-operational. Independent of Spanning Tree Protocol (STP), AOS implements a proprietary loop-detection mechanism. Periodically, a multicast Protocol Data Unit (PDU) is flooded out on the virtual fabric link (VFL) and MCLAG primary ports. Loop Detection is flagged when the PDU is returned to the transmitting peer, triggering the following actions:

- A log message is sent for loop detect event
- A SNMP trap is generated, and
- The offending port is shut down.

Layer 2-virtualization: Hosts belonging to the same IP subnet communicate at Layer 2, independent of Layer 3 lookup. Any non-unicast frame received on MCLAG port when flooded to the remote MCLAG peer via VFL will not be flooded back to any MCLAG ports in that virtual local area network (VLAN).

LAYER 3-virtualization: Virtual IP (VIP) Interface infrastructure common to the MCLAG peers facilitates local routing between hosts in different subnets connected to MCLAG ports. The two peers are independent routers and don't synchronize routing information.

Each MCLAG peer is an independently managed entity. If one of the primary Chassis Management Module (CMMs) on either of the MCLAG peers fails causing takeover, Layer 2 and Layer 3 traffic will not be affected. However, if the primary MCLAG peer fails, there will be an outage to Layer 2 traffic of less than a second, but Layer 3 traffic may experience a longer outage due to reconvergence of the Address Resolution Protocol (ARPs). To minimize the impact on Layer 3, MAC-retention needs to be enabled on MCLAG peers so that system MAC change will not occur upon node failure.

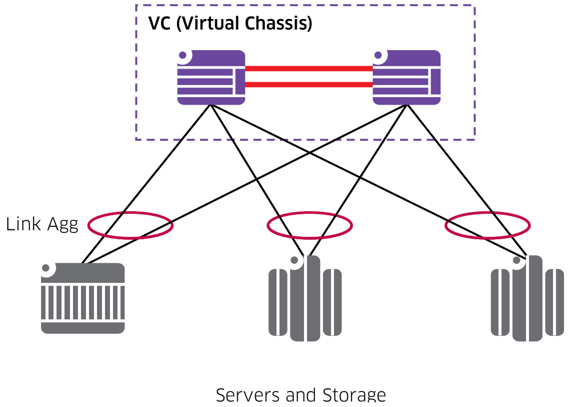
Split-brain: This occurs when the VFL interconnect breaks, but the MCLAG peers are operational. In the case of MCLAG, this can cause Layer 2 flooding in the network.

MCLAG benefits:

- Allows multi-chassis terminated link aggregation groups to be created
- Creates a loop-free edge without STP
- Provides node- and link-level redundancy
- Enables switches to be geo-independent (i.e., don't need to be co-located)
- Inter-connects switches using standard 10G and 40G Ethernet optics
- Supports redundancy and resiliency across the switches

2.2 VC (Virtual Chassis)

Figure 5 VC network design



Similar to MCLAG, this configuration facilitates the dual-homing of servers/storage and access devices with links distributed across virtual chassis (VC) switches. Single interface servers, storage and edge devices can either be connected to any VC switch, or dual-home connections can be established with only one VC switch, if node resiliency is not important. A given switch can only participate in one VC domain group. VC, unlike MCLAG, offers a single management pane for Layer 2/IP configuration. Unlike MCLAG, where STP is disabled on MCLAG ports, it is recommended STP be desabled, but it is optional. In the absence of STP, users can enable the proprietary loop detection feature, which periodically transmits a multicast PDU on the primary LAG port/VFL. Loop detection is flagged when the PDU is returned to the transmitting peer, triggering the following actions:

- A log message is sent for loop detect event
- A SNMP trap is generated, and
- The offending port is shut down.

Layer 2-virtualization: Hosts belonging to the same IP subnet communicate at Layer 2, independent of Layer 3 lookup. Any non-unicast frame received on the LAG port when flooded across the VFL will not be flooded back to any LAG port in that VLAN.

Layer 3-virtualization: Unlike MCLAG, there is no need for VIP VLAN/IP infrastructure. The VC peers, including the primary, synchronize all Layer2/Layer3 information, facilitating local routing between hosts in different subnets connected to VC.

There is an Ethernet Management Port (EMP) port on each VC peer for out-of-band management, but the primary VC chassis is the only centralized point for all IP management. If one of the primary CMMs on either of the VC peers fails, causing takeover, Layer 2 and Layer 3 traffic will be unaffected. However, if the primary VC chassis fails, there will be an outage to Layer 2 traffic of under a second and Layer 3 traffic may experience a longer outage due to flush/relearning of the ARPs. To minimize the Layer 3 impact in such circumstances, MAC-retention needs to be enabled on VC peers so that system MAC change will not occur upon node failure.

Split-brain: This occurs when the VFL interconnect breaks but the VC peers are operational. This can cause a LAYER 2/LAYER 3 storm in the network. AOS uses a proprietary, out-of-band management protocol on the EMP port that detects the operational health of the remote peer. If all VFL links go down, then this protocol will detect and shutdown all user ports on the remote peer to avoid loop. The user ports will automatically come up when VFL connectivity is reestablished.

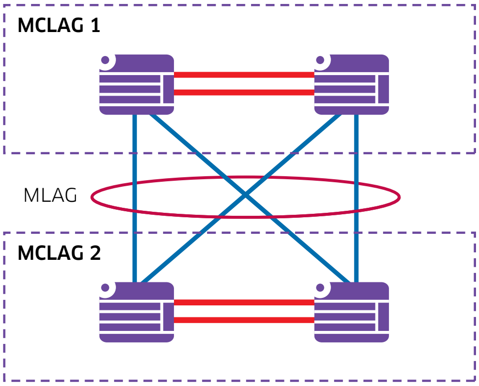
Lossless: VC supports data center bridging exchange/edge virtual bridging (DCBX/EVB) enhanced protocols to automate edge server provisioning and end-to-end lossless path configuration.

Benefits of VC:

- Enables a single point of management, via single IP address
- Provides a centralized control plane for routing and bridging
- Allows multi-chassis terminated link aggregation groups to be created
- Creates loop-free edge without STP
- Provides node-level and link-level redundancy
- Enables the switches to be geo-independent (don't need to be co-located)
- Switches inter-connected using standard 10G and 40G Ethernet optics
- Supports redundancy and resiliency across the switches
- Supports redundancy and resiliency on the VFL used to inter-connect the switches
- Supports full routing, similar to single chassis, over the dual-homed link aggregates
- Enables In-Service Software Upgrade (ISSU) to operate across the chassis
- Prevents split brain loops by using the EMP port for out of band VC control

2.3 MCLAG Mesh

Figure 6. MCLAG mesh network design

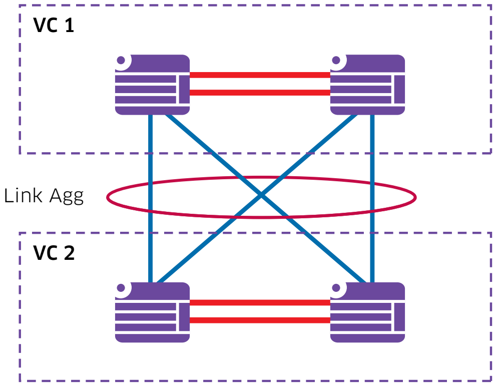


This configuration represents two MCLAG groups connected back-to-back in a mesh. The mesh is limited to two groups. It provides similar benefits and control points as a single MCLAG group. The two groups are connected via MLAG; they cannot be routed interfaces, which could result in the peer MCLAG group apparently talking to a single neighbor, since two physical connections go over the same MCLAG layer 2 logical link. As a result, the peer group may establish an adjacency and exchange routing information with only one of the switches. That would cause only one of the MCLAG chassis to learn about the routes from the upstream router and connectivity would not occur through both multi-chassis peers.

The scalability of a back-to-back MCLAG almost doubles for Layer 2 virtualization.

2.4 VC Mesh

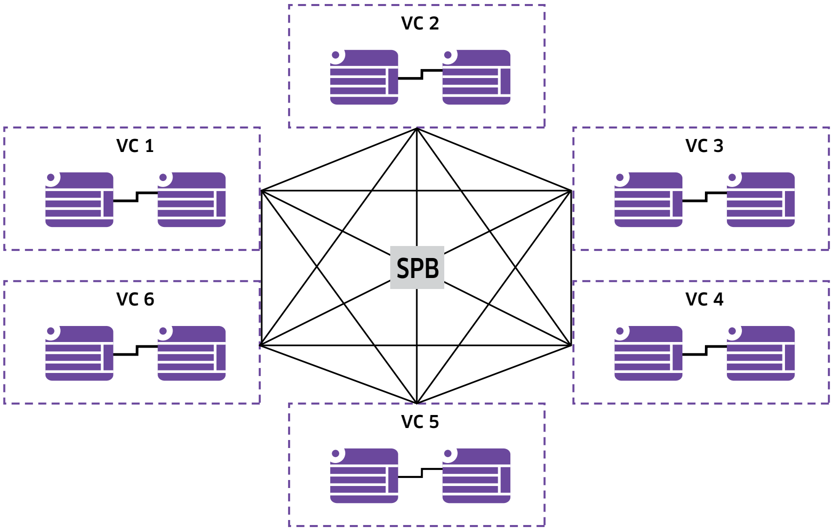
Figure 7. VC mesh network design



This configuration represents two VC groups connected back-to-back in a mesh. The VC mesh is limited to two groups without shortest path bridging (IEEE 802.1aq) (SPB). It provides exactly the same benefits and control points as a single VC group. Unlike MCLAG back-to-back mesh, a VC back-to-back mesh provides Layer 2 and Layer 3 virtualization. The scalability of a back-to-back VC almost doubles compared to a single VC of two.

2.5 VC and SPB Mesh

Figure 8. VC and SPB mesh network design



SPB is a standard protocol developed as an alternative to STP for resilient and scalable connectivity service in the datacenter network. SPB, unlike STP, better uses and distributes traffic in a redundant mesh environment. The SPB-M implementation reduces the need for large MAC address tables in the network core, and increases use of the network over STP. SPB-M supports virtual private network (VPN)-like provisioning of the network, ensuring the data center is secure and allowing multi-tenant configurations. The service instance limitation of 4000 SVLANS is addressed by assigning a 24-bit instance number (I-SIDs) to the customer frames by the Backbone Edge Bridges (BEBs).

A SPB-M network operates on a special set of VLANS called B-VLANs. AOS will support a maximum of 16 SPBM B-VLANs. The SPB network port must be configured as an ISIS-SPB interface. Whenever a network port is configured, all SPBM B-VLANs configured in the system will automatically be added to the port. All nodes must have the same B-VLANs and corresponding ECT parameters for the adjacency to come up. AOS design does not restrict the configuration of other types of VLANs (802.1q or 802.1ad) on the network port.

Customer traffic enters a BEB through access ports. Whenever an access port is created, it is assigned to the default VLAN, 4095. No other usable VLANs (1-4094) can be configured on an access port. These ports are not regular bridging ports and there are no VPAs associated to them. Instead, they are specifically provisioned to accept customer traffic and assign them to a configured service instance. The classification of traffic to services can be specified by creating service access points (SAPs) on the access ports. Each access port has a port profile associated to it. This profile will specify the desired behavior for handling protocol packets received on the port. Packets could be dropped, flooded on the service instance on which it was classified or trapped to CPU and sent to a peering protocol. Both link aggregate ports (static and Link Aggregation Control Protocol-based) and regular Ethernet ports can be configured as an access port.

Dual-homing: Standard 802.3ad LAG-based connectivity for node resiliency is possible on VC nodes. However, either dual home servers/storage with NIC bonding capability (when only one link is active at any given time) or edge switches with proprietary AOS dual-home link (DHL) active-active capability can dual home to standalone SPB nodes.

Layer 2-virtualization: Hosts belonging to same service instance communicate at Layer 2.

Layer 3-virtualization: Under development; functionality will be available in a subsequent release.

Lossless: SPB network supports DCBX/EVB enhanced protocols to automate server provisioning and end-to-end lossless path configuration.

SPB Benefits:

- Offers a multipath loop-free flat fabric
- Makes the VLAN agnostic: VLAN only has local port significance
- Creates up to 16 ECMP paths
- Enables bidirectional deterministic /predictable forwarding (synchronous path)
- Allow for free-form pod/mesh topologies
- Uses existing ISIS protocol and interoperates with existing service provider equipment
- Provides fast, sub-second convergence
- Enables graceful overload control
- Allows each ISID to have 4K VLANs/SVLANs
- Bridges VLANs through the I-SID
- Supports VLAN translation: implicit to SAP configuration
- Allows 1000 ISIDs/BEB
- Scales upto 1000 nodes

3 OMNISWITCH ADDITIONAL BENEFITS

3.1 Virtual Network Profile

The vNP is a universal network profile (UNP) associated with one or more ports. It resides on the OmniSwitch and includes information such as:

- Provisioning requirements
- Access control rights
- Either VLAN or service assignment
- Expected quality of service (QoS) levels
- Application priority

With this knowledge, the vNP can manage applications as services. This unique AOS technology helps VMs to securely bind to a network with service guarantees. The data center backbone can either be a VLAN bridged network or a service domain network. Once a VM is assigned to a vNP, VM traffic is bridged on the classified VLAN, guaranteeing the QoS policies tied to the profile.

VMs in either VLAN bridged domains or in service domains can run on hosts with and without EVB enabled (see following subsection). This results in four types of data center edge ports:

1. UNP-enabled bridge port
2. UNP-enabled service access port
3. EVB-enabled bridge port
4. EVB = enabled service access port

A UNP enabled bridge port can use any of the following parameters to define classification policy:

- MAC
- MAC-range
- MAC + VLAN
- MAC-Range + VLAN
- IP
- IP + VLAN
- VLAN

In addition to these classification rules, VMs can be provisioned into the VLAN tag preassigned by the hypervisor with UNP trust-tag functionality. Each UNP port can be assigned a default-UNP to classify untagged traffic when all classification schemes either fail or are absent.

The classification profiles differ for a VLAN domain and a service domain. VMs in VLAN bridged networks are bound to profiles comprising the VLAN and associated QoS policy list.

Upon enabling “unp dynamic-vlan-configuration”, the switch will create the VLAN dynamically, if absent.

Upon enabling “unp dynamic-profile-config”, the switch will create the vNP profile dynamically. These configurations can be saved and modified by the administrator.

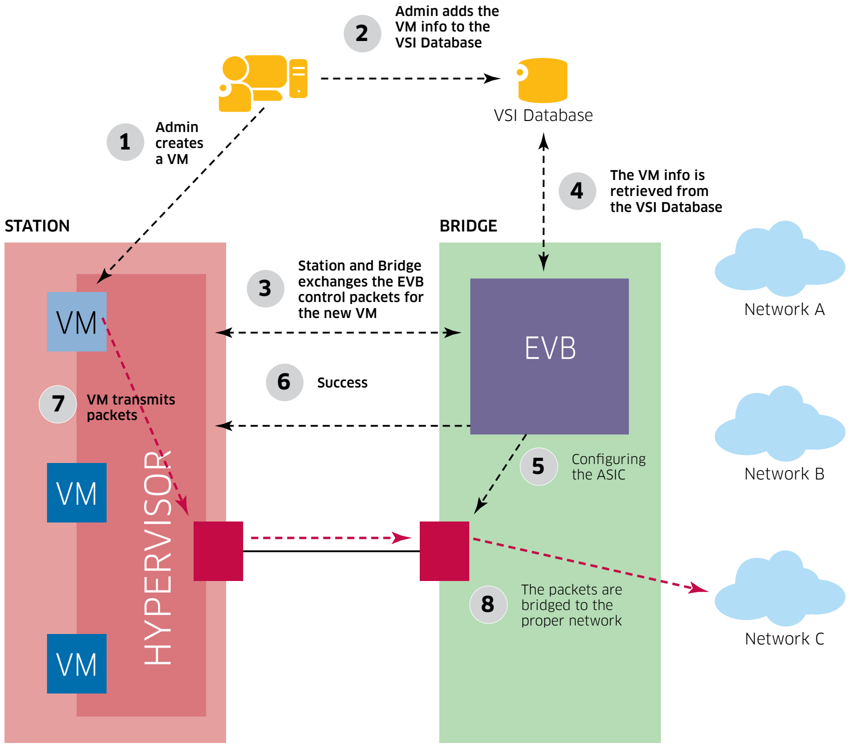
A UNP-enabled service access port uses the same classification policy as a bridge port, but the classification profile is tied to services. Based on the classification policy, the UNP profile will determine the service and the C_VLAN ID into which the VM will be classified.

If the expected service does not exist, and the system level “dynamic-service-configuration” is enabled, and the profile has the information required to create the service, UNP requests the Service Manager to dynamically create this service on the switch.
If the expected service cannot be created, either because there are insufficient system resources or the selected profile does not have the information to create a service or the system level dynamic-service-configuration is not enabled, traffic from VM is discarded through filtering engine.

3.2 EVB (Edge Virtual Bridging, IEEE 802.1Qbg) for Fabric Automation

EVB helps automate the discovery of the VMs and connect them to the proper network domains, either a VLAN bridged network or a BBS service network. When VLAN bridging network is used, AOS EVB does not support GroupID and S-Channel. EVB creates a dynamic EVB VLAN/ VPA, and MVRP can be used to propagate the EVB VLANs to other switches. When SPB service network is used, EVB creates a SAP for each C-VLAN in an Edge Relay (ER). Each SAP is then associated with one of the SPB service instance.

Figure 9. High level overview of the EVB operation


2
Admin adds the VM info to the VSI Database
VSI Database
1
Admin creates a VM
4
The VM info is retrieved from the VSI Database
STATION
BRIDGE
3
Station and Bridge exchanges the EVB control packets for the new VM
VM
EVB
Network A
6
Success
7
VM transmits packets
5
Configuring the ASIC
Network B
VM
HYPERVISOR
8
The packets are bridged to the proper network
Network C
VM

In the data center environment, data center servers run specialized hypervisor software which helps instantiate multiple VMs within a server. These VMs can be dynamically created, deleted or even moved to other servers in the network. Each VM may require different network connections and services. Usually, similar types of VMs are grouped together to form a VM network to control the unicast and multicast domains, and to make a connection to storage area networks (SAN) or wide-area networks (WAN).

Each ER (vSwitch) within the hypervisor can operate in one of two modes: (Virtual Ethernet Port Aggregator (VEPA) or Virtual Ethernet Bridge (VEB).

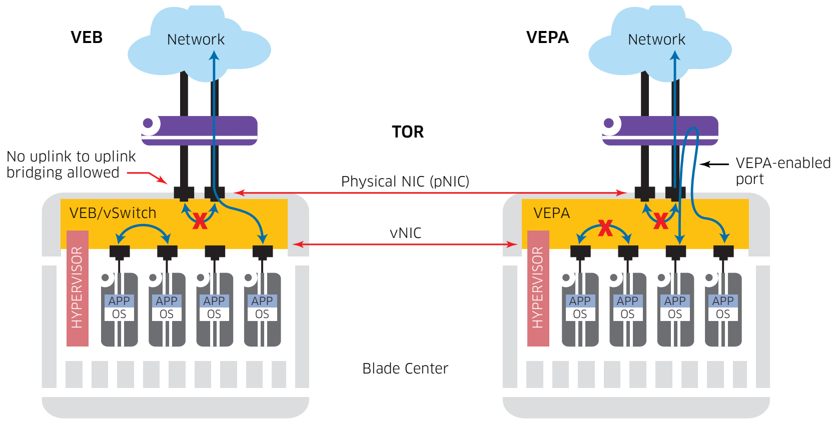
With VEPA, the vSwitch treats the network interface card (NIC) as a singular interface connected to each VM. All outgoing traffic is sent through the NIC to an external switch.

- VEPAs typically don't learn
- The intent is to use the intelligence of the adjacent switch
- VM <-> VM switching is hair-pinned through the bridge for policy, access control lists, security, etc…

With VEB, the vSwitch creates a virtual Ethernet switch inside the host that can bridge data between VMs and send data out to the network as needed.

- VEBs do learn, allow local switching
- The intent is to allow trusted VMs to switch directly
- VM <-> VM switching is allowed, but not uplink <-> uplink, to prevent loops

Figure 10. EVB Mode Operations



VEPA a Virtual Port Capabilities

Table 4. OmniSwitch Interface VEPA/Virtual Port scalability

BOARD TYPE		EVB MODE VEPA	NO. OF VP
OmniSwitch 10K	GNI-C48	Can be enabled per Port	4000
	XNI-U32	Does not support SPB	
	XNI-U16 QNI-U8/U4	Can be enabled per VP	8000
OmniSwitch 6900	XNI-20 XNI-40	Can be enabled per Port(2)	4000

Instead of using typical VLAN and Layer 2 bridging protocol in the data center network, a special backbone bridge service (BBS) can be used to provision the backbone with protocols such as SPB, VPLS, and TRILL.

EVB runs between the EVB station and EVB bridge. It allows the EVB bridge to discover the VMs in the EVB Station and connects the VMs to the proper BBS service instance. This process can be fully automated with the help of the VSI Database.

In AOS, the VSI Database is managed and maintained by the UNP application. The VSI type database is comprised of the following fields:

- VSI manager ID
- VSI type
- VSI type version
- VSI instance ID
- VLAN ID
- VLAN priority
- Group ID
- MAC address

EVB-enabled ports rely on the EVB protocol to exchange key VM (VSI) information. UNP parses the VSI manager ID with a three-field key (VSI type; VSI type version; VSI instance ID).

On an **EVB bridged port**, only the VLAN ID from the database is matched with the VLANTag pre-assigned by the hypervisor with UNP trust-tag functionality. The VM is classified into the matching profile VLAN and associated QoS policy list.

By enabling “dynamic-vlan-configuration”, the switch will create the vlan dynamically if not present.

By enabling “dynamic-profile-config”, the switch will create the vNP profile dynamically.

On an **EVB service access port**, the VLAN ID and Group ID fields are used to determine the matching profile. UNP will request the service manager to dynamically create this service on this switch if:

- the expected service does not exist;
- the system level dynamic-service-configuration is enabled; and
- the profile has the information required to create the service.

After confirming the expected service exists or is dynamically created, UNP checks if the virtual port (VP) with the expected C_VLAN ID (e.g., 1/1:100) exists. If a VP with the expected C_VLAN ID exists, UNP will check if this VP is connected to the expected service. If a VP exists but is associated to a service other than the expected service, UNP will request the service manager to create a new VP to associate this VP with the expected service. If no VP exists with the expected C_VLAN ID, UNP will request the service manager to create a VP with the expected C_VLAN ID and associate this VP to the expected service.

Traffic from VM is discarded though filtering engine if the service cannot be created, either because a) there is not enough system resource, or b) the selected profile does not have the information to create a service, or c) the system level dynamic-service-configuration is not enabled.

Upon availability of both the SAP and the service, UNP requests the service manager to associate the MAC address to the VP for the service. UNP also requests service manager to learn the MAC on the SAP with the expected C_VLAN ID.

In summary, AOS supports the following host station/bridge configurations:

Case1: Station supports EVB, and bridge is configured with SPB

- The AOS EVB supports this case.
- The XNI-U32 of OmniSwitch 10K cannot run SPB.

Case2: Station supports EVB, and bridge is not configured with SPB

- When SPB is not used, it is assumed to be VLAN bridged.
- GroupID cannot be used.
- It supports only single ER per port (No S-channel or S-VLAN).
- MVRP can be used to propagate dynamic VLANs and VPAs created by the AOS EVB.

Case3: Station does not support EVB, and bridge is configured with SPB

- The tagged or untagged frames from the station need to be bridged to the proper SPB service instance.

Case4: Station does not support EVB, and bridge is not configured with SPB

- This is a typical LAYER 2 network. UNP solution will cover this case.

3.3 Application (VM) Visibility

OmniVista 2500 provides a single pane of management for VMs across the network. It provides visibility of VMs and their point of association to the network. This gives network administrator a dashboard that includes:

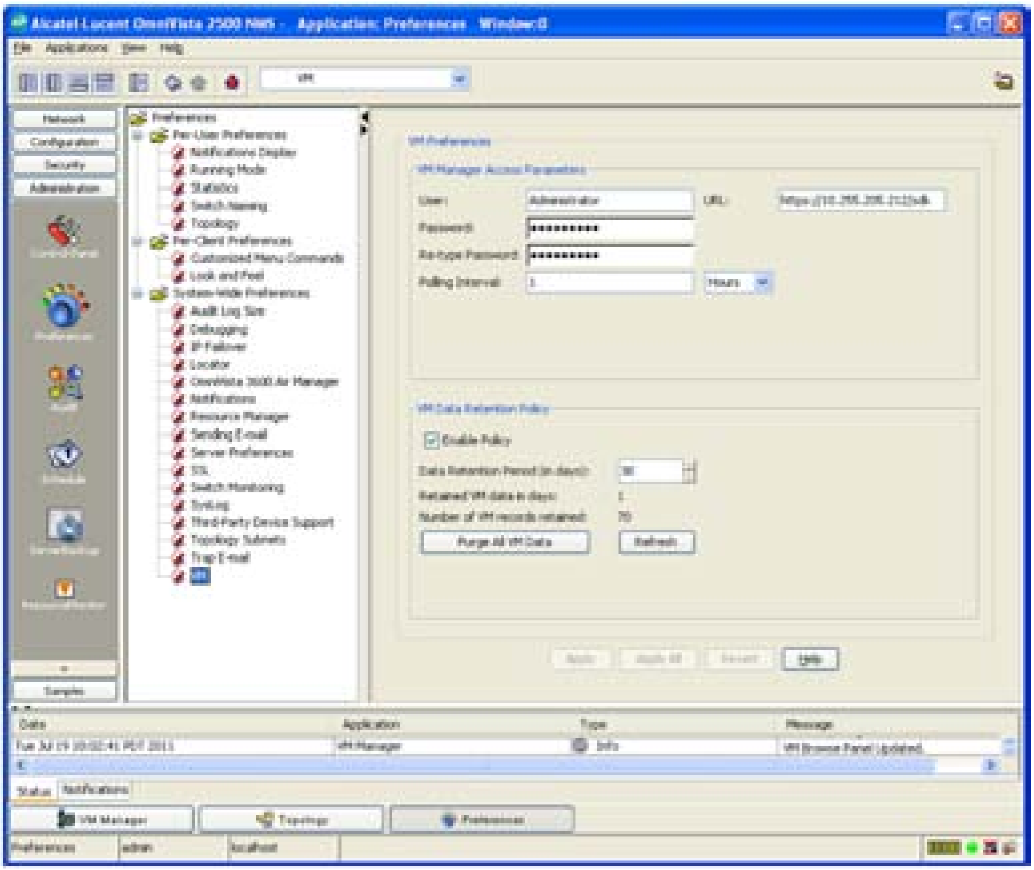
- Virtual machine locations
- Virtual machine types
- The switch ports to which the virtual machines are connected
- The duration of the connections
- Which vNP the virtual machine is using

It is immensely valuable for IT personnel to be able to narrow down whether an issue originates from the network, a server or an application. This is because the network is always the first suspect, due to either overloading or connectivity.

VMM is an embedded application within OmniVista, comprised of:

i) vCenter integration: VM discovery, VM polling and event listener service

Figure 11. VMM vCenter Access Policy Window



VM Preferences

VM Manager Access Parameters

User: Administrator URL: https://10.255.205.212/sdk

Password: ••••••••••

Re-type Password: ••••••••••

Polling Interval: 1 Hours

VM Data Retention Policy

☑ Enable Policy

Data Retention Period (in days): 30

Retained VM data in days: 1

Number of VM records retained: 70

Purge All VM Data | Refresh

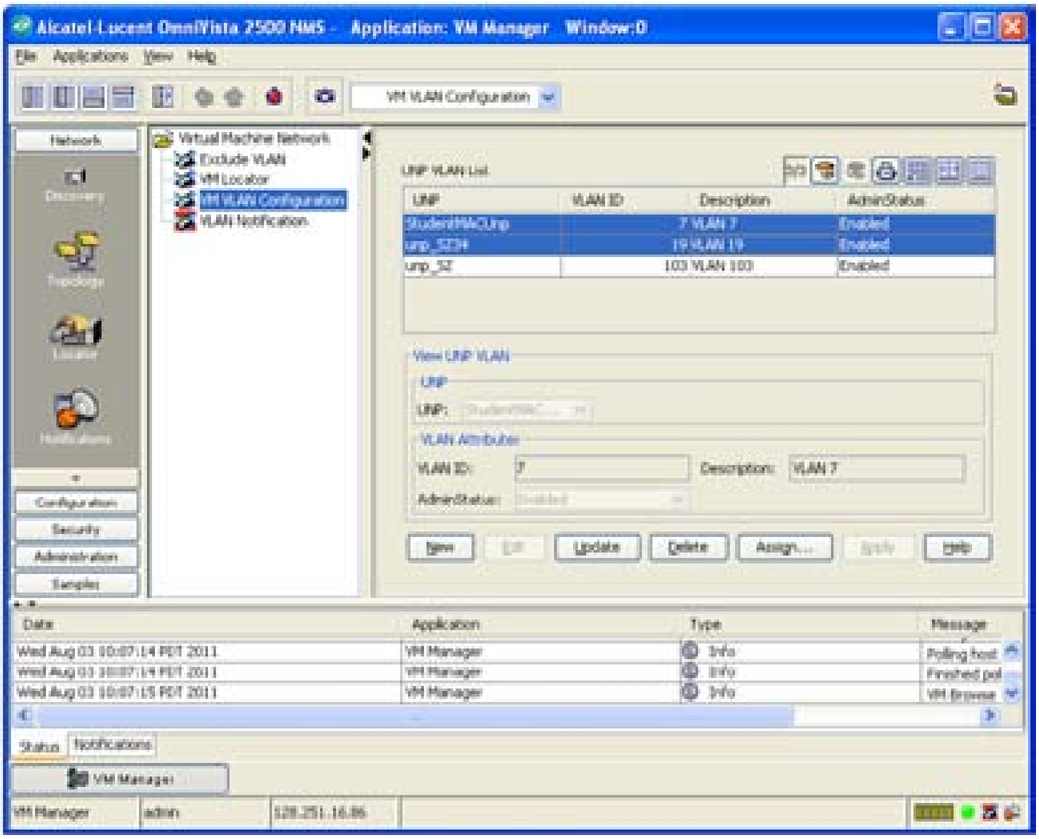
Apply | Apply All | Revert | Help

Date	Application	Type	Message
Tue Jul 19 10:02:41 PDT 2011	VM Manager	Info	VM Browse Panel Updated.

Preferences | admin | localhost

ii) vNP configuration: Profiles of VLAN and UNP configuration that specify the configuration associated with each VM VLAN. The profiles can be assigned to switches, as and when needed, using "Assign" button in this panel.

Figure 12. VMM Virtual Machine Vlan/vNP Configuration Window



UNP VLAN List

UNP	VLAN ID	Description	AdminStatus
StudentMACUnp	7	VLAN 7	Enabled
unp_SZ34	19	VLAN 19	Enabled
unp_SZ	103	VLAN 103	Enabled

View UNP VLAN

UNP

UNP: StudentMAC...

VLAN Attribute

VLAN ID: 7 Description: VLAN 7

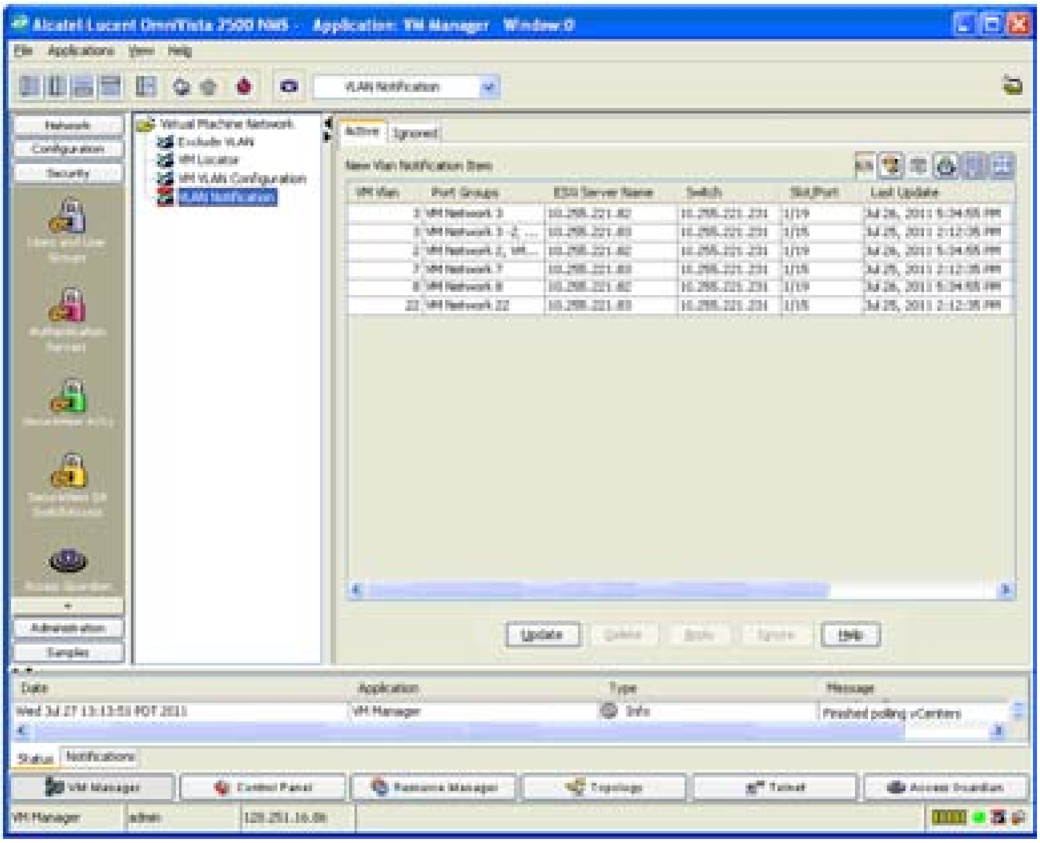
AdminStatus: Enabled

Date	Application	Type	Message
Wed Aug 03 10:07:14 PDT 2011	VM Manager	Info	Polling host
Wed Aug 03 10:07:14 PDT 2011	VM Manager	Info	Finished pol
Wed Aug 03 10:07:15 PDT 2011	VM Manager	Info	VM Browse

VM Manager | admin | 128.251.16.86

iii) VLAN Notification: This panel displays a list of VM instances where Network is missing some configuration to effectively handle VMs attached. The admin can proactively take action to correct missing configurations.

Figure 13. VMM Virtual Machine Vlan Notification Window



New Vlan Notification Item

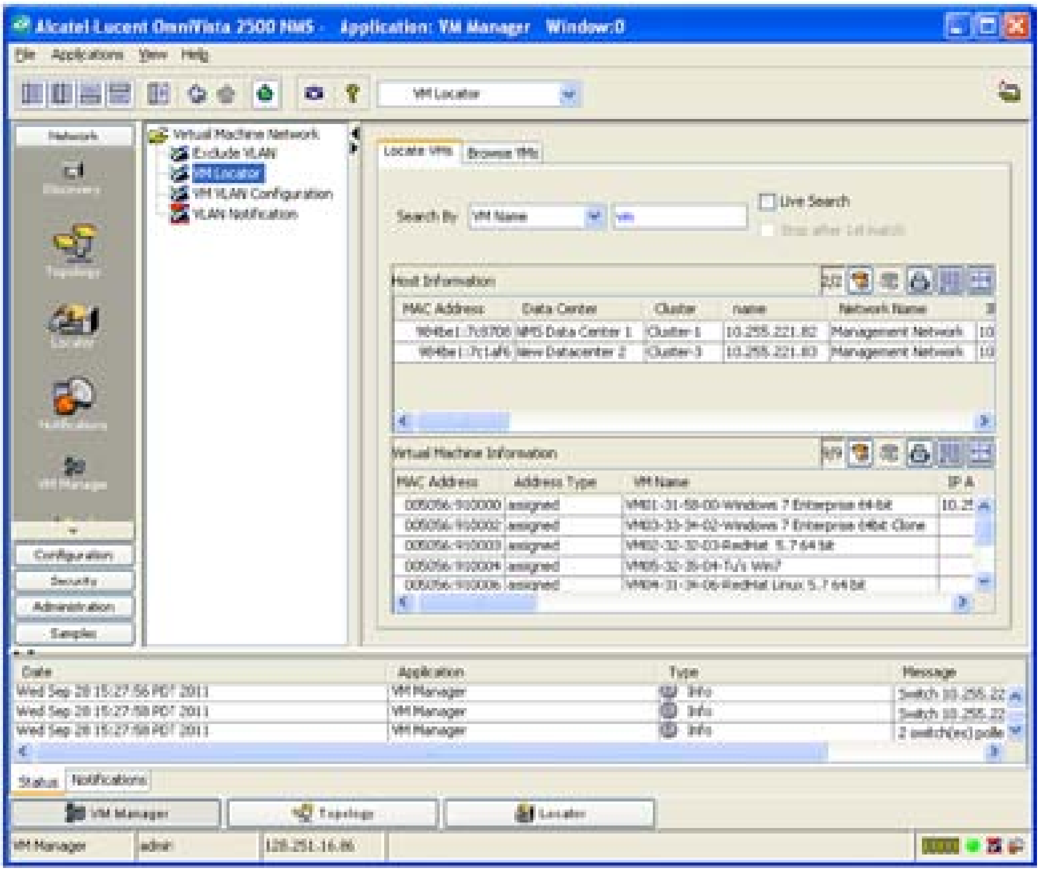
VM Vlan	Port Groups	ESX Server Name	Switch	Slot/Port	Last Update
3	VM Network 3	10.255.221.82	10.255.221.231	1/19	Jul 26, 2011 5:34:55 PM
3	VM Network 3 -2, ...	10.255.221.83	10.255.221.231	1/15	Jul 25, 2011 2:12:35 PM
2	VM Network 2, VM...	10.255.221.82	10.255.221.231	1/19	Jul 26, 2011 5:34:55 PM
7	VM Network 7	10.255.221.83	10.255.221.231	1/15	Jul 25, 2011 2:12:35 PM
8	VM Network 8	10.255.221.82	10.255.221.231	1/19	Jul 26, 2011 5:34:55 PM
22	VM Network 22	10.255.221.83	10.255.221.231	1/15	Jul 25, 2011 2:12:35 PM

Date	Application	Type	Message
Wed Jul 27 13:13:53 PDT 2011	VM Manager	Info	Finished polling vCenters

VM Manager | admin | 128.251.16.86

iv) OmniVista Locator: This provides the ability to search for specific VM using various search criteria, allowing browsing for VMs.

Figure 14. VMM Virtual Machine Network Locator Window



Search By: VM Name | vm

□ Live Search

□ Stop after 1st match

Host Information 2/2

MAC Address	Data Center	Cluster	name	Network Name
984be1:7c8708	NMS Data Center 1	Cluster-1	10.255.221.82	Management Network
984be1:7c1af6	New Datacenter 2	Cluster-3	10.255.221.83	Management Network

Virtual Machine Information 5/9

MAC Address	Address Type	VM Name	IP A
005056:910000	assigned	VM01-31-58-00-Windows 7 Enterprise 64-bit	10.2
005056:910002	assigned	VM03-33-34-02-Windows 7 Enterprise 64bit Clone	
005056:910003	assigned	VM02-32-32-03-RedHat 5.7 64 bit	
005056:910004	assigned	VM05-32-35-04-Tu's Win7	
005056:910006	assigned	VM04-31-34-06-RedHat Linux 5.7 64 bit	

Date	Application	Type	Message
Wed Sep 28 15:27:56 PDT 2011	VM Manager	Info	Switch 10.255.22
Wed Sep 28 15:27:58 PDT 2011	VM Manager	Info	Switch 10.255.22
Wed Sep 28 15:27:58 PDT 2011	VM Manager	Info	2 switch(es) polle

VM Manager | admin | 128.251.16.86

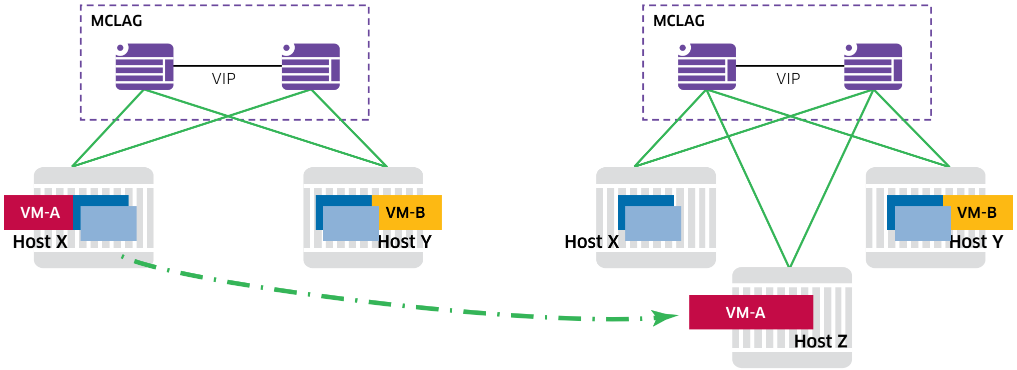
3.4 Application (VM) Mobility (vMotion)

VM mobility in the data center can occur for several reasons; some vMotion policies can be automated, including:

- Server performance degradation
- Server capacity limits breached for memory and disk space utilization
- Power failure/ disaster
- Maintenance window
- Upgrades

In the MCLAG domain, VIP is the common IP interface between the two MCLAG switches and serves as the gateway for connected VMs. With this configuration, ARP requests received by either MCLAG peer will be responded to by both peers with VIP MAC. The hosts will remain associated with the multi-chassis virtual IP interface (IP, MAC-VIP).

Figure 15. VM mobility in MCLAG architecture

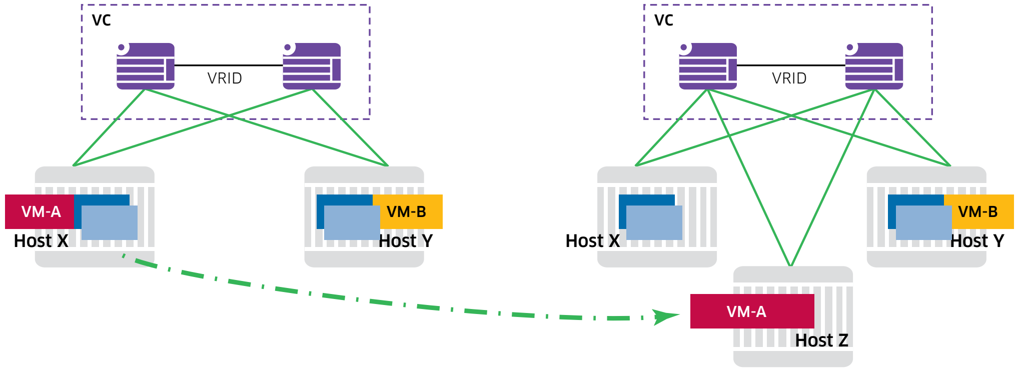


In MCLAG architecture, VMs can either move within a subnet or across subnets. In Figure 15 VM-A is communicating with VM-B. VM-A is manually moved to Host Z, which is newly provisioned for improved performance. When VM-A lands on Host Z, it sends a gratuitous ARP and is operational in less than a second with little to no impact on the user.

Note: when traffic arrives from the upstream network, the MCLAG peers send traffic directly to connected hosts without crossing the VFL, which limits the user traffic on this link whenever possible.

In the VC domain, Virtual Router Redundancy Protocol (VRRP) ID (or VID) is the common IP interface between the two VC switches and serves as the gateway for VMs connected. The switches in the VC domain are configured with active-active VRRP enabled. In this way, ARP request received by either VC peer will be responded by both peers with VRRP MAC. The hosts will remain associated to the VC IP interface (VID, VRRP-MAC).

Figure 16. VM mobility in VC architecture



In VC architecture, VMs can move either within the same subnet or across subnets. In the example shown in Figure 16 VM-A is communicating with VM-B. VM-A is manually moved to Host Z, which is newly provisioned for improved performance. When VM-A lands on Host Z, it sends a gratuitous ARP and is operational within in under a second with little or no impact to the user.

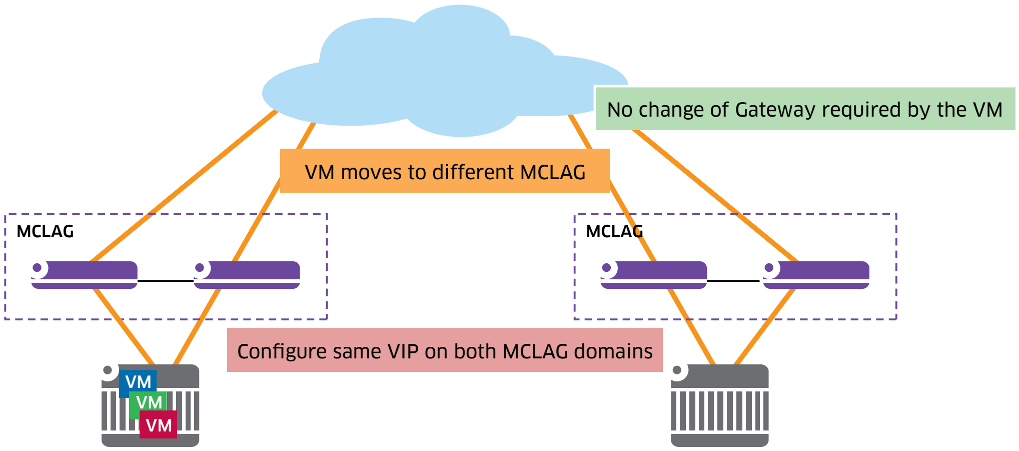
Note: when traffic arrives from the upstream network, the MCLAG peers send traffic directly to connected hosts without crossing the VFL, which limits user traffic on this link whenever possible.

*VM mobility in VC and SPB architectures is currently limited to hosts within the same subnet.

3.5 Resolution to Sub-optimal Routing in Flat Architectures

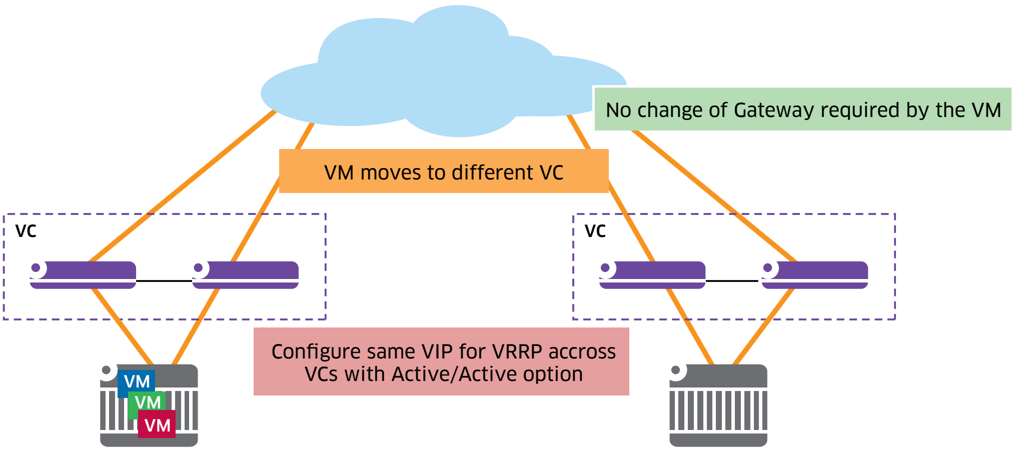
MCLAG domains running in remote sites can be connected either via routed interfaces or bridged interfaces, creating a large flat network. When VMs migrate to another host – and if the gateway does not exist on the local MCLAG domain – the VM continues to use old MCLAG switches for routing. This results in the WAN infrastructure being used sub-optimally and produces undeterministic latency. This issue can be avoided by having VIP (VM gateway) configured to achieve optimal routing on the destination switches.

Figure 17. Addressing sub-optimal routing for mobility across MC-LAG domains


No change of Gateway required by the VM
VM moves to different MCLAG
MCLAG
MCLAG
Configure same VIP on both MCLAG domains
VM
VM
VM

Similarly to MCLAGs, VC domains running in remote sites can be connected either via routed interfaces or bridged interfaces, creating a large flat network. When VMs migrate to another host – and if the gateway does not exist on the local VC domain – the VM continues to use old VC switches for routing. This results in the WAN infrastructure being used sub-optimally and produces undeterministic latency. This issue can be addressed by having VID (VM gateway) configured to achieve optimal routing on the destination switches.

Figure 18. Addressing sub-optimal routing for mobility across VC domains


No change of Gateway required by the VM
VM moves to different VC
VC
VC
Configure same VIP for VRRP accross VCs with Active/Active option
VM
VM
VM

4 STORAGE CONVERGENCE

Table 5. A comparison of storage technologies

STORAGE TECHNOLOGIES	ISCI	NAS	FCOE	RC
IP support	Yes	Yes	No	No
Guaranteed delivery	Yes (TCP)	Yes (TCP)	Yes (DCB)	Yes (Native)
Transmission speed	1GE/10GE/40GE	1GE/10GE/40GE	10GE+	2G/4G/8G/16G FC
Target market	S/M/L	S/M	M/L	M/L/XL
Growth	Moderate	High	High	Flat
Cost	Low/Mid	Low	High	High

Technologies such as SAN and Infiniband that offer differentiated services require specialized network transport equipment and IT personnel to manage. Application mobility across the data center requires storage mobility for optimal use, resulting in architectures that are static and expensive. FC/Infiniband mobility across WAN is a challenge (DWDM only option).

Ethernet speeds and feeds have far surpassed FC and Infiniband offerings, and latency in Ethernet networks is approaching that of Infiniband. Ethernet benefits now make a strong case for the adoption of one converged fabric. To maximize utilization and monetize investments, service providers and IT teams are looking to converged Ethernet to both carry server, storage and application data and to improve delivery services. Generic Ethernet is a best-effort network that does not guarantee delivery of packets. Ethernet standards have, therefore, been ratified with protocols defined to deliver end-to-end lossless behavior.

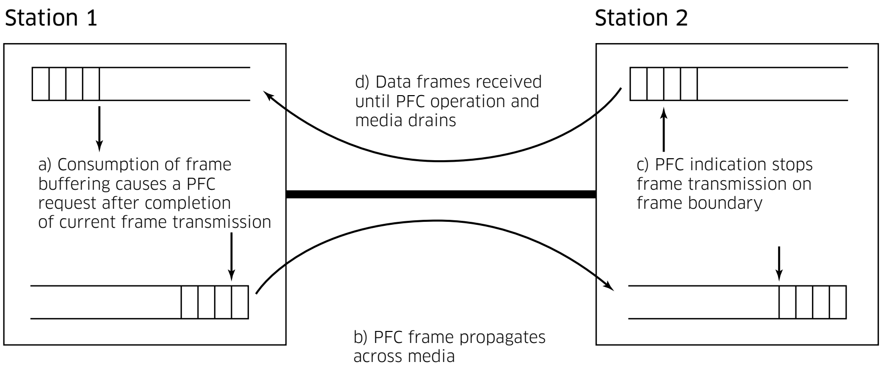
The Fiber Channel Backbone - 5 (FC-BB-5) standard specifies that Fiber Channel over Ethernet is intended to operate over an Ethernet network that does not discard frames in the presence of congestion. The IEEE 802.1 Data Center Bridging (DCB) Task Group has defined a set of standards that enhance existing 802.1 bridge definitions to provide a converged network that allows multiple applications to run over a single physical infrastructure. The DCB standards include Priority-based Flow Control (PFC), Enhanced Transmission Selection (ETS), and the Data Center Bridging Capabilities Exchange protocol (DCBX).

PFC (IEEE 802.1Qbb)

PFC provides more granular flow control, allowing the switch to pause certain traffic types based on 802.1p values in the VLAN tag. To assure data frames are not lost due to lack of receive buffer space, recipients ensure trigger of a PFC frames upon congestion, while there is sufficient receive buffer space (headroom) to absorb the data that may be received while the remote system reacts to the PFC operation. PFC is not necessarily restricted to Fiber Channel over Ethernet (FCoE) networks. It could be used for loss-sensitive traffic in any network where traffic is separated into different priorities. The use of PFC with FCoE traffic provides a functional equivalent to Fiber Channel's buffer-to-buffer credit mechanism:

- OmniSwitch 10k provides 8 lossless priorities per port
- OmniSwitch 6900 provides 128 lossless priorities per system

Figure 19. PFC operational states


Station 1
Station 2
d) Data frames received until PFC operation and media drains
a) Consumption of frame buffering causes a PFC request after completion of current frame transmission
c) PFC indication stops frame transmission on frame boundary
b) PFC frame propagates across media

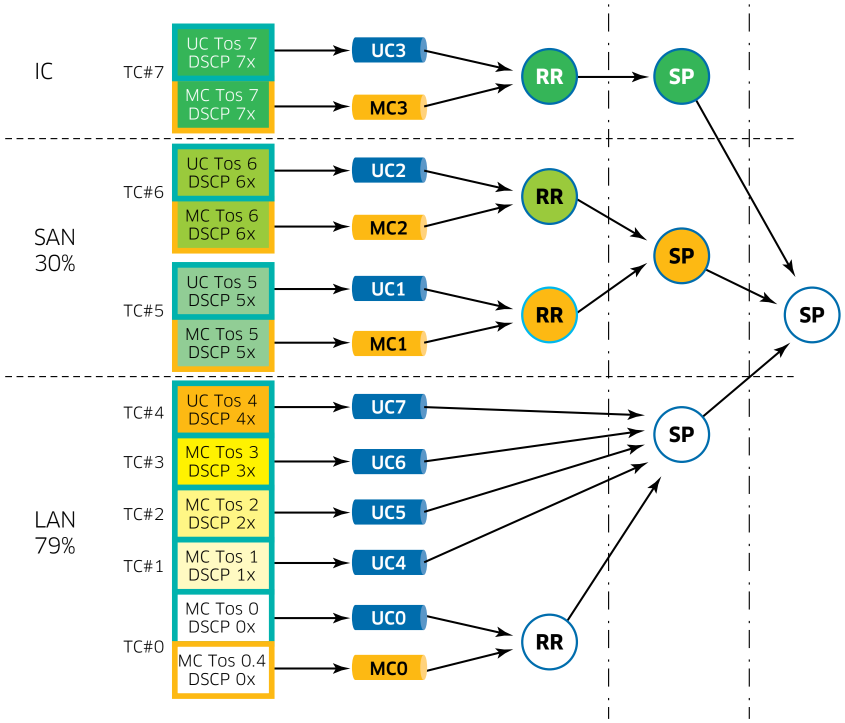
ETS (IEEE 802.1Qaz)

"The Data Center bridging ETS standard (802.1Qaz) specifies mechanisms to support allocation of bandwidth amongst traffic classes. When the offered load in a traffic class does not use its allocated bandwidth, enhanced transmission selection will allow other traffic classes to use the available bandwidth. The bandwidth-allocation priorities coexist with strict priorities. Networks prioritize traffic to provide different service characteristics to traffic classes."

AOS simplifies user configuration of the network by providing up to 128 user profiles to configure the port lossless properties. Profiles DCB_1 to DCB_10 are based on Standard 802.1Q-REV/D1-5 Appendix I. The default profile is DC_8, which allows best effort strict priority on all traffic classes. DCB_11+ are custom profiles that can be derived from the default profiles.

AOS supports hierarchical scheduling, which supports up to eight traffic classes. Each scheduler will be configured as SP, WERR (ETS) or BE. It provides a fair bandwidth allocation scheduler that takes into account the variable packet sizes when allocating bandwidth. Every front panel port has eight UC queues, of which the lower four UC queues (Q0 ~ Q3) have fixed mapping to S3 scheduling nodes or flexible mapping to S2 nodes. The upper four UC queues (Q4 ~ Q7) have flexible mapping to any of the three S2 nodes. Every front panel port has four MC queues which can be arranged either as one MC group with its own scheduler connected to S1 or via fixed mapping to S3 scheduling nodes. The S3 level schedulers balance unicast traffic with multicast traffic. But because there are only four MC queues, some priority grouping is needed for multicast traffic.

Figure 20. Converged Link Traffic Configuration



DCBX (IEEE 802.1Qaz)

DCBX is a capabilities-exchange protocol used by DCB-capable switches. Neighboring devices use DCBX to exchange and negotiate configuration information and detect misconfigurations. Devices can use DCBX for things such as exchanging information regarding FCoE, to determine whether PFC is enabled on the port, and to learn ETS priority group information, including 802.1p priority values and bandwidth allocation percentages.

The supported configuration of traffic classes (TC) in the system will be managed from higher TC to lower TC, in which highest TC will support the high SP traffic in case it is needed, such as Network control and Internet control. Only the lowest or the contiguous lowest TC will have the best effort/background traffic. Supported configurations for checking this traffic class consistency are:

- High SP- ETS - Low SP
- High SP-ETS
- ETS-Low SP

When two Omni switches with different profiles negotiate using DCBX:

ETS: As the default profiles have willing bit enabled, each node will change their operation settings to match other traffic classes. They are accepted because all default profiles pass the traffic class consistency check.

PFC: As the default profiles have willing bit enabled, both ends will resolve into the configuration of PFC of one end based on MAC Address. The PFC consistency check (all priorities on the traffic class should support same PFC type either lossless or lossy) will pass when connecting two switches with default profiles, as the profiles have been created from an implosion/explosion of priorities with same PFC characteristics.

The lossless benefits of AOS can be extended broadly beyond applications such as FCoE and iSCSI.

5 DC INTERCONNECT / WAN

Enterprises seek to remain operational in the event of any disaster. This fundamental business objective demands organizations don’t deploy or rely on a single silo data center for all operations.

Data center sites that provide backup, load sharing capabilities can be located either in the same city, in adjacent cities or in different countries to form private clouds. Small enterprises don’t have the budget to create backup data centers, they must rely on the public cloud for computational elasticity and benefit from a pay-as-you-grow model, leading to adoption of hybrid clouds.

Data center interconnect networks traditionally cater to high-availability, security and data replication, however, virtualization demands the network be a single fabric end-to-end to allow seamless application and storage migration.

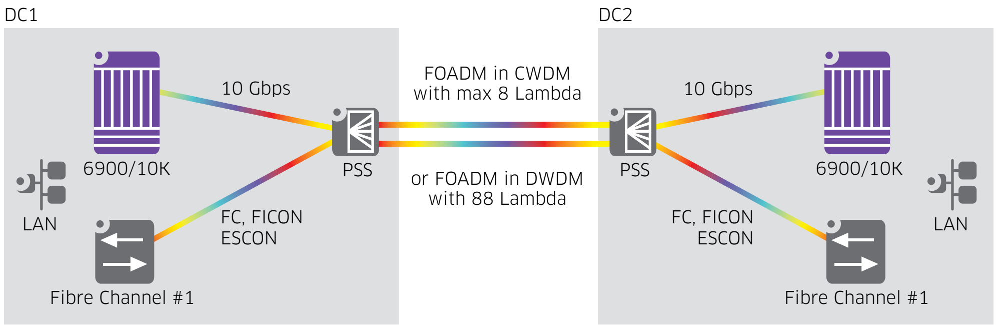
Alcatel-Lucent Enterprise provides data center interconnect solutions comprising IP/MPLS backbone, an optical backbone or a bridge to launch hybrid clouds. Each solution's services are differentiated based on (but not limited to latency) security, capacity and virtualization.

5.1 Optical Interconnect

Alcatel-Lucent 1830 Photonic Service Switch optical interconnect provides:

- Transport fabric between data centers at the speed of light
- Multiple client interfaces (1 Gb, 10 Gb FC, FICON), all of which can be mixed onto one single 10Gbps line
- Each service can be carried on a separate lambda (1 to 88) over a single 10Gbps or 40Gbps line
- Ultra-low latency (~5us) optimized for Data Center interconnect (DCi) applications
- Security can be embedded at the physical layer using symmetric key encryption with minimal impact to latency

Figure 21. DCi Optics Interconnect Solution



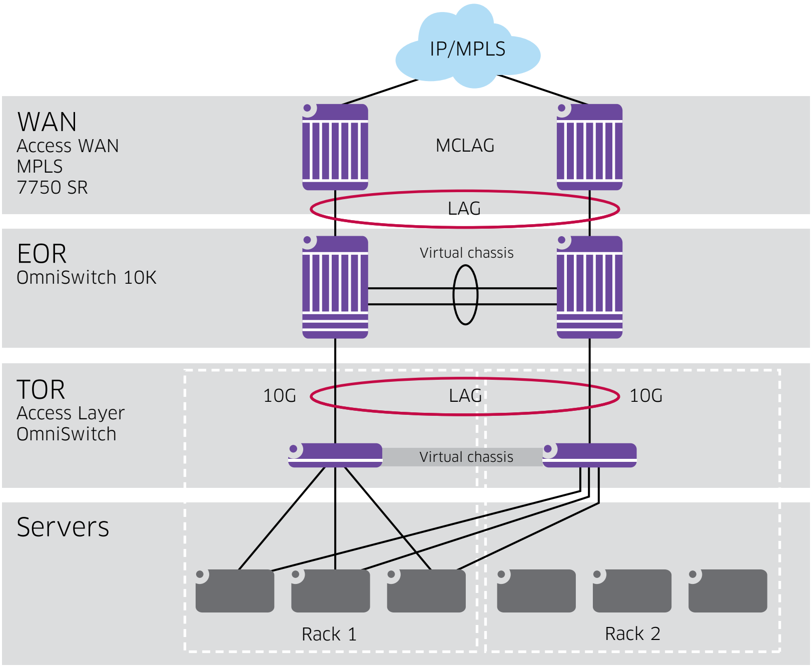
Bidirectional transmission requires two fibers (transmit and receive), however some applications or requirements call for bidirectional transmission over a single fiber. Different wavelengths are used in each direction to achieve this result. The 1830 PSS provides a secure LAYER 2 interconnect fabric.

5.2 IP/MPLS Interconnect

ALU 7750 SR IP/MPLS interconnect provides:

- Service-aware transport layer with IP overlay
- Layer 2 VPN and Layer 3 VPN capabilities, enabling Layer 2/Layer 3 virtualization
- High = speed interconnects: 10Gbps/ 100Gbps
- MCLAG support, for node failure redundancy and distributed load sharing
- Network failure recovery in < 50ms with fast reroute
- Zero-downtime upgrades, as ISSUisenabled

Figure 22. DCi IP/MPLS Interconnect Solution

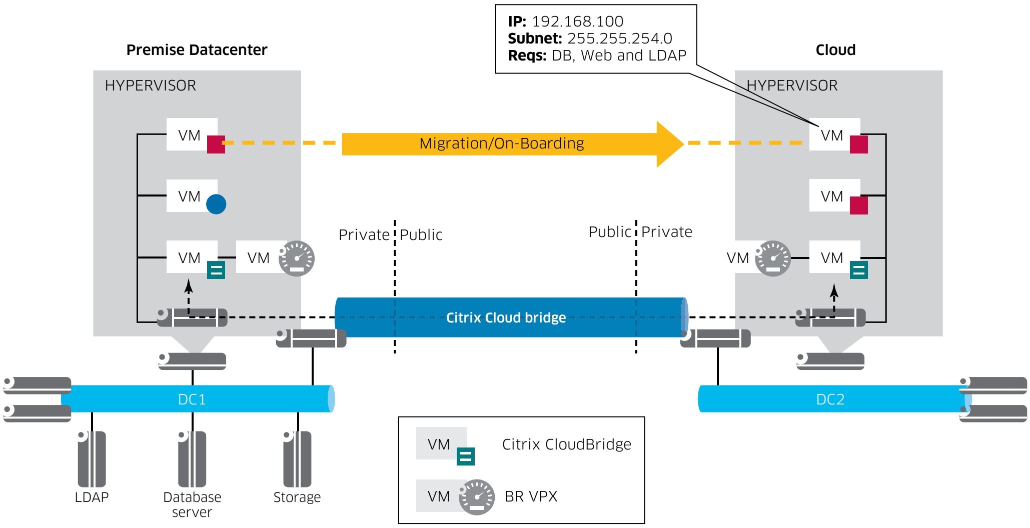


5.3 Hybrid Interconnect

Citrix Cloudbridge interconnect provides:

- Connection between enterprise datacenters and external clouds, making the cloud a secure extension of the enterprise network
- Seamless application visibility and migration offering same level of service by tight integration of solution with AOS/OmniVista
- IPSEC security to ensure data remains secure as it traverses the network between the enterprise and the cloud
- Layer 2 or Layer 3 tunnel to the cloud
- WAN-optimized access for applications using TCP optimizations, data compression and de-duplication techniques

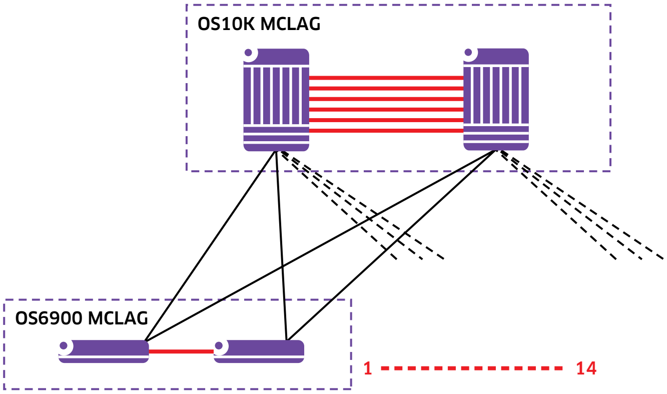
Figure 23. DCi Citrix CloudBridge Interconnect Solution



6 SCALE-UP/SCALE-OUT DESIGNS

6.1 MCLAG: Supports 728 10G dual home server/storage connections

Figure 24. MCLAG Scale-up/scale-out design



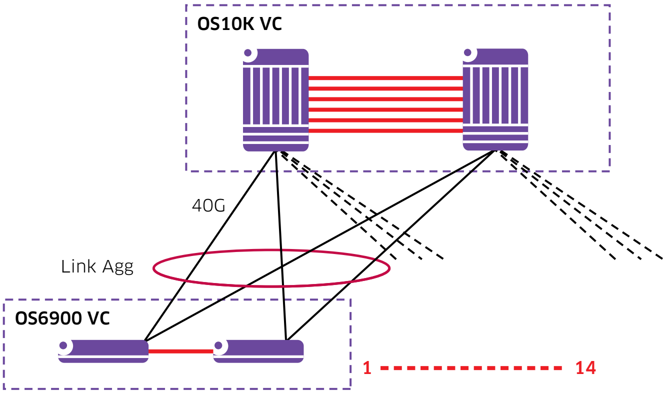
If increasing the oversubscription ratio is appropriate, then the number of server connections can be improved if the VFL/uplink core connections are changed to 10G.

Oversubscription: Assuming the red links are 40GE and the blue links are 10GE, the worst-case oversubscription ratio in this network is 5:1. However, if the blue links are also 40G, the worst-case oversubscription ratio is reduced to 1.3:1.

Latency: In most data center designs, servers that communicate most are localized. Assuming 90 percent of east-west traffic is contained within the OmniSwitch 6900 MCLAG groups and 10 percent of traffic goes northbound to OS10K MCLAG, the latency equals 3microseconds.

6.2 VC: Supports 728 10G dual home server/storage connections

Figure 25. VC scale-up/scale-out design

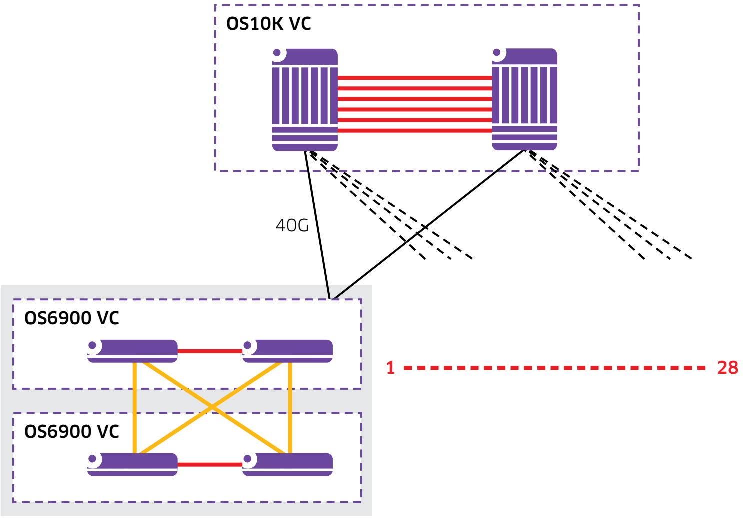


If increasing the oversubscription ratio is okay, then the number of server connections can be improved if the VFL/uplink core connections are changed to 10G. Oversubscription: Assuming the red links are 40GE and the blue links are 10GE, the worst-case oversubscription ratio in this network is 5:1. However, if the blue links are also 40G, the worst case oversubscription is reduced to 1.3:1.

Latency: In most data center designs, servers that communicate most are localized. Assuming 90 percent of east-west traffic is contained within OmniSwitch 6900 VC groups and 10 percent of traffic goes northbound to OmniSwitch 10K VC, then the latency equals 3 microseconds.

6.3 VC: Supports 2576 10G dual home server / storage connections

Figure 26. VC scale out architecture



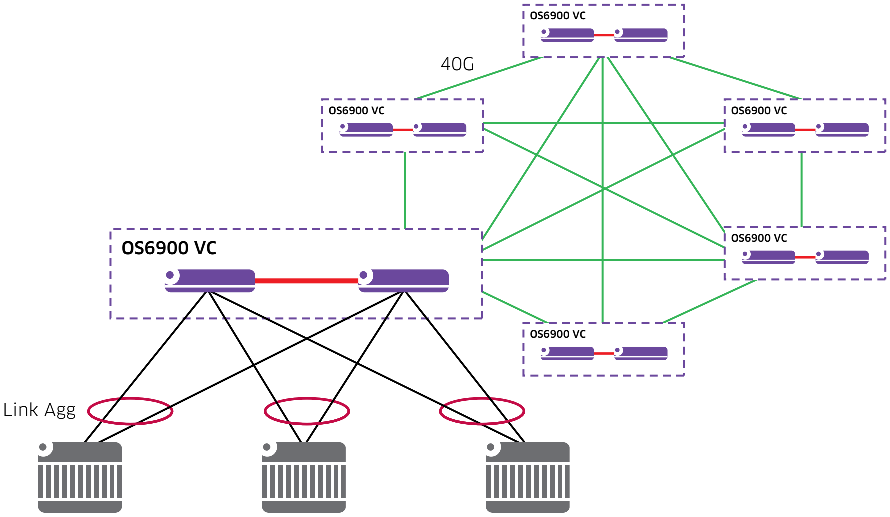
In this VC scale out architecture, the uplinks from the OmniSwitch 6900 to OmniSwitch 10K must be from one single VC system to the core OmniSwitch 10K VC. The two connections cannot be distributed across the two OmniSwitch 6900 VC systems, since it would form a loop.

Oversubscription: Assuming the red and blue links are 40GE and the yellow links are 10GE, the worst-case oversubscription ratio in this network is 5:1. However, if the yellow links are also 40G, the worst-case oversubscription ratio is reduced to 2.5:1.

Latency: In most data center designs, servers that communicate most are localized. Assuming 50 percent of east-west traffic is contained within OmniSwitch 6900 VC groups, another 40 percent of traffic goes across the OmniSwitch 6900 VC mesh and the remaining 10 percent goes northbound to OmniSwitch 10K VC, then latency equals 3.3 microseconds.

6.4 VC and SPB: Supports 200 10G dual home server/storage connections

Figure 27. VC and SPB scale-up/scale-out architecture



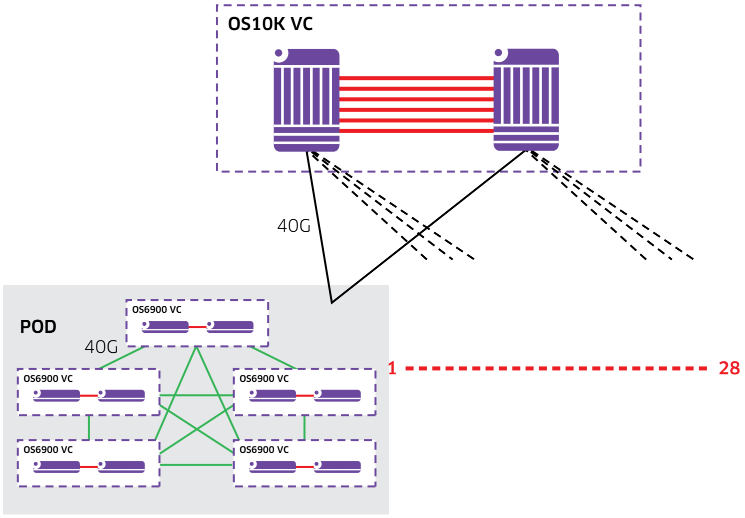
In this architecture, the OmniSwitch 6900 VC groups are connected in a full mesh using SPB, forming a pod.

Oversubscription: Assuming the red and green links are 40GE, the worst-case oversubscription ratio in this network is 1.3:1.

Latency: In most data center designs, servers that communicate most are localized. Assuming 90 percent of east-west traffic is contained within each OmniSwitch 6900 VC group and the remaining 10 percent goes to external OmniSwitch 6900 VC groups, then latency equals 1.3microseconds.

6.5 VC and SPB: Supports 5600 10G dual home server/storage connections

Figure 28. VC and SPB scale-up/scale-out architecture



In this architecture, the OmniSwitch 6900 VC groups are connected in a full mesh using SPB. Such a set is called a pod. Each pod connects dual homed to OmniSwitch 10K VC.

Oversubscription: Assuming the red, blue and green links are 40GE, the worst-case oversubscription ratio in this network is 6.5:1.

Latency: In most data center designs, servers that communicate most are localized. Assuming 50 percent of east-west traffic is contained within each OmniSwitch 6900 VC group, another 40 percent is contained within the pod and the remaining 10 percent goes to northbound to OmniSwitch 10K VC, then latency equals 4.1 microseconds.

7 CONCLUSION

The emergence of technologies such as cloud computing and virtualization have forced organizations to reexamine data center design. To support the demanding availability requirements of today's applications, data centers need to enhance redundancy requirements and build a resilient infrastructure that will meet the needs of today and tomorrow.

Alcatel-Lucent Enterprise network architecture provides a true converged fabric that addresses the needs of existing and future data centers, offering high availability, low latency and complete application visibility and mobility. It offers an innovative switching fabric that enables a range of innovative data center deployment models — from dedicated virtual data centers, to multi-site private cloud, to hybrid cloud environments. Each deployment model can be different in terms of scalability, oversubscription and latency.

Table 6. A comparison of various DC architectures

ARCHITECTURE/ SCALABILITY	MCLAG	VIRTUAL CHASSIS	SPB	VIRTUAL CHASSIS AND SPB
Base Architecture	<256 servers	<512 servers	>1000 servers	>1000 servers
Scale-Up with Spine-Leaf / Mesh Architectures	<1000 servers	<3000 servers	>1000 servers	>1000 servers

* These numbers are based on the assumption that all server/storage connections to the DC are dual home active-active 10G.

DC ECOSYSTEM REFERENCE PAPERS

Data Center Switching Solution

- http://enterprise.alcatel-lucent.com/docs/?id=19297

Virtual Desktop (VDI) / Optimized Network for Citrix Xen Desktop

- http://enterprise.alcatel-lucent.com/docs/?id=22020

WAN Optimization/ Intelligent Load balancing with Citrix Netscaler

- http://enterprise.alcatel-lucent.com/docs/?id=22267

Converged Network Adapters (CNA)

- http://enterprise.alcatel-lucent.com/docs/?id=21980

High-Performance Architecture for Large Scale Data Analytics

- http://enterprise.alcatel-lucent.com/docs/?id=22187

Alcatel·Lucent
Enterprise