Cisco VSPEX Solution with Microsoft Fast Track 3.0 Small Implementation

March 2013

Contents

Executive Summary

The Microsoft Private Cloud Fast Track Program enables customers to get up and running with a preconfigured private cloud. A joint effort between Microsoft and its hardware partners, the Fast Track Program helps organizations quickly develop and implement private clouds while reducing both cost and risk. The program is built on a reference architecture that combines Microsoft software, consolidated guidance, and validated configurations with partner technology—such as computing power, network and storage architectures, and value-added software components.

The Customer Scenarios

The customer scenarios for Microsoft Private Cloud Fast Track's (Fast Track) Small Implementations highlight the advanced performance capabilities and advanced features of Windows Server 2012. The suitability of these customer scenarios for small implementations depends on the specific hardware configuration. Microsoft has been working with the server industry to enable partners to create a new generation of simpler, high-availability solutions that deliver Small Implementations as Cluster-in-a-Box (CiB) or as consolidation appliance solutions at a lower pricing point.

In the server appliance market, the customer scenarios that are a particularly strong match for the Small Implementations architecture are defined in the following sections. These scenarios help system builders identify requirements that will differentiate their solution for target customers, such as by adding features that improve the Windows Server out-of-box-experience (OOBE), or by adding management or other applications in order to simplify customer deployment and management.

Business and Branch Scenarios

The business and branch scenarios use an on-premises, in-a-box consolidation appliance solution, where all of the server and storage needs for a small business or a branch office are provided in a single consolidated, "prepackaged" system, as described in the following sections.

Business In-a-Box, Consolidation Appliance

The business-in-a-box scenario assumes that the consolidation appliance is being used as an on-premises Hyper-V appliance locally at a small or medium-sized business. Examples of this scenario include a doctor's office, individual retail store, or a lawyer's office.

In this scenario, the solution is enabled in an office environment or in an equipment room that has access to an Internet service provider (ISP) network connection. The Hyper-V appliance can enable a workload that is capable of supporting a variety of applications, such as point of sale (PoS), inventory, documents, and records, or any other line-of-business (LOB) application. In this scenario, the IT administrator is expected to be a part-time generalist, and administration might not be that person's primary job function. They might get assistance from a value added reseller (VAR).

Branch In-a-Box, Consolidation Appliance

The branch-in-a-box scenario assumes that the consolidation appliance is being used as an on-premises Hyper-V appliance at a branch office that has a central office infrastructure. Thus, the solution emphasizes the importance of remote management. Examples of this scenario are retail chain stores and bank branches.

In this scenario, the solution is enabled in an office environment or in an equipment room that is domain-connected to the main office. The Hyper-V appliance can enable replicated workloads per branch, with line-of-

business (LOB) applications, and with a file server with cached data. In this scenario, the IT administrator works centrally from their headquarters, primarily providing remote support, with local assistance when required.

File Server Scenario

In this scenario, the solution is designed as a storage building block for the data center —for example, as a dedicated storage appliance. Examples of this scenario are cloud solution builders and enterprise data centers. For example, suppose the solution supports Windows Server 2012 Server Message Block (SMB) 3.0 file shares for Hyper-V or SQL server applications. In this case, the solution would enable the transfer of data from the drives to the network at bus and wired network speeds with CPU utilization that is comparable to Fibre Channel.

In this scenario, the file server is enabled in an office environment in an enterprise equipment room that provides access to a switched network. As a high-performance file server, the solution can support variable workloads, hosted LOB applications, and data. In this scenario, the IT administrators are specialists who are available both onsite and remotely, but they are used by many parts of the organization and may be overcommitted.

Architecture Principles

The architecture principles for Fast Track Small Implementations describe the results that are typically at work in a successful consolidation appliance or a small private cloud solution. These principles are usually achieved in conjunction with one another.

They include:

- Resource pooling
- Elasticity and perception of infinite capacity
- Perception of continuous availability
- Drive predictability

Each of these principles is described in the following sections.

Resource Pooling

Resource optimization is a principle that drives efficiency and cost reduction, and it is achieved primarily through resource pooling. Resource pooling abstracts the platform from the physical infrastructure. It enables the optimization of resources through shared use. When multiple entities share resources, the result is higher resource utilization, which leads to more efficient and effective use of the infrastructure. This optimization can ultimately help drive down costs and improve agility.

Elasticity and the Perception of Infinite Capacity

The solution should appear to have infinite capacity to the consumer. Just as a consumer uses as much electricity as they need from the electric utility provider, so end users can consume cloud-based services on demand.

This utility mindset requires that capacity planning be proactive so that requests can be satisfied on demand. Applying this principle reactively and in isolation often leads to inefficient use of resources and unnecessary costs. Combined with other principles, such as using incentives to encourage desired consumer behavior, this principle allows for a balance between the cost of unused capacity and the need for agility.

Perception of Continuous Availability

The solution should always appear available to the consumer. The consumer never experiences an interruption of service, even if failures occur within the cloud environment. To achieve this perception, a provider must have a

mature service management approach, inherent application resiliency, and infrastructure redundancies in a highly automated environment.

Predictability

From the consumer's viewpoint, the solution should be consistent and have the same quality and functionality each time it is used. For a provider, predictability is realized through the homogenization of the underlying physical servers, network devices, and storage systems. This consistency enables hosted workloads to be supported in a manner that drives service quality.

Conceptual Architecture

Fast Track for Small Implementations features an architecture in which the infrastructure is presented as a number of distinct layers that form a fabric. The conceptual architecture includes:

- **Scale units.** In a modular architecture, a scale unit refers to a redesigned unit of capacity at each layer of the architecture.
- **Fabric infrastructure.** You manage the fabric with tools such as Windows Server 2012 in-box Server Manager, Hyper-V Manager, and Failover Cluster Manager. These tools can be paired with partner technologies to augment the management infrastructure as needed.

Scale Units

Scale units establish the criteria for deploying additional hardware by defining when a given layer will need to be scaled. For instance, a scale unit can be an individual server that is defined based on its CPU and RAM capabilities. When the server reaches its maximum scalable limit, an additional server is required to continue scaling. Windows Server 2012 can significantly increase both the density and scale of virtualization and private cloud infrastructures.
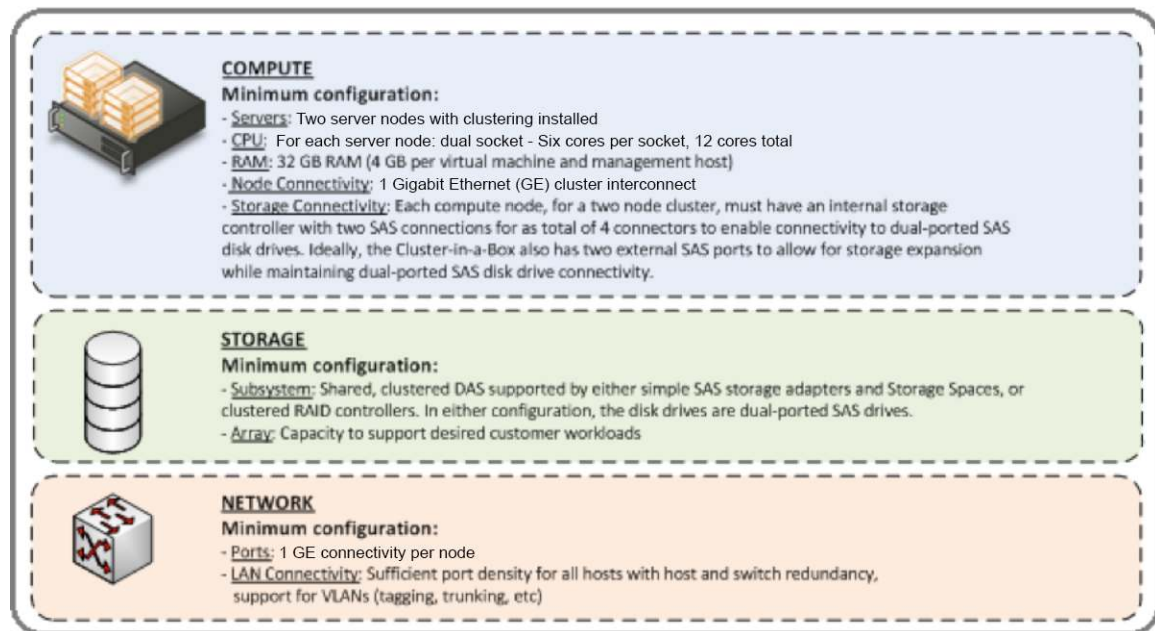
For small implementations, the scale units are for fewer than 75 server virtual machines, and two to four host servers.

The scale limits of all components, both hardware and software, are critical in determining the optimum scale units for the overall architecture. Scale units enable all of the requirements, such as space, power, heating, ventilation, and air conditioning (HVAC), and connectivity, needed for the implementation to be documented. Each scale unit also factors in its associated amount of physical installation and configuration labor.

Fabric Infrastructure

For Fast Track Small Implementations, the fabric infrastructure for the entry level, on-premises Cluster-in-a-Box high-availability design is shown in Figure 1.

Figure 1.   Cluster-in-a-Box Fabric Infrastructure

COMPUTE
Minimum configuration:
- <u>Servers</u>: Two server nodes with clustering installed
- <u>CPU</u>: For each server node: dual socket - Six cores per socket, 12 cores total
- <u>RAM</u>: 32 GB RAM (4 GB per virtual machine and management host)
- <u>Node Connectivity</u>: 1 Gigabit Ethernet (GE) cluster interconnect
- <u>Storage Connectivity</u>: Each compute node, for a two node cluster, must have an internal storage controller with two SAS connections for as total of 4 connectors to enable connectivity to dual-ported SAS disk drives. Ideally, the Cluster-in-a-Box also has two external SAS ports to allow for storage expansion while maintaining dual-ported SAS disk drive connectivity.

STORAGE
Minimum configuration:
- <u>Subsystem</u>: Shared, clustered DAS supported by either simple SAS storage adapters and Storage Spaces, or clustered RAID controllers. In either configuration, the disk drives are dual-ported SAS drives.
- <u>Array</u>: Capacity to support desired customer workloads

NETWORK
Minimum configuration:
- <u>Ports</u>: 1 GE connectivity per node
- <u>LAN Connectivity</u>: Sufficient port density for all hosts with host and switch redundancy, support for VLANs (tagging, trunking, etc)

For small implementations, the server, network, compute, and virtualization layers for the Cluster-in-a-Box fabric infrastructure are described in the following sections.

Servers

Server scale limits can include the number and speed of CPU cores, the maximum amount and speed of RAM, and the number and type of expansion slots. In addition, the number and type of onboard input/output (I/O) ports, as well as the number and type of supported I/O cards, are particularly important. Both Ethernet and Fibre Channel expansion cards often provide multiport options, since a single card can have four ports.

In blade server architectures, there are often limitations in the amount of I/O card and/or supported combinations. It is important to be aware of these limitations, along with the oversubscription ratio between blade I/O ports and any blade chassis switch modules.

Windows Server 2012 Hyper-V increases scalability and expands support for host processors and memory including:

- Support for up to 64 processors and 1 TB memory for Hyper-V virtual machines. A Hyper-V host supports up to 320 logical processors, 4 TB memory, and 1024 active running virtual machines per host, and can be scaled across a maximum of 64 cluster nodes, running up to 4000 virtual machines.

- Advanced server features including the ability to project a virtual Non-Uniform Memory Access (NUMA) topology into a virtual machine to provide optimal performance and workload scalability in large virtual machine configurations.

Windows Server 2012 Hyper-V also provides improvements to dynamic memory, including:

- Minimum memory. Minimum memory allows Hyper-V to reclaim unused memory from virtual machines to allow for higher virtual machine consolidation numbers.

- Hyper-V Smart Paging. Smart Paging bridges the memory gap between minimum and startup memory. This feature allows virtual machines to start reliably when the minimum memory setting has indirectly led to an insufficient amount of available physical memory during restart.

In addition, Windows Server 2012 Hyper-V allows for runtime configuration of memory settings, including increasing the maximum memory and decreasing the minimum memory of running virtual machines.

These updated features allow the virtualization infrastructure to support the configuration of large, high-performance virtual machines that run demanding workloads. For more information about Windows Server 2012 Hyper-V, go to Hyper-V Overview at: http://technet.microsoft.com/en-us/library/hh831531.

Networking

High-availability virtualization solutions frequently use a dedicated host management network to help eliminate competition with guest traffic requirements, and to provide a degree of separation for security purposes. A dedicated management network is typically implemented using one network interface card (NIC) per host and one port per networked device to the management network.

Windows Server 2012 features enable:

- Network adapter load balancing and failover (LBFO), and quality of service (QoS) at the Hyper-V switch level. These features permit the configuration of management networks as virtual interfaces, while helping to ensure the required availability and performance levels. They are useful in scenarios having a lower number of ports and high bandwidth per port such as 10 Gigabit Ethernet (GE).

- Multitenant networking. This feature uses technologies such as virtual local area networks (VLANs) or Internet Protocol security (IPsec) isolation techniques to provide dedicated networks that utilize a single network infrastructure or wire. In addition, isolated networks can be provided where different owners, such as particular departments or applications, have their own dedicated networks.

- IPsec-based isolation can be accomplished using Active Directory® Domain Service (AD DS) and group policy to control firewall settings across the hosts and guests, as well as using IPsec policies to control network communication. For a high-availability virtualization solution, network settings and policies can be defined centrally and applied universally by the management solution.

- VLAN-based network segmentation. For this feature, components including the host servers, host clusters, and network switches must be configured correctly to enable both rapid provisioning and network segmentation.

- With Hyper-V and host clusters, identical virtual networks must be defined on all nodes in order for a virtual machine to fail-over to any node and maintain its connection to the network. At large scale, this configuration task can be accomplished using Windows PowerShell® scripting.

- Support for PVLANs. This feature provides isolation between two virtual machines on the same VLAN when the systems are not required to communicate with each other.

In addition, Windows Server 2012 introduces several networking enhancements including support for single root I/O virtualization (SR-IOV), third-party extensions to the Hyper-V extensible switch, QoS minimum bandwidth, network virtualization, data center bridging (DCB), and remote direct memory access (RDMA) network connectivity to support low-latency connectivity to remote resources.

For more information about Windows Server 2012 networking capabilities, visit:
http://download.microsoft.com/download/7/E/6/7E63DE77-EBA9-4F2E-81D3-9FC328CD93C4/WS 2012 White Paper_ Networking.pdf.

Storage

The storage architecture, including the storage and supporting storage networking, is a critical design consideration for high-availability virtualization solutions. The overall cost of storage can be significant because storage tends to be costly compared to other infrastructure components. For the Small Implementations architecture, the storage can be enabled using shared, clustered direct-attached storage (DAS). It can also be provided using a traditional Storage Area Network (SAN).

The Small Implementations architecture enables the rapid provisioning and deprovisioning of virtual machines. This requires tight integration with the storage architecture and robust automation.

Windows Server 2012 Hyper-V features include:

- An update to the virtual hard disk (VHD) format called VHDX. VHDX provides higher capacity (up to 64 terabytes of storage) and additional protection from corruption from power failures. It prevents performance degradation on large-sector physical disks by optimizing structure alignment.
- Support for offloaded data transfer (ODX) for advanced storage arrays that can support VHDX. ODX uses a token-based mechanism for reading and writing data within or between intelligent storage arrays. VHDX files connected to the virtual machine as virtual SCSI devices or by using DAS, in conjunction with ODX-capable hardware, can take advantage of this new capability.
- Virtual Fibre Channel. This feature allows virtual machines to have unmediated access to SAN logical unit numbers (LUNs). It enables scenarios including running the Windows Failover Cluster Management feature inside the guest operating system of a virtual machine connected to shared Fibre Channel storage. Virtual Fibre Channel supports multipath I/O (MPIO), N_Port ID Virtualization (NPIV) for many to one mappings, and up to four virtual Fibre Channel adapters per virtual machine.

For more information about Windows Server 2012 storage capabilities, visit:
http://download.microsoft.com/download/3/8/F/38F07CCB-B752-45DE-9747-247BAE5E2974/WS 2012 Data Sheet_Storage.pdf.

Virtualization

The virtualization layer provides for the decoupling of hardware, operating systems, data, applications, and user state. It opens a wide range of options for better management and distribution of workloads across the physical infrastructure. Hypervisor-based virtualization technologies enable solution capabilities, such as enabling the virtualization layer to migrate running virtual machines from one server to another without downtime, and many other features.

Virtualization provides an abstraction layer that moves the majority of management and automation to software, instead of requiring manual operations to be performed on the physical hardware. Similar to the hardware, the automation, management, and orchestration layers must be able to manage virtualization layer capabilities in a way that maintains desired states and proactively addresses decaying hardware, or other issues that would otherwise cause faults or service disruptions.

Windows Server 2012 Hyper-V introduces improvements in both virtualization features and scale. Together, these features provide significant enhancements to the capabilities of the Small Implementations architecture.

For more information about Windows Server 2012 server virtualization capabilities, visit:
http://download.microsoft.com/download/5/D/B/5DB1C7BF-6286-4431-A244-438D4605DB1D/WS 2012 White Paper_Hyper-V.pdf.

Fast Track Program Overview

Each Microsoft Private Cloud Fast Track Program Version 3 outlines the high-level architectural vision that is intended to enable partners to rapidly develop end-to-end, integrated, and tested virtualization or private cloud solutions for small and medium-sized businesses, and for the enterprise and data center. Each solution that the partner develops must meet or exceed Microsoft validation standards.

The Fast Track Program has three main branches, as shown in Figure 2.

Figure 2.   Private Cloud Fast Track Branches

For small implementations, the server, network, compute, and virtualization layers for the Cluster-in-a-Box fabric infrastructure are described in the following sections.

Fast Track Reference Architecture

Each Fast Track branch employs a reference architecture that define the requirements needed to design, build, and deliver virtualization and private cloud solutions for small, medium-sized, and large implementations. Examples of these Fast Track reference architectures are shown in Figure 3.

Figure 3.   Examples of Fast Track Reference Architectures
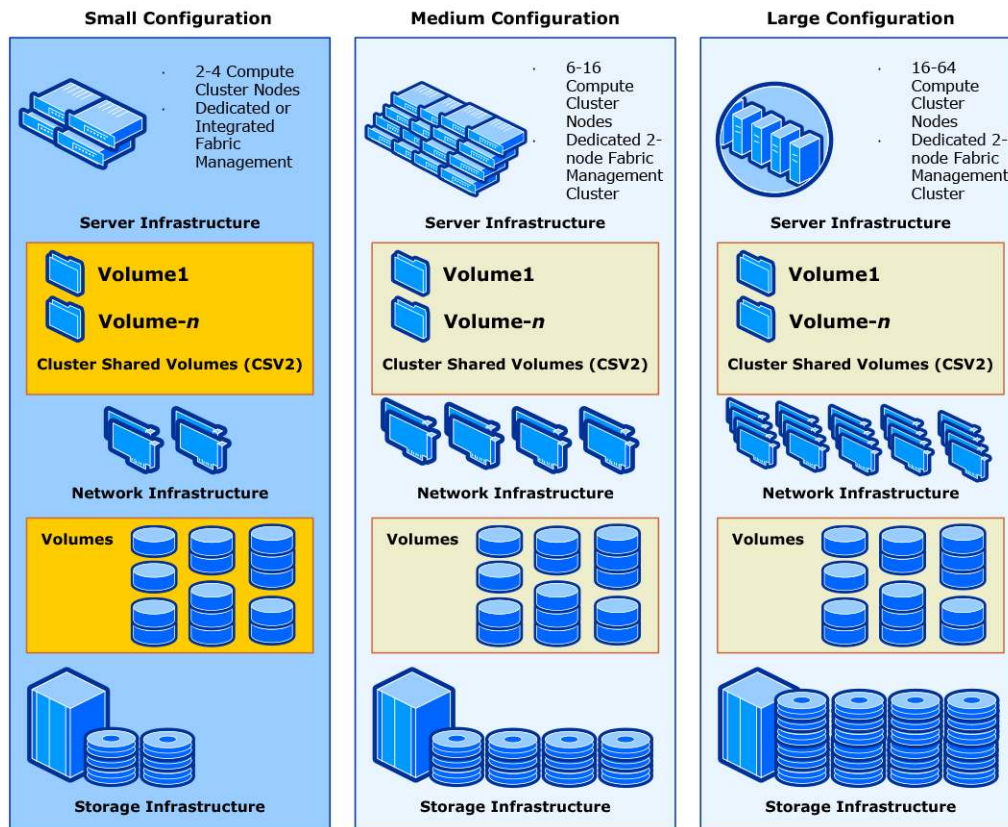
Each Fast Track reference architecture combines concise guidance with validated configurations for the compute, network, storage, and virtualization layers. Each architecture presents multiple design patterns for enabling the architecture, and each design pattern describes the minimum requirements for validating each Fast Track solution.

This document describes the Fast Track reference architecture for small implementations. It presents the Small Implementations architecture first at a high level, and then in-depth, including detailed technical guidance and the validation requirements for building a solution.

Fast Track Validation Requirements

The Microsoft Private Cloud Fast Track Program Version 3 describes the minimum requirements Microsoft will use to validate Partner solutions that were built using the design patterns described in this document. These validation requirements are organized into categories according to the criteria specified below.

Validation Criteria

The Microsoft Private Cloud Fast Track Program enables partners to build solutions based on a series of minimum requirements for computing, network, storage, and virtualization that are validated by Microsoft. All hardware and infrastructure validation requirements are vendor-agnostic.

Each validation requirement is categorized as one of the following:

Mandatory—Mandatory best practice, agnostic solution. These requirements are necessary in order to pass the Microsoft validation.

Recommended—Recommended best practice. These requirements describe industry-standard best practices that are strongly recommended. However, implementing these requirements is at the discretion of each partner. They are not required in order to pass the Microsoft validation.

Optional—Optional best practice. These requirements are voluntary considerations that can be implemented in the solution at the discretion of each partner.

The Cisco® and EMC solution implements all mandatory requirements and uses recommended and optional requirements when appropriate.

Windows Certification
In Fast Track for Small Implementations, it is mandatory for each architecture solution has to pass these validation requirements:

- Windows Hardware certification
- Failover Clustering validation
- If a third-party clustered RAID controller is used, the solution must also pass the clustered RAID Controller validation.

Each of these validations is described in the following sections.

Windows Hardware Certification

The architecture solution must receive validation through the Microsoft Certified for Windows Server 2012 Program in order to be presented in the Windows Server Catalog. The catalog contains all servers, storage, and other hardware devices that are certified for use with Windows Server 2012 and Hyper-V.

The Certified for Windows Server 2012 logo demonstrates that a server system meets Microsoft's highest technical bar for security, reliability, and manageability, and any required hardware components that support all the roles, features, and interfaces supported by Windows Server 2012.

The logo program and support policy for failover clustering solutions requires that all individual components that comprise a cluster configuration earn the appropriate "Certified for" or "Supported on" Windows Server 2012 designations before being listed in their device specific category in the Windows Server Catalog.

For more information, go to the Windows Server Catalog at: http://www.windowsservercatalog.com. Under Hardware Testing Status, click Certified for Windows Server 2012. The two primary entry points are the WHCK process and the SysDev Dashboard portal for starting the logo certification process.

The Cisco and EMC solution is based on the Cisco UCS C220 M3 Rack-Mount Server with the Cisco UCS 1225 Virtual Interface Card (converged adapter) connected to the EMC VNXe3300 iSCSI Target. All components are logo-certified for Windows Server 2012.

Failover Cluster Validation

For Windows Server 2012, failover clustering can be validated using the in-box Cluster Validation Tool to confirm network and shared storage connectivity between the nodes of the cluster. The tool runs a set of focused tests on the set of servers to be used as nodes in a cluster. This failover cluster validation process tests the underlying

hardware and software directly and individually obtains an accurate assessment of whether the failover clustering has the ability to support a given configuration.

Cluster validation is used to identify hardware or configuration problems before the cluster goes into production. This helps to ensure that a solution is truly dependable. Note that cluster validation can also be performed on configured failover clusters as a diagnostic tool.

The Cisco and EMC solution has run and passed the cluster validation wizard. This wizard does not put any limitations on the amount of memory or number of disks in the configuration, so both the amount of memory in the servers and disks in the storage array can be increased or decreased and still pass validation, allowing a customer to size this solution according to business needs.

In addition, the failover cluster must be tested and pass the Failover Cluster Validation in order to receive end customer support from Microsoft Customer Support Services (CSS).

Customer Experience

The Fast Track Version 3 Program helps to ensure that partner solutions built using Microsoft products and technologies provide our customers with a satisfying out-of-box user experience. The OOBE is intended to enable customers to trust our products and services, and it results in increased loyalty to Microsoft and its partners.

SKU-Based Design

The major focus of SKU-based design is to enable small and mid-market customers to rapidly implement virtualization and consolidation solutions. These architectures employ transformative technologies that deliver a robust, elastic, and highly available IT "ready to order" solution, rather than following a traditional private cloud design. For example, these designs typically target market segments that are characterized by less sophisticated onsite IT departments, rather than large-scale enterprise data centers.

The key differentiators for the SKU-based design are the ability to:

- Provide a quick and easy method for ordering a single SKU or part number for a given solution
- Deliver a solution that arrives onsite preloaded with operating system that is ready to run

The WS2012 VSPEX solution is available from multiple distributor sources. In North America, Avnet has created a single SKU for ordering the solution.

Ordering sku is: VSPEXM100FastTrack

The e-mail address for inquiries would be vspex@avnet.com. The telephone number is 1-800-409-1483, option #4.

Windows Server 2012 Licensing

The Fast Track Small Implementations architecture uses either Windows Server 2012 Standard Edition or Windows Server 2012 Datacenter Edition for enablement.

For more information about Windows Server 2012, visit: http://go.microsoft.com/?linkid=9799219.

Windows Server 2012 editions packaging and licensing has been updated to simplify purchasing and reduce management requirements, as shown in the Table 1. Windows Server 2012 editions are differentiated only by virtualization rights: two virtual instances for Standard edition and unlimited virtual instances for Datacenter

edition. Running instances can exist either in a physical operating system environment (POSE) or a virtual operating system environment (VOSE).

Table 1.   Licensing of Windows Server 2012 Editions

Edition	Running Instances In Pose	Running Instances In Vose	Edition
Datacenter	1	Unlimited	Datacenter
Standard	1 (When a customer is running all allowed virtual instances, the physical instance may only be used to manage and service the virtual instances.)	2	Standard

For more information, see the Windows Server 2012 Editions licensing overview at:
http://www.microsoft.com/en-us/server-cloud/windows-server/2012-editions.aspx.

For information about licensing in virtual environments, see the Microsoft Volume Licensing Brief at:
http://www.microsoft.com/en-us/download/details.aspx?displaylang=en&id=15113.

The Cisco and EMC solution is uses the Windows Server 2012 Datacenter Edition to provide the right to run an unlimited number of instances of Windows Server on any node in the cluster.

Reference Architecture In-Depth

As mentioned earlier, there are two reference architectures for Fast Track Small Implementations. The first is based on a Cluster-in-a-Box design that utilizes low-cost storage options connected to clustered RAID controllers. The second architecture is built on a storage solution that off-loads storage processing to a SAN. The Cisco and EMC solution is based on this secondary design.

Secondary Design Pattern: Clustered SAN

The Clustered SAN design pattern uses the highly available Windows Server 2012 Hyper-V clustered architecture with traditional SAN storage. The Clustered SAN design pattern enables the storage network and network paths to be combined over a single medium, which requires less infrastructure by offering a converged network design. The design pattern employs an Ethernet infrastructure that serves as the transport for the management and failover networks, and provides logical separation between these networks.

The Clustered SAN design pattern is shown in Figure 4.

Figure 4.   Clustered SAN Design Pattern
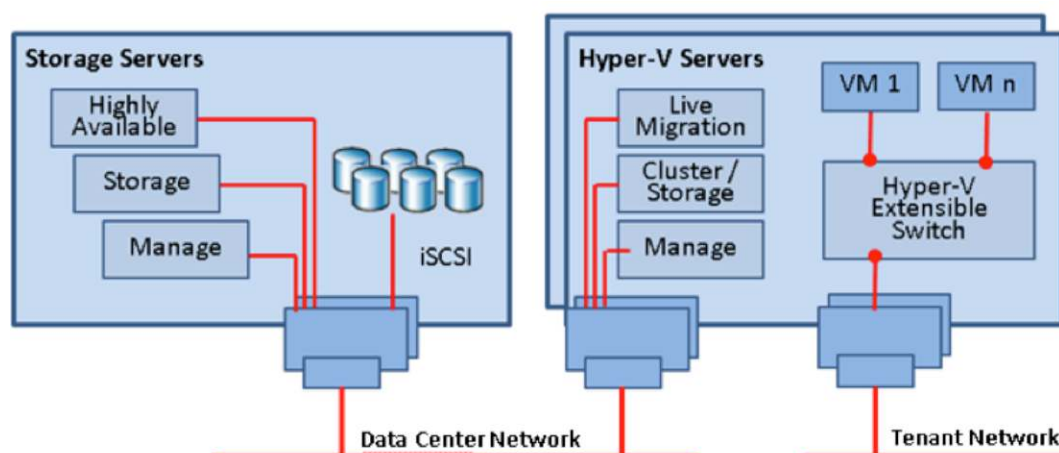
In addition to the converged design mentioned above, the Fast Track architecture allows for the flexibility to create "nonconverged" implementations in which internal traffic is segmented through multiple network connections. For example, a non-coverged implementation would have two network connections to the data center network, one for storage and another for data center management functions. This is considered an extension of the "mandatory" sections within the Clustered SAN design, and as such, is inclusive of the pattern.

This topology utilizes a traditional SAN-based solution with two to four server nodes connected and clustered. The virtual machines all run within the Hyper-V cluster and utilize the networking infrastructure, whether they use a converged or nonconverged design.

As Figure 5 illustrates, the Cisco and EMC solution uses local on-motherboard LAN connections for management of the hosts. All other networking is handled via a converged fabric that configures the redundant connections into multiple, individual LANs for use by the different functions—for example, live migration and iSCSI.

Figure 5.   Cisco and EMC Reference Design Pattern
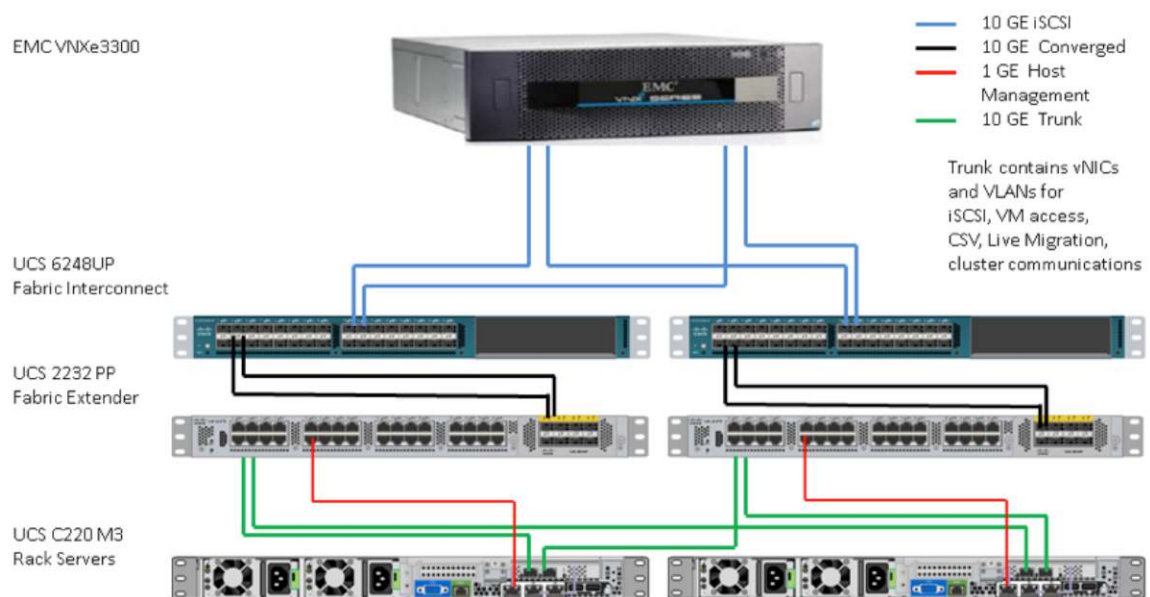
Compute

The Cisco Unified Computing System™ (Cisco UCS) combines Cisco UCS C-Series Rack Servers and B-Series Blade Servers with networking and storage access into a single converged system that simplifies management and delivers greater cost efficiency and agility with increased visibility and control. The latest expansion of the Cisco UCS portfolio includes the new Cisco UCS C220 M3 Rack Server (one rack unit [1RU]) and Cisco UCS C240 M3 Rack Server (2RU) and the Cisco UCS B200 M3 Blade Server. These three new servers increase compute density through more cores and cache balanced with more memory capacity, disk drives, and faster I/O. Together these server improvements and complementary Cisco UCS advancements deliver the best combination of features and cost efficiency required to support IT's diverse server needs. The Cisco and EMC solution is built upon the Cisco UCS C220 M3 Rack Server.

The Cisco UCS C220 M3 Rack Server, shown in the Figure 6, is designed for performance and density over a wide range of business workloads, from web serving to distributed databases to virtualization hosts. Building on the success of the Cisco UCS C200 M2 Rack Server, the enterprise-class Cisco UCS C220 M3 server further extends the capabilities of the Cisco UCS portfolio in a 1RU form factor with the addition of the Intel Xeon processor E5-2600 product family, which delivers significant performance and efficiency gains. In addition, the Cisco UCS C220 M3 server offers up to two Intel Xeon® processor E5-2600s, 16 DIMM slots, eight disk drives, and two 1 Gigabit Ethernet LAN-on-motherboard (LOM) ports, delivering outstanding density and performance in a compact package.

The Cisco UCS C220 M3 interfaces with Cisco UCS using another unique Cisco innovation: the Cisco UCS 1225 Virtual Interface Card (VIC). The Cisco UCS 1225 VIC is a virtualization-optimized Fibre Channel over Ethernet (FCoE) PCI Express (PCIe) 2.0 x 16 10-Gbps adapter designed for use with Cisco UCS C-Series servers. The VIC is a dual-port 10 Gigabit Ethernet PCIe adapter that can support up to 256 PCIe standards-compliant virtual interfaces, which can be dynamically configured so that both their interface type (network interface card [NIC] or host bus adapter [HBA]) and identity (MAC address and worldwide name [WWN]) are established using just-in-time provisioning. In addition, the Cisco UCS 1225 can support network interface virtualization and Cisco Data Center Virtual Machine Fabric Extender (VM-FEX) technology.

The Cisco and EMC solution uses Cisco UCS C220 M3 Rack Servers with the Intel Xeon processor E5-2600 product family. The E5-2600 product family has one or two processors from 1.8 to 3.3 GHz in speed and six or eight cores for each processor. This solution was validated with dual E2650 2.0-GHz processors with eight cores each.

**Figure 6.**   Cisco UCS C220 M3 Rack-Mount Server

The Cisco UCS C220 M3 offers up to 256 GB of RAM, up to eight drives or solid-state drives (SSDs) and two 1 Gigabit Ethernet LAN interfaces built into the motherboard. This solution was validated with 64 GB of RAM and with two 67-GB drives in a RAID 1 configuration. Additional memory can be configured if it is desired to run additional VMs on each host.

Cisco UCS Servers Change the Economics of the Data Center
IT infrastructure matters now more than ever, as organizations seek to achieve the full potential of infrastructure-as-a-service (IaaS), bare metal, virtualized servers, and cloud computing. Cisco continues to lead in data center innovation with the introduction of new building blocks for Cisco UCS that extend its exceptional simplicity, agility, and efficiency (see Figure 7). Cisco demonstrates leadership with new innovations such as the third-generation Cisco UCS C220 M3 Rack-Mount Server.

Figure 7.   Cisco UCS Components
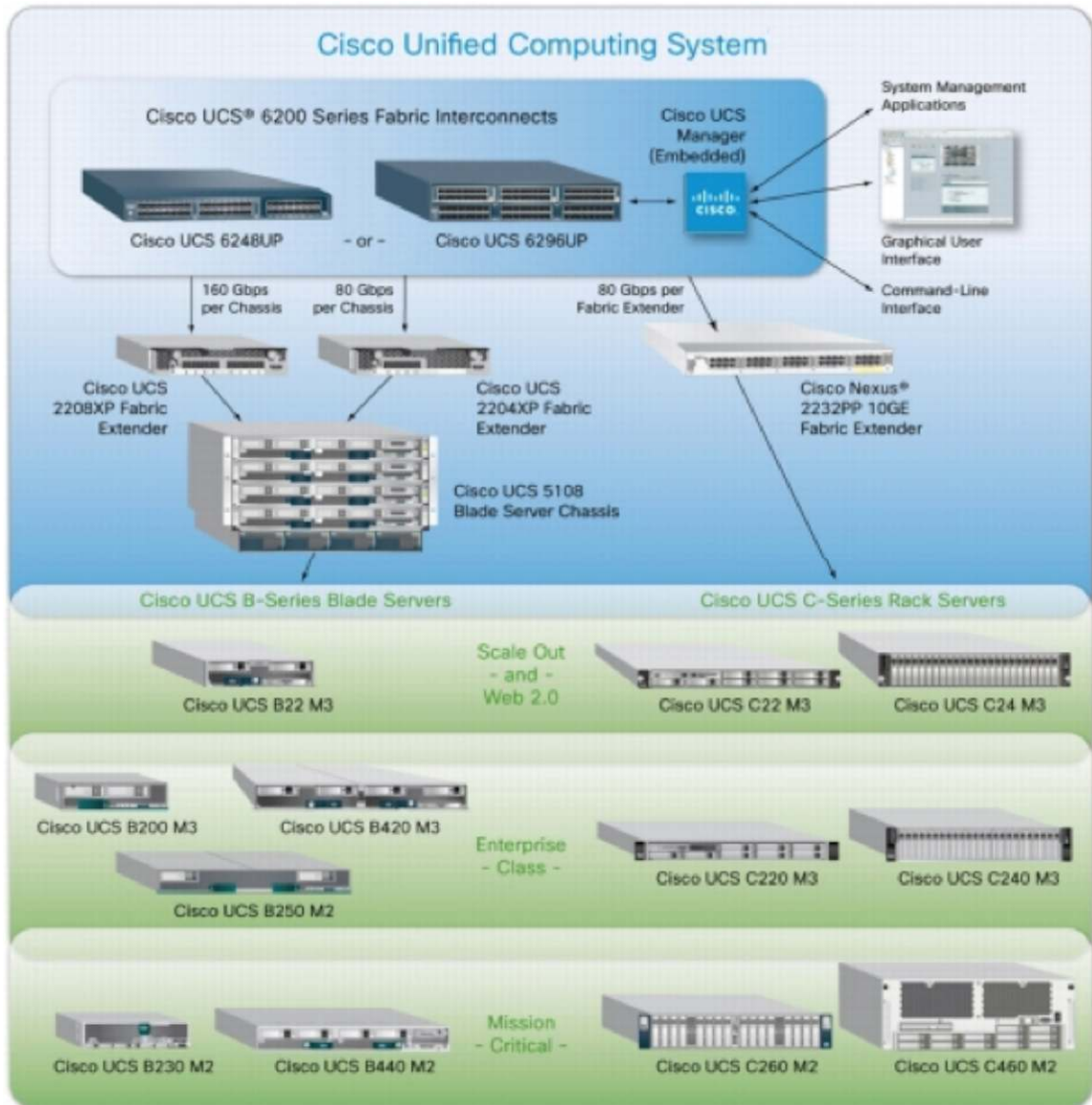
Cisco innovations, such as Cisco UCS Manager, allow administrators to create a software definition for a desired server (using Cisco service profiles and templates) and then instantiate that server and its I/O connectivity by associating a service profile with physical resources. This approach contrasts with the traditional approach of configuring each system resource manually, one at a time, through individual element managers. In contrast to the products of other vendors, Cisco service profiles can be moved from rack server to rack or blade server, or between blade or rack servers in different chassis. In other words, Cisco UCS Manager and service profiles are form-factor agnostic and can bridge blade chassis boundaries.

Other Cisco UCS building blocks include enhanced server I/O options and expanded Cisco UCS fabric interconnects that extend scalability and management simplicity for both blade and rack systems across bare-metal, virtualized, and cloud-computing environments. Cisco helps ensure that nearly all parts of Cisco UCS offer

investment protection and are backward compatible. For example, fabric extenders can be upgraded using the same fabric interconnects and the same Cisco UCS 1225 VIC. Fabric interconnect hardware can be upgraded independently of fabric extenders and blade chassis. Cisco continues to innovate in all these areas, helping ensure that both now and in the future, more powerful rack servers with larger, faster memory have adequate I/O bandwidth and compute power. Cisco completes this vision through continuous innovation in VIC, fabric extender, fabric interconnect, blade server, blade chassis, and rack server technologies and form-factor-agnostic Cisco UCS Manager Software.

The Cisco UCS C220 M3 is part of a large family of rack servers: the Cisco C-Series Rack Servers. Cisco UCS C-Series servers extend unified computing innovations to an industry-standard form factor to help reduce total cost of ownership (TCO) and increase business agility. Designed to operate both in standalone environments and as part of Cisco UCS, the Cisco UCS C-Series servers employ Cisco technology to help customers handle the most challenging workloads. The Cisco UCS C-Series complements a standards-based unified network fabric; Cisco Data Center VM-FEX virtualization support;, Cisco UCS Manager Software; Cisco fabric extender and fabric interconnect architectures; and Cisco Extended Memory Technology. Again, Cisco is innovating across all these technologies. With Cisco UCS architectural advantages, software advances, continuous innovation, and unique blade server and chassis designs, Cisco UCS is the first truly unified data center platform. In addition, Cisco UCS can transform IT departments through policy-based automation and deep integration with familiar systems management and orchestration tools.

High-Availability Requirements

The architecture requires that the server configuration provide for high availability and resiliency.

The Cisco and EMC solution provides high-availability throughout its design, ensuring there is no single point of failure in the solution. Dual-power domains are supported by servers, fabric interconnects, and storage systems. Multiple interconnects exist between servers, servers and storage, and within the storage infrastructure itself. Storage is provisioned from RAID protected pools, helping to ensure that single drive failures do not impact availability. Hot-spare drives are also configured within the storage infrastructure to mitigate exposures caused by single drive failures. All components are built into a Microsoft Server 2012 Failover Cluster environment that can accommodate anywhere from two to 64 nodes. The solution described in this paper is validated with four nodes, but the configuration can be easily made larger or smaller by the simple addition or removal of server and storage resources without invalidating the validation.

Microsoft Failover Cluster Service (FCS) helps you:

- Reduce planned maintenance downtime by eliminating user disruptions during scheduled maintenance periods. Virtual machines can be live-migrated with no downtime to another host in the cluster in order to do maintenance on a physical host. Once maintenance is completed, the physical host automatically rejoins the cluster and virtual machines can be moved onto it.

- Address the causes of unplanned downtime by scanning, isolating, and responding to unexpected server problems before they happen. FCS is constantly scanning the status of resources in the cluster—networks, storage, servers, virtual machines, and so forth. It automatically tries to resolve issues before declaring a failure, at which time it will automatically make changes in its configuration to ensure availability, such as transferring ownership of a disk drive from one host to another if a drive connection fails, while at the same time keeping the virtual machine available for use.

- Achieve high availability of servers, applications, and services by responding to data corruption dynamically and transparently without disrupting applications and services for users.

Network

The internal network also called the fabric is used to host and run the virtual machines. The high-availability network architecture is required to host and run the virtual machines, and to provide connectivity between the two cluster nodes in support of Windows Failover Clustering. The network architecture provides connectivity for service availability in the event of network interface failure.

The Cisco and EMC solution provides the necessary network connectivity. Two LAN-on-motherboard (LOM) 1 Gigabit Ethernet ports provide the connection for physical host management. Cisco UCS 1225 Virtual Interface Card (PCIe) is a 10 Gigabit Ethernet converged network adapter (CNA) that provides for creation of all other requisite NICs.

The Cisco UCS C220 M3 Rack-Mount Server has various CNA options. The UCS 1225 Virtual Interface Card option is used in this Cisco and EMC solution.

The Cisco Unified Computing System is a next-generation data center platform that unites computing, networking, storage access, and virtualization resources into a cohesive system designed to reduce TCO and increase business agility. The system integrates a low-latency, lossless 10 Gigabit Ethernet (10 GE) unified network fabric with enterprise-class, x86-architecture servers. The system is an integrated, scalable, multichassis, and/or rack platform in which all resources participate in a unified management domain.

Cisco UCS 6200 Series Fabric Interconnects

The Cisco UCS 6248UP is a member of the Cisco UCS 6200 Series of fabric interconnects. Cisco UCS 6200 Series Fabric Interconnects are a core part of the Cisco Unified Computing System, providing both network connectivity and management capabilities for the system. The Cisco UCS 6200 Series offers line-rate, low-latency, lossless 10 GE; Fibre Channel over Ethernet (FCoE); and FC functions.

The Cisco UCS 6200 Series provides the management and communication backbone for the Cisco UCS B-Series blade servers, the 5100 Series blade server chassis, and the Cisco UCS C-Series Rack servers that are properly configured to participate in the domain. All chassis, and therefore all blades, and all rack mount servers attached to the Cisco UCS 6200 Series fabric interconnects become part of a single, highly available management domain. In addition, by supporting unified fabric, the Cisco UCS 6200 Series provides both LAN and storage area network (SAN) connectivity for all blades and rack mounts within its domain.

From a networking perspective, the Cisco UCS 6200 Series uses a cut-through architecture that supports deterministic low-latency, line-rate 10 GE on all ports; switching capacity of 2 terabits (Tb); and 320 Gbps bandwidth per chassis, independent of packet size and enabled services. The product family supports Cisco low-latency, lossless 10 GE unified network fabric capabilities that increase the reliability, efficiency, and scalability of Ethernet networks. The fabric interconnect supports multiple traffic classes over a lossless Ethernet fabric from the blade or rack mount through the interconnect. Significant TCO savings come from an FCoE-optimized server design in which NICs, host bus adapters (HBAs), cables, and switches can be consolidated.

The Cisco UCS 6200 Series hosts and runs Cisco UCS Manager in a highly available configuration that enables the fabric interconnects to fully manage all Cisco UCS elements. Connectivity to the Cisco UCS C220 M3 Servers is maintained through the Cisco UCS 2232PP Fabric Extender mounted in the rack.

The Cisco UCS 6200 Series interconnects support out-of-band management through a dedicated 10/100/1000-Mbps Ethernet-management port, as well as in-band management. Cisco UCS Manager is typically deployed in a

clustered active-passive configuration on redundant fabric interconnects that are connected through dual 10/100/1000 Ethernet clustering ports.

The Cisco UCS 6248UP 48-port Fabric Interconnect shown Figure 8 is a one-rack unit (1RU) 10 Gigabit Ethernet, FCoE, and FC switch that offers up to 960 Gbps throughput and up to 48 ports. The switch has 32 1/10-Gbps fixed Ethernet, FCoE, and FC ports and one expansion slot.

Figure 8.   Cisco UCS 6248UP 48-Port Fabric Interconnect

Cisco UCS 2232PP Fabric Extenders

The Cisco Nexus® 2000 Series Fabric Extender, also known as FEX, is a highly scalable and flexible server networking solution that works with Cisco Nexus Series and Cisco UCS 6200 Series devices to provide high-density, low-cost connectivity for server aggregation. Scaling across 1 Gigabit Ethernet, 10 Gigabit Ethernet, unified fabric, rack, and blade server environments, the 2000 Series Fabric Extender is designed to simplify data center architecture and operations.

The 2000 Series Fabric Extender integrates with its parent switch, a Cisco Nexus or UCS 6200 Series device, to allow automatic provisioning and configuration taken from the settings on the parent device. This integration allows large numbers of servers and hosts to be supported using the same feature set as the parent device, including security and quality-of-service (QoS) configuration parameters, with a single management domain. The Fabric Extender and its parent switch enable a large multipath, loop-free, active-active data center topology without the use of Spanning Tree Protocol (STP).

The Cisco Nexus 2000 Series Fabric Extender forwards all traffic to its parent Cisco Nexus Series device over 10-Gigabit Ethernet fabric uplinks, which allows all traffic to be inspected by policies established on the Cisco Nexus Series device.

The Cisco Nexus 2000 Series Fabric Extenders behave as remote line cards for a parent Cisco Nexus switch or the UCS 6200 Series Fabric Interconnects. The fabric extenders essentially behave as extensions of the parent Cisco switch fabric. The parent switch is extended to connect to the server either as a remote line card or logically partition or virtualize adapter ports to connect to any type of server–rack and/or blades, with Cisco Adapter FEX and VM-FEX technologies.

The 1 RU Cisco Nexus 2232PP Fabric Extender, shown in Figure 9 has eight 10 Gigabit Ethernet fabric interfaces for uplink connections and 32 10 Gigabit Ethernet host server interfaces.

Figure 9.   Cisco UCS 2232PP

Cisco UCS 1225 Virtual Interface Card

A Cisco innovation, the Cisco UCS 1225 Virtual Interface Card is a virtualization-optimized Fibre Channel over Ethernet (FCoE) PCI Express (PCIe) 2.0 x 16 10-Gbps adapter designed for use with Cisco UCS C-Series Rack-Mount Servers. The virtual interface card is a dual-port 10 Gigabit Ethernet PCIe adapter that can support up to 256 PCIe standard-compliant virtual interfaces, which can be dynamically configured so that both their interface type (network interface card [NIC] or host bus adapter [HBA]) and identity (MAC address and worldwide name [WWN]) are established using just-in-time provisioning. In addition, the Cisco UCS 1225 can support network interface virtualization and Cisco VM-FEX technology.

Unique to the Cisco Unified Computing System, the Cisco UCS 1225 is optimized for virtualized environments, for organizations that seek increased mobility in their physical environments, and for data centers that want reduced TCO through NIC, HBA, cabling, and switch reduction.

The Cisco UCS 1225 can present up to 256 virtual interfaces on a given server. The 256 virtual interfaces can be dynamically configured by Cisco UCS Manager as either Fibre Channel or Ethernet devices (see Figure 10). With Cisco UCS 1225, deployment of applications that require or benefit from multiple Ethernet and Fibre Channel interfaces is no longer constrained by the available physical adapters.

To an operating system or a hypervisor running on a Cisco UCS C-Series Rack-Mount Server, the virtual interfaces appear as regular PCIe devices. In a virtualized environment, Cisco VM-FEX technology allows virtual links to be centrally configured and managed without the complexity that traditional approaches interpose with multiple switching layers in virtualized environments. I/O configurations and network profiles move along with virtual machines, helping increase security and efficiency while reducing complexity.

Another significant virtualization innovation is a technology known as hypervisor bypass. The Cisco UCS 1225 has built-in architectural support enabling the virtual machine to directly access the adapter such as Single Root I/O Virtualization (SR-IOV) or Cisco's VM-FEX. I/O bottlenecks and memory performance can be improved by giving virtual machines direct access to hardware I/O devices, eliminating the overhead of embedded software switches.

The Cisco UCS 1225 also brings adapter consolidation to physical environments. The adapter can be defined as multiple different NICs and HBAs. For example, one adapter can replace two quad-port NICs and two single-port HBAs, resulting in fewer NICs, HBAs, switches, and cables.

Figure 10. Cisco UCS 1225 Converged Adapter
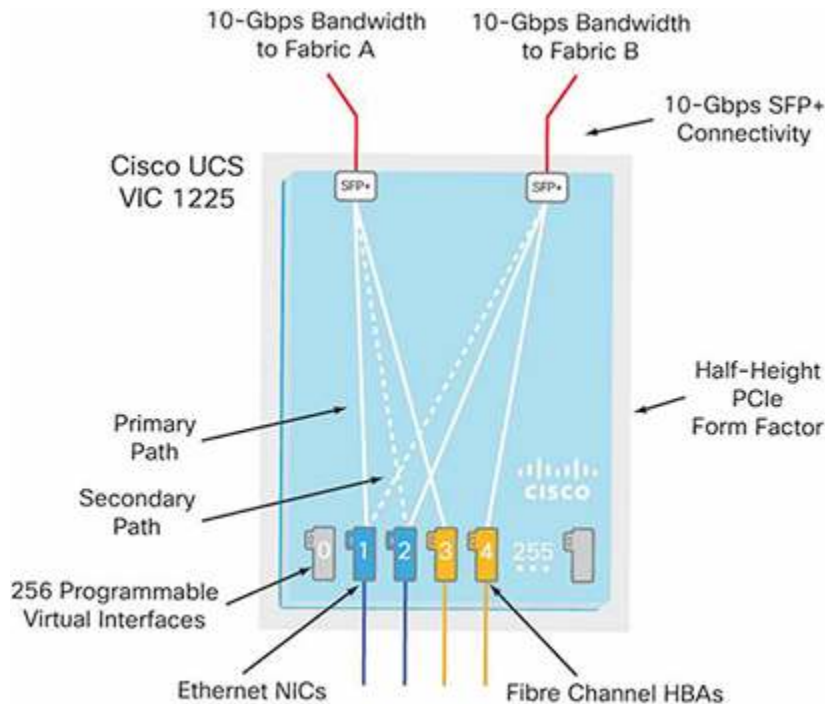
CSV Network Requirements

When Failover Clustering is enabled for Cluster Shared Volumes (CSV), the network may be used for communication between the nodes in support of the cluster file system. For information about the network traffic in a CSV enabled environment, watch for updates and information at http://technet.microsoft.com/en-us/library/hh831579.

When CSV is used, each node must have a network adapter capable of supporting CSV communication. In brief, the required configuration for CSV communication includes:

- **Hardware and system settings.** The storage configuration and hardware on the failover cluster are required to be identical. The cluster nodes used for live migration are required to have processors by the same manufacturer. As an alternative, the hardware and system settings are required to be as similar as possible to minimize potential problems.

- **Security policies.** Do not apply IPsec policies on the private network for live migration. This can significantly affect the performance of live migration. IPsec is required only when the live migration traffic needs to be encrypted.

- **IP subnet configuration.** For live migration, ensure that the virtual network on the source and destination nodes in the failover cluster is connected through the same IP subnet. This enables the virtual machine to retain the same IP address after live migration. For each network on a failover cluster in which CSV is enabled, all nodes must be on the same logical subnet. This means that multisite clusters that use CSV must use a VLAN.

To provide the correct access and appropriate performance, the Cisco and EMC solution provides a dedicated cluster communications network. The speed capability of the network matches the performance expectations of the solution.

High Availability and Number of Network Adapters
In order to add network availability and resiliency to the solution, the network design must have redundant paths to each server. Although the minimum requirement for the Small Implementations architecture is two physical network interfaces, one of those connections must be dedicated to failover cluster communications. This means that the other network interface must handle the rest of the network traffic. When more than two physical network interfaces are enabled, the remaining connections can be used to increase network availability.

The use of the NIC teaming capability in Windows Server 2012 combines two or more network interfaces to provide redundancy for the virtual machines. NIC teaming can enable multiple, redundant NICs and connections between servers and network switches. Teaming can be enabled via hardware or software-based approaches and enable multiple scenarios including path redundancy, failover, and load balancing.

The Windows Server 2012 SMB 3.0 file-sharing feature introduces a capability called multichannel. The SMB multichannel feature is similar to NIC teaming in that the SMB client and server can logically combine more than one network interface to provide increased scalability and additional network fault resiliency. This means that if a solution provides multiple network interfaces, the SMB client and server can take advantage of the additional resources.

The Cisco and EMC solution is configured with seven networks. The first network is the LAN-on-motherboard network that is used strictly for management of the physical hosts. The other six NICs are as follows:

- **VM access:** This network is a 10 GE connection used for access to communicate with the virtual machines.
- **Cluster communications:** This network is a 1 Gigabit Ethernet connection used exclusively for cluster communications among the nodes of the host cluster.
- **Live migration:** This network is a 10 Gigabit Ethernet connection defined with quality of service to be used for live migration of virtual machines from one host to another. It also is a secondary channel for cluster communications.
- **CSV:** This network is a 10 Gigabit Ethernet connection to be used for Cluster Shared Volume (CSV) traffic. It also is a secondary channel for cluster communications.
- **iSCSI-A and iSCSI-B:** These networks are 10 Gigabit Ethernet connections defined for redundant access via multipath I/0 (MPIO) to the VNXe300 storage array. They are defined with a quality of service level to ensure maximum performance.

Redundancy in the network is provided by virtual port channels in the networking infrastructure to which this environment would be connected. If virtual port channels are not used, the configuration can be changed to include dual networks for VM access, live migration, and CSV. These dual networks would make use of Windows Server 2012 Network Teaming capabilities to protect against single points of failure in the network.

Additional networks can easily be defined and added to the environment depending on the business needs of the customer. With no rewiring required, Cisco USC Manager can configure new service profiles that define different network configurations for the physical UCS C220 M3 Rack-Mount Servers. The Hyper-V Manager can be used to define virtual switches on the new networks.

Network Virtualization

Network virtualization is the process of combining hardware and software network resources and network functionality into a single, software-based administrative entity known as a virtual network. Network virtualization involves platform virtualization and it is often combined with resource virtualization.

Network virtualization is categorized as:

- External. Combines many networks or parts of networks into a virtual unit.
- Internal. Provides network-like functionality to the software containers on a single system.

Whether network virtualization is internal or external depends on the implementation provided by the vendors that support the technology.

Various equipment and software vendors offer network virtualization by combining any of the following:

- Network hardware such as switches and NICs
- Networks including VLANs and containers (virtual machines)
- Network storage devices
- Network media such as Ethernet and Fibre Channel

For more information about storage virtualization, visit the Storage Virtualization: The SNIA Technical Tutorial at: http://www.snia.org/education/storage_networking_primer/stor_virt/.

Virtual Machine Access Network

Virtual machine networks are dedicated to virtual machine LAN traffic. The Cisco and EMC solution has defined a 10 Gigabit Ethernet connection for access to the virtual machines.

Live Migration Network

Windows Server 2012 dramatically increases the performance of live migration, and it enables concurrent live migrations without limitation. The number of live migrations that can be performed is limited only by the investment in network infrastructure.

During live migration, the content of the memory of a running virtual machine that resides on the source node is transferred to the destination node over a LAN connection. In this case, a dedicated network is required to enable rapid live migration.

The Cisco and EMC solution has defined a 10 Gigabit Ethernet network that is dedicated to the host partition for live migration traffic. It has also been defined with quality of service to ensure optimal performance. It also serves as a secondary path for cluster communications should the cluster communications network fail.

Management Network

The management network is required to manage hosts using a dedicated network to prevent competition with guest traffic. This dedicated network provides a degree of separation for security and ease of management. The network typically dedicates one NIC per host and one port per network device to the management network.

Additionally, many server manufacturers also provide a separate out-of-band management capability that enables remote management of server hardware outside the host operating system.

The Cisco and EMC solution uses the Cisco UCS C220 M3 LAN on motherboard (LOM) exclusively for managing the host servers with out-of-band management. This is accessed through the Cisco Unified Computing System Manager console. This is a 1 Gigabit Ethernet connection and is a totally separate network from all the other networks defined on the converged network adapter.

Storage

For the storage architecture, third-party external storage systems such as a SAN are used to provide shared storage. These external storage platforms can use Fibre Channel, iSCSI (block-based), or SMB (file-based) connectivity for guest clustering. The SAN can utilize any supported drive type and can employ storage tiering.

For a SAN, storage network and network paths may be isolated using dedicated I/O adapters. Failover and scalability are achieved on the storage network using MPIO, while the TCP/IP network uses NIC teaming.

The Cisco and EMC solution implements a VNXe3300 SAN-based solution using a high-availability (HA) design to help ensure that business-critical data is always accessible. VNXe3300 offers N+1 redundant architecture, which provides data protection against any single component failure. With redundant components, including dual-storage-processors and dual-ported disk drives, the VNXe3300 system can overcome many different types of multiple component failure scenarios.

Storage Processors

In the VNXe3300 array, storage resources are distributed between the two storage processors; however, a storage resource is assigned to only one storage processor at a time. For example, a shared folder created on Storage Processor A will not be associated with Storage Processor B unless an storage processor failover occurs.

A storage processor fails-over when it reboots, experiences a hardware or software failure, or when a user places it in Service Mode. In this case, the storage resource fails-over to the peer storage processor. The surviving storage processor assumes ownership and begins servicing host I/O requests. When one storage processor services all I/O requests, performance between hosts and the VNXe3300 system can be degraded. Table 2 describes the storage processor events that can cause a failover.

Table 2.   Events That Cause Storage Processor Failover

Event	Response
Storage processor rebooting	The system or a user rebooted the storage processor. If the storage processor is healthy when it comes back online, the storage servers will fail back to it, and the system will return to normal. Check the System Health page to ensure that the storage processor is operating normally.
Storage processor in Service Mode	The system or a user placed the storage processor in Service Mode. A storage processor automatically enters Service Mode when it is unable to boot due to a hardware or system problem. Use the service actions on the Service System page to try to fix the problem. If the storage processor is healthy, you can reboot it to return it to Normal Mode.
Storage processor powered down	A user powered down the storage processor.
Storage processor failed	The storage processor failed and must be replaced.

Failover of the Management Interface

The management services for the VNXe3300 system run on one storage processor at a time, and it does not matter on which storage processor they run. In the event of a storage processor failure, the management server fails-over to the peer storage processor, and the management stack starts on the peer storage processor. Assuming both storage processors' management ports are cabled and on the same network, this process is not visible to the user, other than for a brief time when the management stack restarts on the peer storage processor.

If Unisphere is open in a web browser at the time of the storage processor failure, you see a pop-up message indicating a loss of connection to the failed storage processor. Another pop-up message appears when the connection is reestablished with the peer storage processor. The management stack remains on this storage processor even after the failed storage processor returns to a healthy state.

Power Modules

VNXe3300 systems have redundant power supplies, power cables, and battery backup units (BBUs) that protect data in the event of internal or external power failures. The VNXe3300 system employs dual-shared power supplies. If one power supply fails, the other one provides power to both storage processors.

The BBU does not function like an uninterruptable power supply (UPS), because it is not designed to keep the storage system up and running for long periods in anticipation of power being restored. With the BBU, the power to the storage processor is maintained. This protects the data in the cache and dumps it to the internal SSD vault when there is power loss. The BBU is sized to support the connected storage processor and is required to maintain the write cache long enough for it to be stored to the vault.

Memory

Each storage processor has its own dedicated system memory. The VNXe3300 has 12 GB per storage processor. This system memory is divided into storage processor memory and cache memory. Write cache memory is mirrored from one storage processor to its peer storage processor. Figure 11 shows a conceptual view of the VNXe3300 memory.

Figure 11.   Mirroring Write Cache Memory
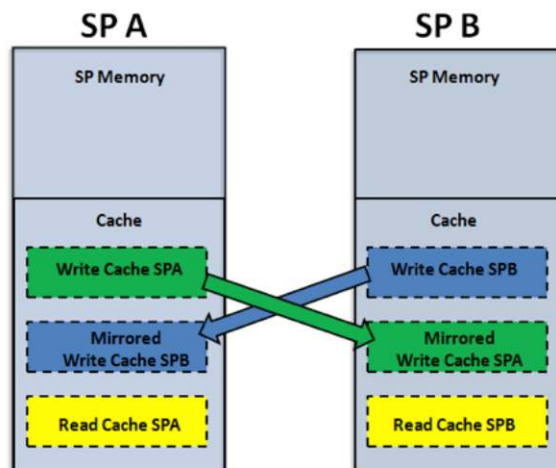
The storage processor cache is used for read and write caching. Read cache is for data that is held in memory in anticipation of it being requested in a future read I/O. Write cache stores write request data waiting to be written to a drive.

In addition to a storage processor's own read and write cache, cache memory contains a mirror copy of its peer storage processor's write cache. This is an important availability feature. The majority of all cache memory is allocated to write cache, with the remainder allocated to read cache. This occurs because the I/O for a read operation is often already in cache, unless the read operation is performing sequential read I/O.

Write Cache

The VNXe3300 write cache is a mirrored write-back cache. For every write, the data is stored in cache and copied to the peer storage processor. Then, the request is acknowledged to the host. In this process, the write cache is fully mirrored between the VNXe3300 system's storage processors to help ensure data protection through redundancy. In addition, requests are acknowledged before they are written to disk.

When the VNXe3300 system is shut down properly, the storage processor cache is flushed to backend drives and disabled. If a storage processor fails and then reboots, the cache is kept intact through all system resets.

If there is a power loss, each storage processor uses battery power to write its copy of the cache to its SSD (a flash-based hard drive), which does not need any power to retain the data. Upon reboot, the storage processor cache contents are restored on both storage processors. The two storage processors then determine the validity of the contents. Normally, both copies are valid. In the event that one storage processor has a newer version of the cache contents (or if one of them is invalid), the storage processor with the latest valid copy synchronizes its contents with the peer storage processor before reenabling the cache and allowing access to storage resources.

Fail-Safe Networking

All ports on the VNXe3300 system are configured automatically with Fail-Safe Networking (FSN). To take advantage of FSN, you must cable both storage processors the same way. There is a primary physical port on one storage processor and a secondary physical port on its peer storage processor. If the primary physical port or link fails, FSN fails over to the corresponding port on the peer storage processor. The data then gets routed internally through the inter-storage processor communications link, from the corresponding port on the peer storage processor to the storage processor associated with the storage resource. For example, if a given storage resource is accessed through eth2 on Storage Processor A and this port fails, FSN will fail-over traffic to eth2 port on Storage Processor B.

In the example shown in Figure 12, a shared folder was established on Storage Processor A (SP A) and is accessible through network port eth2. When the VNXe3300 system detects that the network link failed on SP A, it reroutes the shared folder data path from Storage Processor A to network port eth2 on Storage Processor B (SP B). It then routes the data internally from SP B to SP A. Once the failed network link recovers, the data path reverts to its original route.

Figure 12.   Fail-Safe Networking Invoked
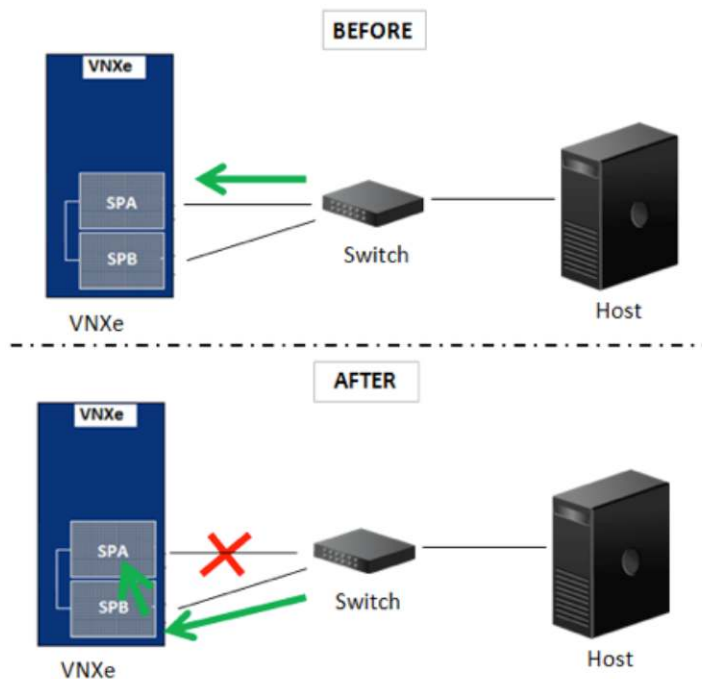
Implementation for iSCSI storage

To help ensure that there are redundant paths between the host and storage system, there must be a failover path in case the primary path fails. In an HA configuration, you should set up I/O connections from a host to more than one port on a storage processor and configure I/O connections between the host and peer storage processor as an additional safeguard. Having a host connect to more than one of the storage system's front-end ports is called multipathing.

When implementing an HA network for iSCSI storage, the following rules apply:

- A single iSCSI storage resource on a VNXe3300 system is presented on only one storage processor at a time.
- You can configure up to four IP interfaces for an iSCSI storage server. These IP interfaces are associated with two separate physical interfaces on the same storage processor.

Network switches should be configured on separate subnets. Servers cannot be attached directly to a VNXe3300 system.

SAN Architecture Design

The Cisco and EMC small implementation is configured with an iSCSI SAN. The Cisco UCS C-220 Rack-Mount Servers boot from iSCSI LUNs, and all cluster shared volumes are also presented via iSCSI. In addition to providing shared storage for the Hyper-V cluster, this also enables virtual machines direct access to iSCSI LUNs for failover clusters implemented at the virtual machine layer.

Performance

Storage performance is a complicated mix of drive, interface, controller, cache, protocol, SAN, host bus adapter (HBA), driver, and operating system considerations. Typically, the overall performance of the storage architecture is measured in terms of maximum throughput, maximum I/O operations per second (IOPS), and latency or response time. While each of the factors is important, IOPS and latency are highly relevant to server virtualization.

Many modern SANs use a combination of high-speed disks, slower-speed disks, and large memory caches. Storage controller cache can improve performance during burst transfers or when the same data is accessed frequently by storing it in the cache memory, typically several orders of magnitude faster than the physical disk I/O. However, cache is not a substitute for adequate disk spindles because cache is ineffective in aiding heavy write operations.

The VNXe3300 supports a range of drive technologies and RAID protection schemes. For the proposed solution, Cisco and EMC have implemented a base configuration that utilizes a single drive type and RAID protection scheme. The solution implements a total of 30 300-GB 15k RPM drives in a RAID 5 configuration. This design layout is shown in Figure 13.

Figure 13.   EMC VNXe3300 SAN Storage System
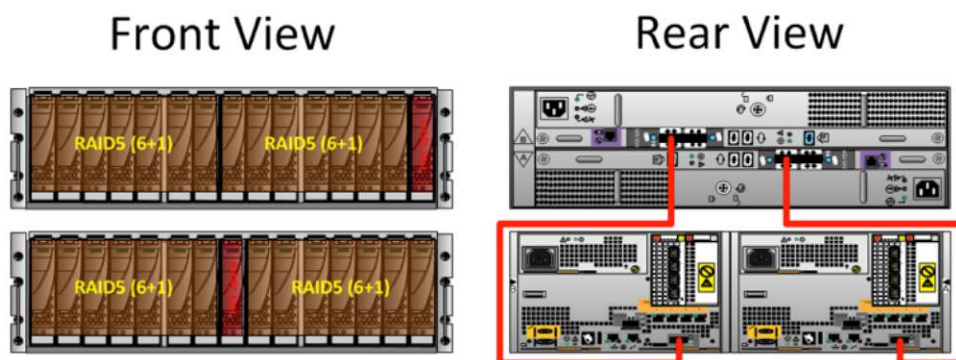
RAID 5 is implemented in multiple 6+1 drive sets (7 physical drives). For a total of 30 drives, this is implemented as 4 x 6+1 RAID sets. The remaining 2 drives are configured as Hot Spares, shown in red in Figure 13.

RAID 5 stripes data at a block level across several disks and distributes parity among the disks. With RAID 5, no single disk is devoted to parity. This distributed parity protects data if a single disk fails. Failure of a single disk reduces storage performance, so you should replace a failed disk immediately.

Performance of a RAID 5 stripe provides high performance for read workloads as all drives provide independent access patterns. Write workloads impose a RAID 5 penalty of a factor of 4 IOPs for each write IOP. This is a result of the read/modify/write operations that are implemented in RAID 5 solutions.

It is anticipated that the storage solution as provided will be necessary to service the expected cumulative read/write workload within this solution.

Drive Types

The type of hard drive utilized in the host server or in the storage array has a significant effect on the overall storage architecture performance. The critical performance factors for hard disks are:

- Interface architecture—for example, U320 SCSI, SAS, SATA

- Rotational speed of the drive, such as 7200, 10k, or 15k RPM
- Average latency in milliseconds

Other factors, such as the cache on the drive and support for advanced features, can also improve performance.

As with the storage connectivity, high IOPS and low latency are more critical than maximum sustained throughput when it comes to host server sizing and guest performance. The best practice is to select drives with the highest rotational speed, and lowest latency. Utilizing 15k RPM drives over 10k RPM drives can result in up to 35 percent more IOPS per drive.

The workloads targeted to run within the virtual machines can be used to determine acceptable disk subsystem latency.

The EMC VNXe3300 can support up to 150 drives, and supports multiple drive types including multiple form factors of 3.5–in. and 2.5–in. drives. All drives are SAS drives, and are implemented on dual-redundant, 6- Gbps SAS back ends.

The Cisco and EMC solution implements a configuration utilizing 30 x 3.5–in., 300-GB, 15k RPM drives.

RAID Array Design

The RAID type should provide both high availability and high performance, even during disk failures and RAID parity rebuilds. In general, RAID 10 (0+1), or a proprietary hybrid RAID type, is recommended for virtual machines volumes. RAID 1 is also acceptable for host boot volumes, although many proprietary RAID types and additional SAN capabilities may be employed. In general, the RAID type must be able to tolerate a single drive failure and not sacrifice performance for capacity.

Table 3. RAID Array Designs

Raid Type	Protects	Recommended Use	Disk Configuration VNXE3300
RAID 5* **Striped, distributed, parity protected**	Against single disk failure	Transaction processing; is often used for general-purpose storage, relational databases, and enterprise resource systems	(4+1), (6+1) A minimum of 5 disks for (4+1) or 7 disks for (6+1) must be allocated each time you allocate to a pool.
RAID 6** **Striped, distributed, double parity protected**	Against double disk failure	Same uses as RAID 5, only where increased fault tolerance is required	(4+2) A minimum of 6 disks must be allocated each time you allocate to a pool.
RAID 10 **Mirror protected**	Against multiple disk failures, as long as the disk failures do not occur in the same mirrored pair	RAID 10 may be more appropriate for applications with fast- or high-processing requirements, such as enterprise servers and moderate-sized database systems	(3+3) A minimum of 6 disks must be allocated each time you allocate to a pool.

* Flash drives as configured in RAID 5 (4+1) are only available for VNXe3150 and VNXe3300 systems

** RAID 6 is used by default for all NL-SAS drives in the VNXe3300 systems.

RAID 5

RAID 5 stripes data at a block level across several disks and distributes parity among the disks. With RAID 5, no single disk is devoted to parity. This distributed parity protects data if a single disk fails.

The failure of two disks in a RAID 5 disk group causes data loss and renders any storage in the RAID group unavailable. The failed disks must be replaced and the data then restored from a disk-based backup or accessed via a manual failover to a replicated system.

RAID 6

RAID 6 is similar to RAID 5; however, it uses a double-parity scheme that is distributed across different disks. This offers extremely high fault tolerance and disk-failure tolerance. This configuration provides block-level striping with parity data distributed across all disks. Arrays can continue to operate with up to two failed disks. If a single additional disk fails before the rebuild is complete, double parity gives time to rebuild the array without the data being at risk.

The failure of three disks in a RAID 6 disk group causes data loss and renders any storage in the RAID group unavailable. As with RAID 5, the failed disks must be replaced and the data then restored from a disk-based backup or accessed via a manual failover to a replicated system. RAID 6 provides high performance and reliability at medium cost, while providing lower capacity per disk.

RAID 10

This configuration requires a minimum of six physical disks to implement in VNXe3300 systems, where three mirrored sets in a striped set together provide fault tolerance. Although mirroring provides fault tolerance, failed disks must be replaced immediately and the array rebuilt.

A minimum of six disks can be allocated at a time to a pool, with three used strictly for mirroring. In other words, to provide redundancy, three disks out of every six are duplicates of the other three disks. This configuration is used in custom pools that are created using SAS disks.

Hot Sparing

Hot spares are spare disks that you can allocate when or after you configure your RAID groups. A hot spare replaces a failed disk in the event of a disk failure. When a disk drive fails, the VNXe system rebuilds data to the hot spare disk, using the remaining disks in the RAID group. When the failed disk is replaced, the VNXe3300 system copies the data and parity from the hot spare to the new drive. This process is called equalization. After equalization is completed, the hot spare returns to its default status and becomes ready for any future disk failure event.

Multipathing

Multipathing ensures that there is no single point of failure in the path to the storage. In addition, employing Microsoft Multipath I/O (MPIO) and Failover Clustering together as complimentary technologies mitigates the risk of a system outage both at the hardware and application levels.

Multipathing solutions use redundant physical path components, including adapters, cables, and switches, to create logical paths between the server and the storage device. If one or more of these components fails, thereby causing the path to fail, multipathing logic uses an alternative path for I/O allowing the applications to maintain uninterrupted access. Each NIC (for iSCSI) or HBA is connected using redundant switch infrastructures to provide continued access to storage in the event of a failure in a storage fabric component. Microsoft MPIO supports iSCSI, Fibre Channel, and SAS connectivity by establishing multiple sessions or connections to the storage array.

When SAN storage is used, storage vendors typically build a device specific module (DSM) containing their own MPIO software on top of the Microsoft Multipath I/O software. Each DSM and HBA has its own unique multipathing options and recommended number of connections.

Failover times vary by the storage vendor. They can be configured by using timers in the Microsoft iSCSI software initiator driver, or by modifying the Fiber Channel HBA driver parameter settings.

The VNXe3300 SAN array is equipped with four 10 Gigabit Ethernet connections that are used for iSCSI connections. There are two connections on each storage processor to ensure redundancy in case of a failure. These connections are connected to the server environment using Microsoft's MPIO feature to ensure no single point of failure from the host to the storage.

Storage Tiering

Storage tiering is used to physically partition data into multiple distinct classes such as price or performance. Data may be dynamically moved among classes in a tiered storage implementation based on access, activity, or other considerations. Storage tiering is normally achieved through a combination of varying types of disks that are used for different data types—for example, production, nonproduction, or backups.

For the Small Implementations configuration, the VNXe3300 storage array has been configured with a single class of storage technology. No automated storage tiering is included in the base configuration.

Deployment Automation

One of the key technologies used in the Cisco/EMC solution is the use of automation tools that are available as part of the Windows Server 2012 operating environment, and with tools available from Cisco.

Cisco UCS PowerTool

Cisco has placed comprehensive infrastructure management, including the management of Cisco UCS B-Series Blade Servers, Cisco UCS C-Series Rack-Mount Servers, and stand-alone Cisco UCS C-Series servers, under the control of Microsoft Windows PowerShell by developing a user-friendly CLI-based Cisco UCS management tool call Cisco UCS PowerTool. Microsoft Windows PowerShell is the task automation framework used across all Microsoft operating systems and applications, as well as a growing number of third-party platforms. With Cisco UCS PowerTool, you can more easily create scripts for Cisco UCS management tasks by using an extensive library of purpose-built commands called cmdlets. Cisco UCS PowerTool is based on the Cisco UCS Manager XML API framework and relies on standard Microsoft Windows PowerShell design principles, including:

- Inline help support
- Full pipelining support
- Fully classed object definition
- All legal verbs

With Cisco UCS PowerTool, your operations team can more easily tie together the management of storage components, computing components, and software applications into a custom, end-to-end management solution that is easy to use and easy to script. By using Cisco UCS PowerTool with similar tools based on Microsoft Windows PowerShell from third-party vendors to manage all components in the data center, your team can:

- Vastly streamline operations across the entire system
- Manage updates within the allotted time windows
- Protect systems and data
  - By simplifying update management, you help ensure that updates, patches, and firmware are applied regularly
  - Safeguard important information
  - Help ensure regulatory compliance
- Reduce operating and capital expenses

- Support continuity of IT services and business operations by reducing downtime and outages
- Significantly reduce your server-to-personnel ratio

The following table shows a sample of the operational agility that was achieved by managing Cisco UCS with Cisco UCS PowerTool.

Table 4.   Example of Management of a Cisco UCS Data Center with Cisco UCS PowerTool

Resource	Capacity
Administrators	1
Server remote monitoring agents	13
Server problems resolved	7
DIMM replacements	32
Physical locations	2
Number of servers	835
Number of firmware endpoint (BIOS, local storage controllers, network adapters, and Cisco UCS management infrastructure	28,800
Time to update firmware endpoints by using Microsoft PowerShell	Approximately 6 hours
Firmware updated without service disruption in a production environment	Yes

Cisco and EMC Bill of Materials

Table 5 lists the Cisco UCS hardware and supporting materials in the Cisco and EMC Microsoft Private Cloud for Small Implementations solution. Table 6 lists the EMC hardware and software in the solution.

Table 5.   Cisco Bill of Materials

Item	Qty
Hardware	
Cisco C220 M3 Rack-Mount Servers with 64 GB of memory and 2 Intel E5-2650 CPUs	4
Redundant Power Supplies (for the Cisco C220 M3)	4
Cisco UCS 1225 Virtual Interface Cards (adapters)	4
Cisco UCS 2232PP Fabric Extenders	2
Redundant Power Supplies (for Cisco UCS 2232)	2
Cisco UCS 6248UP Fabric Interconnects	2
Redundant Power Supplies (for the fabric interconnects)	2
3m LC-LC Fiber Optic Cables	8
1 ft. Cisco Catalyst 6000 Series Cables	2
3m Cisco Catalyst 6000 Series Cables	10
1m Twinax Cables to connect UCS 2232 to Fabric Interconnect	4
3m Twinax Cables to connect C220 to UCS 2232	8
SFP-10G-SR Fiber Tranceivers	8
GLC-T Transceivers	12
KVM cable for connecting keyboard, video, and mouse to the C220 servers	1

Table 6. EMC Bill of Materials

Item	Qty
Hardware	
VNXE3300 Rack	1
300-GB 15K SAS Drive	22
10-B Ethernet Optical Ultraflex IO Module	2
VNXE3300 3U DAE; 15X3.5 w/rack	1
VNXE3300; 2XSP DPE;15X3.5 DS;8X300GBSAS;AC; w/rack	1
RACK-40U-60 pwr cord US	1
Software	
VNXE3300 Base OE V2.0 (EMC ECOSYS) =IC	1
VNXE3300 Software Features	1

Management Systems Architecture

Active Directory Domain Services

Active Directory Domain Services (AD DS) is a required foundational component that is provided as a component of Windows Server 2012. Previous versions are not directly supported for all workflow provisioning and de-provisioning automation. It is assumed that AD DS deployments exist at the customer site and deployment of these services is not in scope for the typical deployment.

- AD DS in guest virtual machine. For standalone, business in-a-box configurations, the preferred approach is to run AD DS in a guest virtual machine, using the Windows Server 2012 feature that allows a Windows Failover Cluster to boot prior to AD DS running in the guest.

- Forests and domains. The preferred approach is to integrate into an existing AD DS forest and domain, but this is not a hard requirement. A dedicated resource forest or domain may also be employed as an additional part of the deployment. The Cluster-in-a-Box architecture supports multiple domains or multiple forests in a trusted environment using two-way forest trusts.

- Trusts (multidomain or inter-forest support). The Cluster-in-a-Box architecture enables multidomain support within a single forest in which two-way forest (Kerberos) trusts exist between all domains.

The Cisco and EMC solution is designed to integrate with an existing AD DS infrastructure. If this is for a new installation that does not have an AD infrastructure, virtual machines can be built as virtual machines on the cluster hosts. If you are using a 2008 or earlier AD DS infrastructure, the virtual machines running AD DS should not be configured as highly available virtual machines.

Server Manager

Server Manager is a management console in Windows Server 2012 that is used to provision and manage both local and remote Windows-based servers from desktops, without requiring either physical access to servers, or the need to enable Remote Desktop Protocol (RDP) connections to each server.

Server Manager has been completely redesigned for Windows Server 2012, to support remote, multi-server management, and help increase the number of servers that an administrator can manage.

Server Manager is recommended for managing remote servers in a mid-sized organization. It can be used to manage a larger numbers of servers, although the number of events that are collected and displayed need to be monitored and limited. For example, thousands of event log entries can result in delayed responses from Server Manager.

Server Manager Task Description

Server Manager makes server administration more efficient by allowing administrators to do tasks using a single tool. In Windows Server 2012, both standard users of a server and members of the Administrators group can perform management tasks in Server Manager. By default, standard users are prevented from performing some tasks.

For more information on managing Windows Server 2012 with Server Manager, visit: http://technet.microsoft.com/en-us/library/hh831456.

Conclusion

The Cisco and EMC Microsoft Private Cloud solution provides a highly scalable and reliable platform for a variety of virtualized workloads. The goal of this program is to help you quickly deploy a private cloud environment within your enterprise without the expense or risk associated with designing and building your own custom solution.

Appendix A: Reference Architecture Overview

Fast Track Version 3 for Small Implementations describes a reference architecture that uses multiple design patterns to provide hardware partners with more opportunities to match their hardware and software solutions.

The Fast Track Small Implementations architecture uses either Windows Server 2012 Standard Edition or Windows Server 2012 Datacenter Edition for enablement. For more information, see the Windows Server 2012 Editions licensing overview at: http://www.microsoft.com/en-us/server-cloud/windows-server/2012-editions.aspx. For information about licensing in virtual environments, see the Microsoft Volume Licensing Brief at: http://www.microsoft.com/en-us/download/details.aspx?displaylang=en&id=15113.

Design Patterns

The Fast Track reference architecture is defined for use with Windows Server 2012 Hyper-V. The Fast Track architecture takes advantage of the Windows Server 2012 advanced hardware capabilities and scenarios for small implementations. Due to the wide range of high-end servers, as well as lower-cost alternatives, these advanced capabilities have been summarized into multiple design patterns. The design patterns defined for the Small Implementations architecture include the Cluster-in-a-Box and Clustered SAN configurations.

The following sections provide a high-level overview of Design Pattern 2 as the pattern used by the Cisco and EMC solution, including the Windows Server 2012 features that are used and the specific hardware that is required.
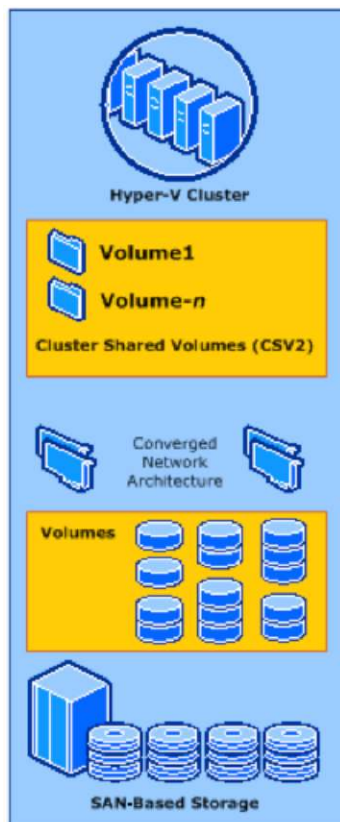
Design Pattern 2: Clustered SAN Overview

For Small Implementations architecture, the SAN design pattern can offer partners more flexibility to build solutions that meet end customer requirements. Key drivers for this SAN design include the need for an integrated hardware solution that provides simplified operations, reduction of device and port counts, and unification of the network and SAN fabric. The decision points for this design pattern include the tradeoff of making a larger capital expenditure to reduce the operational expenses for hardware management.

Similar to Cluster-in-a-box, the SAN design pattern provides a complete solution that can survive the failure and repair of single components. The compute, network, and storage requirements are the same as those defined in the Cluster-in-a-Box design, but with greater flexibility in the system enclosure choice and, more specifically, in the choice of the shared storage component.

A clustered SAN design uses the network to transport both LAN and storage traffic. This requires logical separation between these two network uses. This design pattern can use high-density blade compute enclosures with advanced hardware capabilities that support a clustered infrastructure, networking, and SAN storage. Failover and scalability are achieved through clustered blade enclosures and network architectures.

The Clustered SAN design pattern infrastructure is shown in Figure 14.

Figure 14. Clustered SAN Design Pattern

The key design considerations for the Clustered SAN design pattern are described next.

Compute
The architecture uses blade or rack-mounted servers to enable data center solutions such as a dedicated storage appliance.

Network
The high-availability network architecture consists of the connectivity between the two cluster nodes in support of Windows Failover Clustering. This is the network path for the workload inside the virtual machine. It is dependent on the types of workloads running and more bandwidth for the virtual machine traffic may be required.

Windows Failover Clustering identifies service and node outages and provides for failover during a service interruption. The network architecture includes node-to-node network connectivity and external network connectivity on separate physical connections.

The SAN design pattern allows for use of the network for access to the storage through transports such as iSCSI. In this instance, the use of the network for storage access must be separated from the cluster communications and the external, hosted network connectivity.

Storage

For the storage architecture, third-party external storage systems such as a SAN can be used to provide shared storage. These external storage platforms may use Fibre Channel or iSCSI (block-based storage) for clustering.

For a SAN, the storage network and network paths are isolated using dedicated I/O adapters. Failover and scalability are achieved on the storage network using MPIO, while the TCP/IP network uses NIC teaming. In this pattern, two 1 Gigabit Ethernet minimum or greater network adapters are used for primary connectivity to the shared storage network using Fibre Channel and/or iSCSI to enable advanced configurations.

Components and Features

While not a comprehensive list, the components that are expected for the SAN design pattern are described in the Table 7.

Table 7.   Expected Components for the Clustered SAN Design Pattern

Expected Components	Optional Components
• 1 GE or higher network connectivity • Redundant paths for storage networking components	• Storage array support for offloaded data transfer (ODX) • Additional Fibre Channel HBAs as required to support complex virtual Fibre Channel configurations within virtual machines • Addition of the following: • Single root I/O virtualization (SR-IOV) network card support • Remote direct memory access (RDMA) network connectivity • Certified Hyper-V Extensible vSwitch extension • Network adapter support for data center bridging (DCB) • SMI-S compliant management interfaces for storage components (hardware-based storage virtualization)

The Windows Server 2012 features and technologies for the Clustered SAN design pattern are described in Table 8.

Table 8.   Windows Server 2012 Features for the Clustered SAN Design Pattern

Windows Server 2012 Feature	Key Scenarios	References
Increased VP: LP Ratio		http://blogs.technet.com/b/windowsserver/archive/2012/04/05/windows-server-8-beta-hyper-v-amp-scale-up-virtual-machines-part-1.aspx
Increased Virtual Memory and Dynamic Memory	Support for up to 1-TB memory inside virtual machines.	http://technet.microsoft.com/en-us/library/hh831766.aspx
Virtual Machine Guest Clustering enhancements (Fibre Channel)	Support for virtual machine guest clusters using iSCSI connections with the Hyper-V Fibre Channel adapter to connect to shared storage.	http://technet.microsoft.com/en-us/library/hh831413.aspx

Windows Server 2012 Feature	Key Scenarios	References
Hyper-V Extensible Switch	A virtual Ethernet switch that allows for third-party filtering, capturing and forwarding extensions to be added to support additional virtual switch functionality to the Hyper-V platform.	http://technet.microsoft.com/en-us/library/hh831452.aspx http://blogs.technet.com/b/server-cloud/archive/2011/11/08/windows-server-8-introducing-hyper-v-extensible-switch.aspx http://channel9.msdn.com/posts/Edge-Byte-Windows-Server-8-Extensible-Switch-in-Hyper-V-Interview-with-Bob-Combs http://technet.microsoft.com/en-us/query/hh598173 http://msdn.microsoft.com/en-us/library/windows/hardware/hh582268(v=vs.85).aspx http://msdn.microsoft.com/en-us/library/windows/hardware/hh598183(v=vs.85).aspx http://msdn.microsoft.com/en-us/library/windows/hardware/hh598286(v=vs.85).aspx http://msdn.microsoft.com/en-us/library/windows/hardware/hh598296(v=vs.85).aspx
Encrypted Cluster Volumes	Enables support for BitLocker encrypted CSV2 volumes.	http://technet.microsoft.com/en-us/library/hh831713.aspx
Cluster Aware Updating	Provides the ability to apply updates to running failover clusters through coordinated patching of individual failover cluster nodes.	http://technet.microsoft.com/en-us/library/hh831694.aspx http://www.microsoft.com/download/en/details.aspx?id=29015 http://msdn.microsoft.com/en-us/library/windows/desktop/Hh706740(v=vs.85).aspx
Offloaded Data Transfer (ODX)	Support for storage level transfers using ODX technology (SAN feature).	http://technet.microsoft.com/en-us/library/hh831628.aspx http://blogs.technet.com/b/server-cloud/archive/2011/10/11/windows-server-8-hyper-v-overview.aspx
Single Root I/O Virtualization (SR-IOV) Network Support	Gives virtual machines near-native I/O against the physical network interface.	http://blogs.technet.com/b/jhoward/archive/2012/03/12/everything-you-wanted-to-know-about-sr-iov-in-hyper-v-part-1.aspx http://technet.microsoft.com/en-us/library/hh440148(v=vs.85).aspx
4k Physical Disk Support	Support for native 4k disk drives on hosts.	http://technet.microsoft.com/en-us/library/hh831459.aspx
Diskless Network Boot with iSCSI Target	The new iSCSI Target feature provides the network boot capability on commodity hardware using an iSCSI boot-capable network adapter or a software boot loader can be used (such as iPXE or netBoot/i).	http://technet.microsoft.com/en-us/library/hh831563.aspx http://blogs.technet.com/b/storageserver/archive/2011/05/04/diskless-servers-can-boot-and-run-from-the-microsoft-iscsi-software-target-using-a-regular-network-card.aspx
QoS Minimum Bandwidth (Fair Share)	Minimum bandwidth assigns a certain amount of bandwidth to a given type of traffic and ensures each type of network traffic receives up to its assigned bandwidth.	http://technet.microsoft.com/en-us/library/hh831511.aspx
Virtual Machine Storage Enhancements (VHDX)	Support for VHDX disks up to 64 TB in size.	http://technet.microsoft.com/en-us/library/hh831446.aspx
NIC Teaming (LBFO) Support	Support for switch independent and dependent load distribution using physical and virtual network connections.	http://technet.microsoft.com/en-us/library/hh831648.aspx http://go.microsoft.com/fwlink/?LinkId=215654
IPsec Offload	Supports network adapters equipped with hardware that reduces the CPU load by performing the computationally intensive work.	http://technet.microsoft.com/en-us/library/dd125367(v=WS.10).aspx
Datacenter Bridging	Enables hardware support for clustered fabrics, allowing bandwidth allocation and priority flow control.	http://technet.microsoft.com/en-us/library/hh849179

Appendix B: Virtualization Architecture

The virtualization architecture is provided at the server, network, and storage layers. Virtualization supports the pooling of resources at each layer and abstraction between the layers for greater efficiency. For the Small Implementations architecture, virtualization is enabled using Windows Server 2012, Hyper-V.

The Hyper-V virtualization architecture is summarized in the following sections.

Windows Server 2012, Hyper-V

The drivers for using Windows Server 2012, Hyper-V include providing effective solutions that enable higher asset utilization, improve system manageability, reduce energy consumption, and minimize data center and branch office facilities space, thereby lowering the total cost of ownership (TCO).

Hyper-V Dynamic Memory

Dynamic memory is a Hyper-V feature that allows the physical memory to be used more efficiently. This feature enables Hyper-V to use memory as a shared resource that can be reallocated automatically among running virtual machines. Dynamic memory adjusts the amount of memory available to each virtual machine based on the workload, including changes in memory demand and other factors.

For more information about Hyper-V Dynamic Memory, go to the Hyper-V Dynamic Memory Overview at http://technet.microsoft.com/en-us/library/hh831766.aspx.

Specific applications or workloads, particularly those with a built-in memory management capability, such as SQL Server or Exchange Server, might require workload-specific guidance. As an example, for SQL Server best practices guidance, go to Running SQL Server with Hyper-V Dynamic Memory at: http://msdn.microsoft.com/en-us/library/hh372970.aspx

Hyper-V Networking

The Hyper-V host cluster requires different types of network access, as described in Table 9.

Table 9.   Host Cluster Networks

Network Access Type	Purpose of the Network Access Type	Network traffic requirements	Recommended network access
Virtual Machine Access	Workloads running on virtual machines usually require external network connectivity to service client requests.	Varies	Public access that can be teamed for link aggregation or to fail-over the cluster.
Cluster and Cluster Shared Volumes	Preferred network used by the cluster for communications to maintain cluster health. Also, used by CSV to send data between owner and non-owner nodes. If storage access is interrupted, this network is used to access CSV or to maintain and back up CSV.		
The cluster should have access to more than one network for communication to ensure the cluster is highly available.	Usually low bandwidth and low latency. Occasionally, high bandwidth.	Private access	
Live Migration	Transfer virtual machine memory and state.	High bandwidth and low latency during migrations	Private access
Storage	Access storage through iSCSI or SAS (SAS does not need a network adapter).	High bandwidth and low latency	Usually, dedicated and private access. Refer to your storage vendor for guidelines.

Hyper-V Host Failover Cluster Design

Highly available host servers are one critical component of a dynamic, virtual infrastructure. A Hyper-V host failover cluster is a group of independent servers that work together to increase the availability of applications and services. The clustered servers (nodes) are connected physically. If one of the cluster nodes fails, another node begins to provide service. In the case of a planned live migration, users experience no perceptible service interruption.

Host Failover Cluster Topology

The server topology consists of one Hyper-V host cluster for the consolidated fabric that includes the management component. The Hyper-V host cluster is required to provide resource availability for the virtual machines that host the various parts of the management stack.

Hyper-V Guest Virtual Machine Design

Standardization is a key tenet of private cloud architectures and virtual machines. A standardized collection of virtual machine templates can both drive predictable performance and greatly improve capacity planning capabilities. An example of a basic virtual machine template library is defined in Table 10.

Table 10. Example of Virtual Machine Template Library

Template	Specs	Network	Operating System	Unit Cost
Template 1–Small	1 vCPU, 2-GB Memory, 50-GB Disk	VLAN 20	Window Server 2012	1
Template 2–Med	2 vCPU, 4-GB Memory, 100-GB Disk	VLAN 20	Window Server 2012	2
Template 3–X-Large	4 vCPU, 8-GB Memory, 200-GB Disk	VLAN 20	Window Server 2012	4
Template 1–Small	1 vCPU, 2-GB Memory, 50-GB Disk	VLAN 20	Window Server 2003 R2	1
Template 2–Med	2 vCPU, 4-GB Memory, 100-GB Disk	VLAN 20	Window Server 2003 R2	2
Template 3–X-Large	4 vCPU, 8-GB Memory, 200-GB Disk	VLAN 20	Window Server 2003 R2	4
Template 4–Small	1 vCPU, 2-GB Memory, 50-GB Disk	VLAN 10	Windows Server 2008	1
Template 5–Med	2 vCPU, 4-GB Memory, 100-GB Disk	VLAN 10	Windows Server 2008	2
Template 6–X-Large	4 vCPU, 8-GB Memory, 200-GB Disk	VLAN 10	Windows Server 2008	4

Appendix C: Cluster Shared Volumes

CSV provides shared access to the disk and a storage path for I/O fault tolerance (dynamic I/O redirection). In cases where the virtual machines have a direct I/O connection to the storage volumes, they only use the dedicated storage paths for disk I/O. When there is a failure of the storage path on one node, I/O traffic is rerouted from the primary node to a secondary node using SMB. Redirecting I/O to the secondary node provides a temporary failover path, while the primary storage path is brought back online. There is a minimal performance impact when the I/O runs this state.

In addition, any cluster communications network can be used with CSV. However, in the event of failover, the performance impact can be reduced when there are high-speed network connections between nodes, such as RDMA or 10 Gigabit Ethernet.

Note that there is a small amount of data that runs through the coordinator node, including metadata operations such as opening/closing files, shrinking/growing files, changing the security or attributes of the file, and so on.

CSV Limits

CSV uses Microsoft New Technology File System (NTFS) and it has no special hardware requirements beyond supported block-based shared storage. The limitations listed in Table 11 are imposed by the NTFS file system and inherited by CSV.

Table 11. CSV Limits

CSV Parameter	Limitation
Maximum Volume Size	256 TB
Maximum Number of Partitions	128

CSV Parameter	Limitation
Directory Structure	Unrestricted
Maximum Files per CSV	4+ Billion
Maximum VMs per CSV	Unlimited

CSV maintains metadata information about the volume access and requires that some I/O operations take place over the cluster communications network. One node in the cluster is designated as the coordinator node and is responsible for the disk operations.

CSV Volume Sizing

Because all cluster nodes can access all CSV volumes simultaneously, in a cluster with CSV, use standard LUN allocation methodologies based on performance and capacity requirements of the workloads running within the virtual machines.

Start by isolating the virtual machine operating system I/O from the application data I/O on separate LUNs, and address application-specific I/O considerations. For example, segregate databases and transaction logs and create storage volumes and/or storage pools that factor in the I/O profile, such as random read and write operations versus sequential write operations.

CSV architecture differs from other traditional clustered file systems in that it is free from typical scalability limitations. Prior to scaling the number of Hyper-V nodes or virtual machines on a CSV volume, ensure that the overall I/O requirements of the expected virtual machines running on the CSV are met by the underlying storage system and storage network.

While rare, disks and volumes can enter a state in which a check disk is required. In Windows Server 2012, CHKDSK can be used to perform spot fixes within seconds, and with CSV there is no disconnect at all to the VMs or other applications using the file.

Each enterprise application that is run within a virtual machine might have unique storage recommendations and require the vendor's virtualization-specific storage guidance. This also applies to CSV volumes. Note that CSV supports data deduplication for the operating system.

A storage pool or RAID array can contain many LUNs. If an enterprise application requires specific storage I/O operations per second (IOPS) or disk response times, all of the LUNs in use on the storage pool must be considered. An application that would require dedicated physical disks if run without virtualization, might require dedicated storage pools and CSV volumes running within a virtual machine.

When a dedicated network for CSV is required:

- For maximum flexibility, configure LUNs for CSV with a single volume such that 1 LUN equals 1 CSV.
- For I/O optimization or performance critical workloads, at least four CSVs per host cluster are recommended for segregating operating system I/O, random R/W I/O, sequential I/O, and other virtual machine-specific data.

Follow the storage vendor's recommendations for implementing CSV.

In addition, consider implementing the following CSV best practices:

- Create a standard size and IOPS profile for each type of CSV LUN to utilize for capacity planning. When additional capacity is needed, provision additional standard CSV LUNs.
- Prioritize the network used for CSV traffic. Designate a preferred network for cluster shared volumes communication.
- Add a CSV read cache for greater performance. For more information, go to How to Enable CSV Cache at: http://blogs.msdn.com/b/clustering/archive/2012/03/22/10286676.aspx. Note that if CSV read cache is enabled, data deduplication cannot be used.

CSV Design Configurations

In general, using CSV has many advantages such as improving the high-availability capability and the simplicity of the solution. However, it is not required for Small Implementation architectures. CSV is typically configured to enable special-purpose storage solutions such as virtual desktop infrastructure (VDI).

CSV can be implemented in several design configurations, including:

- Single CSV for each cluster
- Multiple CSV for each cluster
- Multiple I/O -optimized CSVs for each cluster

Each of these CSV design configurations is described in the following sections.

Single CSV for Each Cluster

In the single CSV per cluster design configuration (Figure 15), the storage array is configured to present a single large LUN to all the nodes in the host cluster. The LUN is configured as a CSV in Failover Clustering. All virtual machine-related files belonging to the virtual machines hosted on the cluster are stored on the CSV.
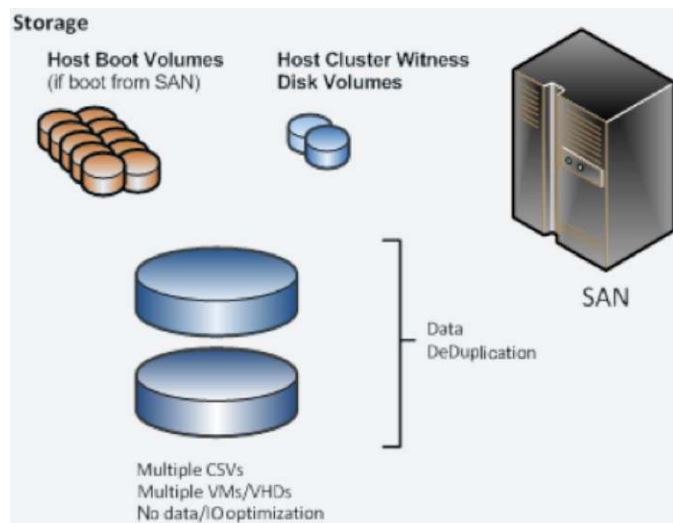
Figure 15. Virtual Machines on a Single Large CSV
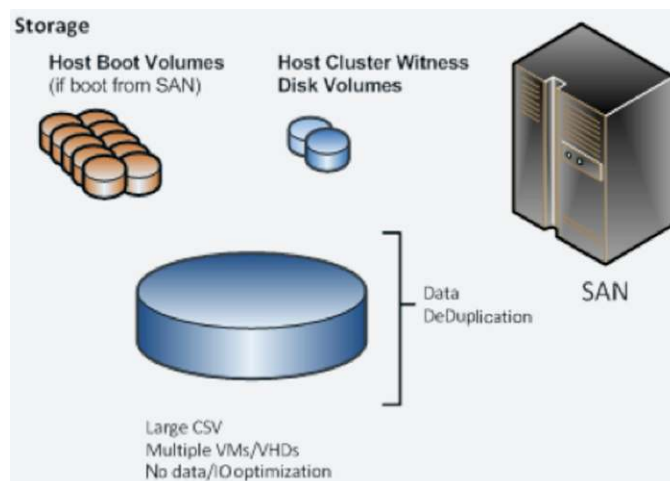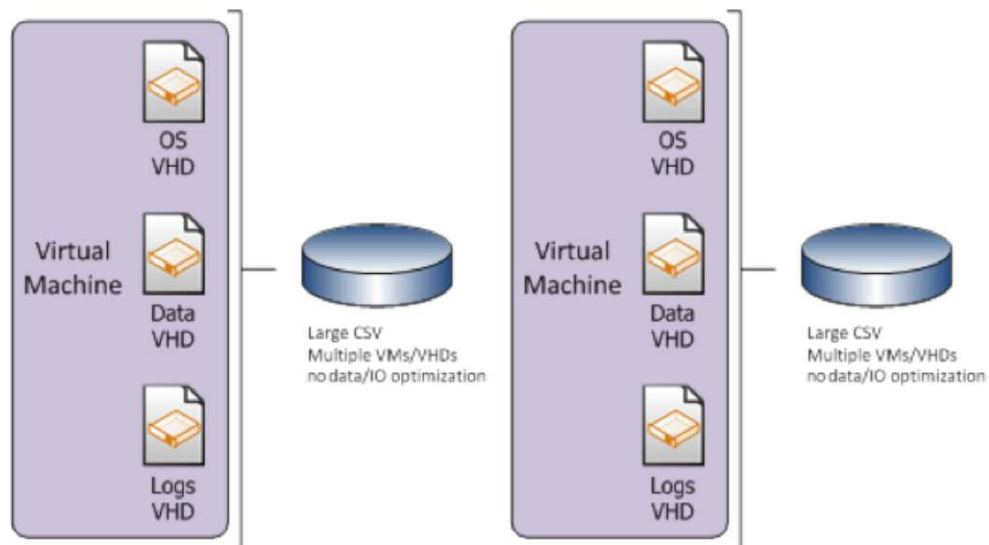

Multiple CSVs for Each Cluster

In the multiple CSV per cluster design configuration (Figure 16), the storage array is configured to present two or more large LUNs to all the nodes in the host cluster. The LUNs are configured as a CSV in Failover Clustering. All virtual machine-related files belonging to the virtual machines hosted on the cluster are stored on the CSVs.

Figure 16. Virtual Machines on Multiple CSVs with Minimal Segregation

For both the single and multiple CSV patterns, each CSV has the same I/O characteristics (Figure 17). Each individual virtual machine has all its associated virtual hard disks (VHDs) stored on one of the CSVs.
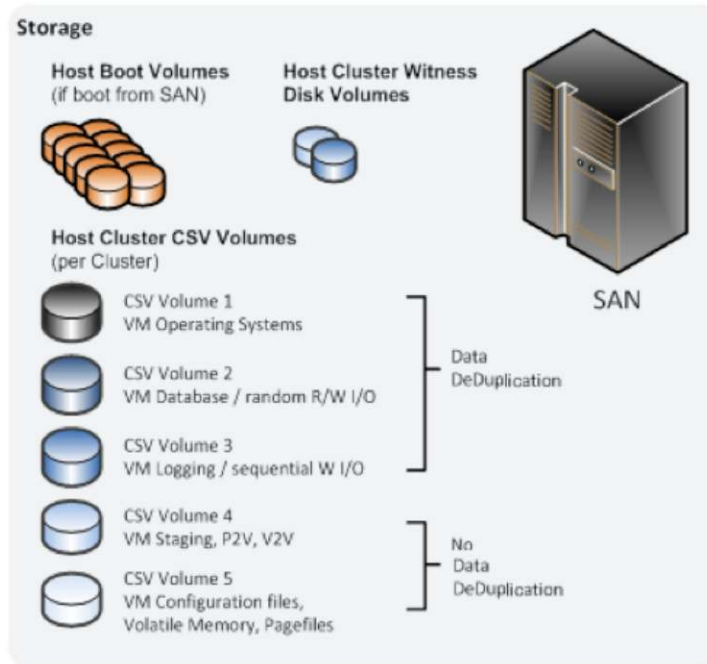
Figure 17. Each Virtual Machine's Virtual Disks Reside Together on the Same CSV


Multiple I/O Optimized CSVs for Each Cluster

In the multiple I/O optimized CSVs per cluster design configuration (Figure 18), the storage array is configured to present multiple LUNs to all the nodes in the host cluster. However, the LUNs are optimized for particular I/O patterns such as fast sequential read performance, or fast random write performance. The LUNs are configured as CSV in Failover Clustering. All VHDs belonging to the virtual machines hosted on the cluster are stored on the CSVs, but each VHD is stored on the appropriate CSV for the given I/O needs.

Figure 18.   Virtual Machines with a High Degree of Virtual Disk Segregation

In the multiple I/O optimized CSVs per cluster design configuration, each individual virtual machine has all its associated VHDs stored on the appropriate CSV per required I/O requirements (Figure 19).

Figure 19.   I/O Optimized CSVs with a High Degree of Virtual Disk Segregation
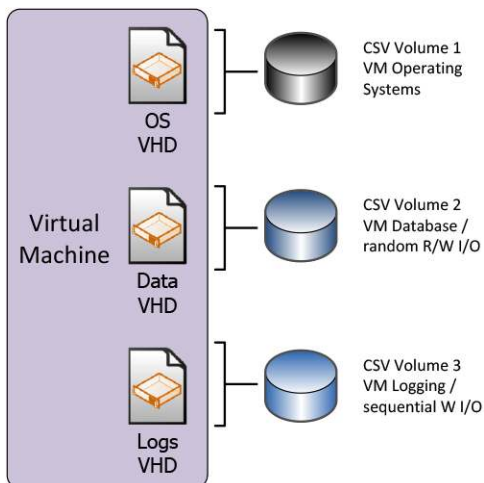
Note that a single virtual machine can have multiple VHDs, and each VHD can be stored on a different CSV (provided all CSVs are available to the host cluster on which the virtual machine is created).

CISCO

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

Printed in USA   C11-716894-02   03/13