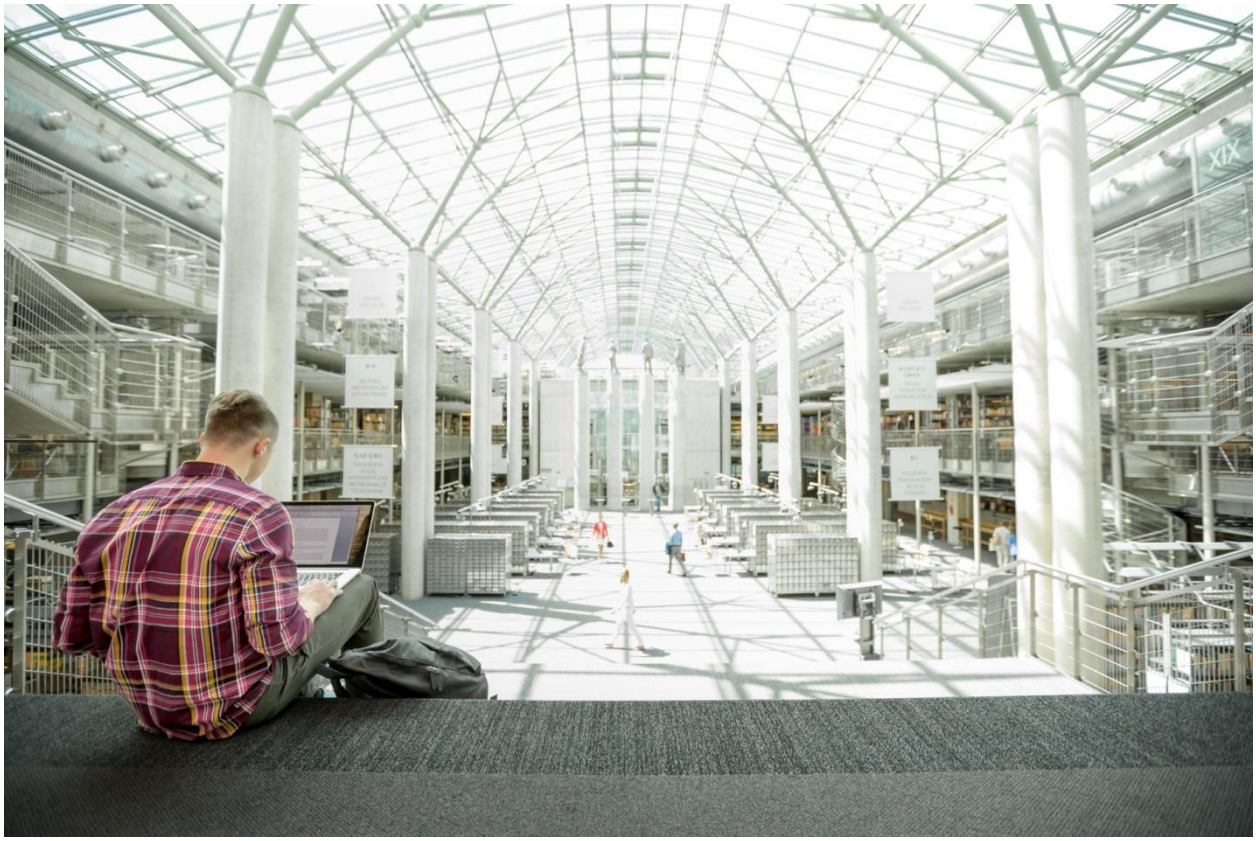


Cisco HyperFlex 3.5 All-Flash Systems for Deploying Microsoft SQL Server 2016 Databases with VMware ESXi 6.5

Deployment best practices and recommendations for Microsoft SQL Server 2016 Databases on Cisco HyperFlex 3.5 All-Flash Systems with VMware ESXi 6.5

Last Updated: December 6, 2018

CISCO VALIDATED DESIGN

About the Cisco Validated Design Program

The Cisco Validated Design (CVD) program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information, visit:

http://www.cisco.com/go/designzone.

Table of Contents

Executive Summary 5
Solution Overview 6
Introduction 6
Audience 6
Purpose of this Document 6
What's New? 6
Technology Overview 7
HyperFlex Data platform 3.5 – All-Flash Storage Platform 7
Architecture 8
Physical Infrastructure 9
Cisco Unified Computing System 9
Cisco UCS Fabric Interconnect 11
Cisco HyperFlex HX-Series Nodes 11
Cisco VIC 1227 and 1387 MLOM Interface Cards 12
Cisco HyperFlex Compute-Only Nodes 12
Cisco HyperFlex Systems Details 13
Why to use HyperFlex All-Flash systems for Database Deployments 21
Solution Design 22
Logical Network design 23
Storage Configuration for SQL Guest VMs 26
Deployment Planning 27
Datastore recommendation 27
SQL Virtual Machine configuration recommendation 28
Achieving Database High Availability 31
Microsoft SQL Server Deployment 38
Cisco HyperFlex 3.5 All-Flash System Installation and Deployment 38
Deployment Procedure 38
Solution Resiliency Testing and Validation 46
Node Failure Test 47
Fabric Interconnect Failure Test 47
Database Maintenance Tests 47
Database Performance Testing 49

Single Large VM Performance 49
Performance Scaling with Multiple VMs 50
Performance testing with HyperFlex Stretched Cluster 52
Common Database Maintenance Scenarios 56
Troubleshooting Performance 58
High SQL Guest CPU Utilization 58
High Disk latency on SQL Guest 58
Summary 59
About the Authors 60
Acknowledgements 60

Executive Summary

Cisco HyperFlex™ Systems deliver complete hyperconvergence, combining software-defined networking and computing with the next-generation Cisco HyperFlex Data Platform. Engineered on the Cisco Unified Computing System™ (Cisco UCS®), Cisco HyperFlex Systems deliver the operational requirements for agility, scalability, and pay-as-you-grow economics of the cloud—with the benefits of on-premises infrastructure. With a hybrid or All-flash-memory storage configurations and a choice of management tools, Cisco HyperFlex Systems deliver a pre-integrated cluster with a unified pool of resources that you can quickly deploy, adapt, scale, and manage to efficiently power your applications and your business.

With the latest All-Flash storage configurations, a low latency, high performing hyperconverged storage platform has become a reality. This makes the storage platform optimal to host the latency sensitive applications like Microsoft SQL Server. This document provides the considerations and deployment guidelines to have a Microsoft SQL server virtual machine setup on an All-Flash Cisco HyperFlex Storage Platform.

Solution Overview

Introduction

Cisco HyperFlex™ Systems unlock the potential of hyperconvergence. The systems are based on an end-to-end software-defined infrastructure, combining software-defined computing in the form of Cisco Unified Computing System (UCS) servers; software-defined storage with the powerful Cisco HX Data Platform and software-defined networking with the Cisco UCS fabric. Together with a single point of connectivity and hardware management, these technologies deliver a pre-integrated and an adaptable cluster that is ready to provide a unified pool of resources to power applications as your business needs dictate.

Microsoft SQL Server 2016 is a relational database engine release from Microsoft. It brings in a lot of new features and enhancements to the relational and analytical engines. It is built to provide a consistent and reliable database experience to applications delivering high performance. Currently more and more database deployments are getting virtualized and hyperconverged storage solutions are gaining popularity in the enterprise space. Cisco HyperFlex All-Flash system is the latest hyperconverged storage solution which was released recently. It provides a high performing and cost-effective storage solution making use of the high speed SSDs locally attached to the VMware ESXi hosts. It is crucial to understand the best practices and implementation guidelines that enable customers to run a consistently high performing SQL server database solution on a hyperconverged All-Flash solution.

Audience

This document can be referenced by system administrators, database specialists and storage architects, who work on planning, designing and implementing Microsoft SQL Server database solution on Cisco HyperFlex All-Flash storage solution.

Purpose of this Document

This document discusses reference architecture and implementation guidelines for deployment of SQL Server 2016 database instances on Cisco HyperFlex All-Flash solution.

What's New?

The list below provides few highlights of the solution discussed in this CVD.

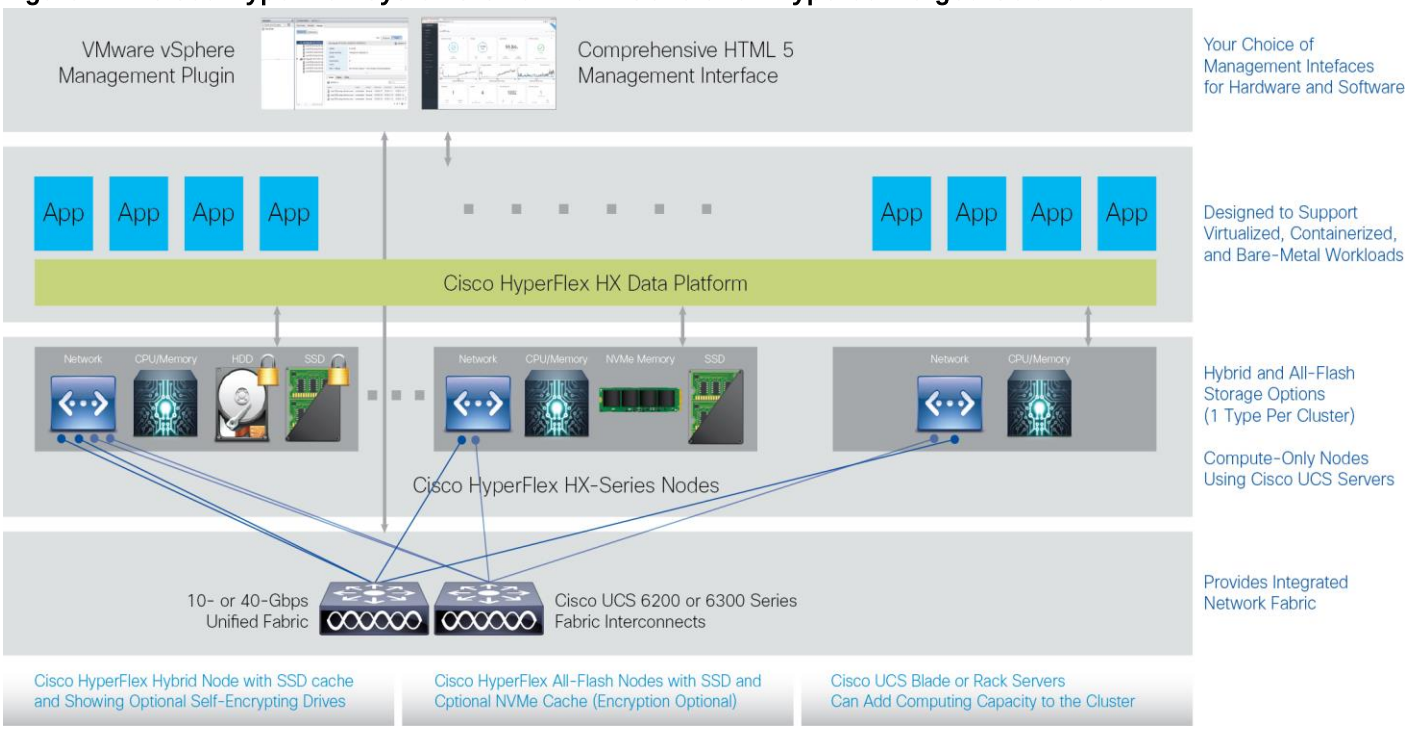
- Microsoft SQL Server 2016 deployment and validation on the latest HyperFlex version 3.5 with All-Flash cluster using Cisco UCS M5 servers.
- Microsoft SQL Server database performance validation on HyperFlex 3.5 Stretched Cluster feature.

Technology Overview

HyperFlex Data platform 3.5 – All-Flash Storage Platform

Cisco HyperFlex Systems are designed with an end-to-end software-defined infrastructure that eliminates the compromises found in first-generation products. Cisco HyperFlex Systems combine software-defined computing in the form of Cisco UCS® servers, software-defined storage with the powerful Cisco HyperFlex HX Data Platform Software, and software-defined networking (SDN) with the Cisco® unified fabric that integrates smoothly with Cisco Application Centric Infrastructure (Cisco ACI™). With All-Flash memory storage configurations, and a choice of management tools, Cisco HyperFlex Systems deliver a pre-integrated cluster that is up and running in an hour or less and that scales resources independently to closely match your application resource needs (Figure 1).

Figure 1 Cisco HyperFlex Systems Offer Next-Generation Hyperconverged Solutions



The Cisco HyperFlex All-Flash HX Data Platform includes:

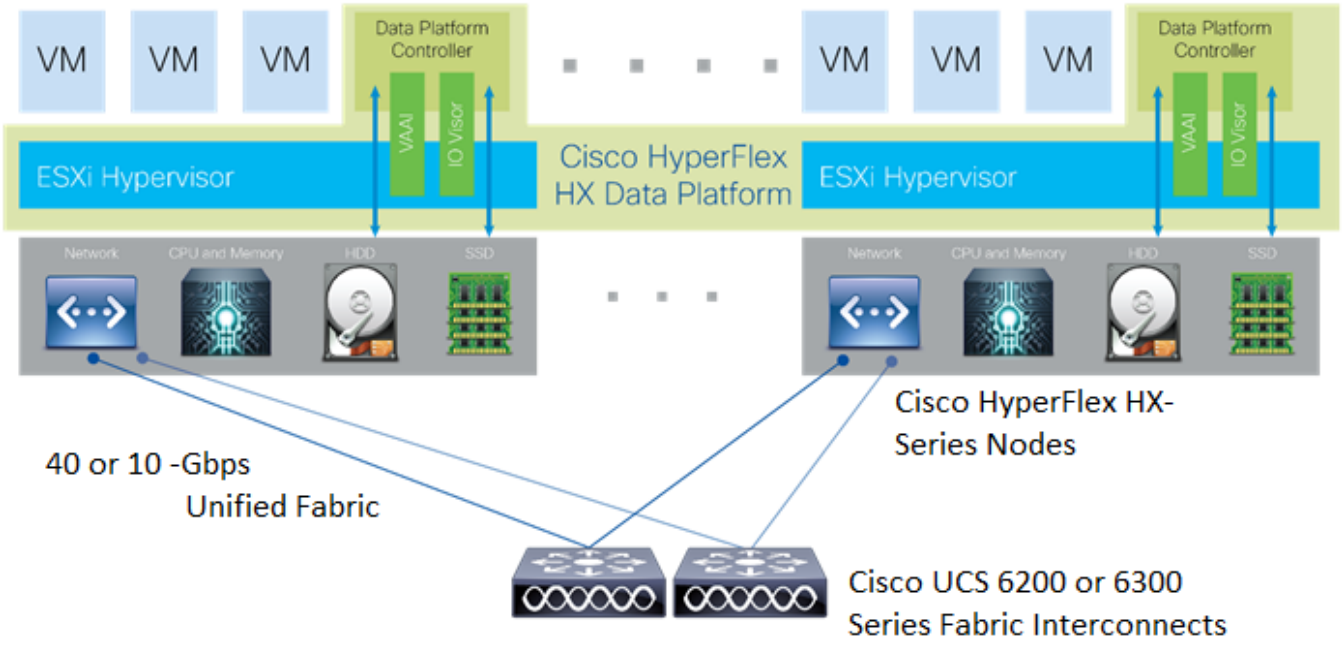
- Enterprise-class data management features that are required for complete lifecycle management and enhanced data protection in distributed storage environments–including replication, always on inline deduplication, always on inline compression, thin provisioning, instantaneous space efficient clones, and snapshots.
- Simplified data management that integrates storage functions into existing management tools, allowing instant provisioning, cloning, and pointer-based snapshots of applications for dramatically simplified daily operations.

- Improved control with advanced automation and orchestration capabilities and robust reporting and analytics features that deliver improved visibility and insight into IT operation.
- Independent scaling of the computing and capacity tiers, giving you the flexibility to scale out the environment based on evolving business needs for predictable, pay-as-you-grow efficiency. As you add resources, data is automatically rebalanced across the cluster, without disruption, to take advantage of the new resources.
- Continuous data optimization with inline data deduplication and compression that increases resource utilization with more headroom for data scaling.
- Dynamic data placement optimizes performance and resilience by making it possible for all cluster resources to participate in I/O responsiveness. All-Flash nodes use SSD drives for caching layer as well as capacity layer. This approach helps eliminate storage hotspots and makes the performance capabilities of the cluster available to every virtual machine. If a drive fails, reconstruction can proceed quickly as the aggregate bandwidth of the remaining components in the cluster can be used to access data.
- Enterprise data protection with a highly-available, self-healing architecture that supports non-disruptive, rolling upgrades and offers call-home and onsite 24x7 support options
- API-based data platform architecture that provides data virtualization flexibility to support existing and new cloud-native data types
- Cisco Intersight is the latest visionary cloud-based management tool, designed to provide a centralized off-site management, monitoring and reporting tool for all of your Cisco UCS based solutions including HyperFlex Cluster.

Architecture

In Cisco HyperFlex Systems, the data platform spans three or more Cisco HyperFlex HX-Series nodes to create a highly available cluster. Each node includes a Cisco HyperFlex HX Data Platform controller that implements the scale-out and distributed file system using internal flash-based SSD drives to store data. The controllers communicate with each other over 10 or 40 Gigabit Ethernet to present a single pool of storage that spans the nodes in the cluster (Figure 2). Nodes access data through a data layer using file, block, object, and API plug-ins. As nodes are added, the cluster scales linearly to deliver computing, storage capacity, and I/O performance.

Figure 2 Distributed Cisco HyperFlex System



In the VMware vSphere environment, the controller occupies a virtual machine with a dedicated number of processor cores and amount of memory, allowing it to deliver consistent performance and not affect the performance of the other virtual machines on the cluster. The controller can access all storage without hypervisor intervention through the VMware VM_DIRECT_PATH feature. In the All-Flash-memory configuration, the controller uses the node's memory, a dedicated SSD for write logging, and other SSDs for distributed capacity storage. The controller integrates the data platform into VMware software using two preinstalled VMware ESXi vSphere Installation Bundles (VIBs):

IO Visor: This VIB provides a network file system (NFS) mount point so that the VMware ESXi hypervisor can access the virtual disk drives that are attached to individual virtual machines. From the hypervisor's perspective, it is simply attached to a network file system.

VMware Storage API for Array Integration (VAAI): This storage offload API allows VMWare vSphere to request advanced file system operations such as snapshots and cloning. The controller causes these operations to occur through manipulation of metadata rather than actual data copying, providing rapid response, and thus rapid deployment of new application environments.

Physical Infrastructure

Cisco Unified Computing System

The Cisco Unified Computing System (Cisco UCS) is a next-generation data center platform that unites compute, network and storage access. The platform, optimized for virtual environments, is designed using open industry-standard technologies and aims to reduce the total cost of ownership (TCO) and increase the business agility. The system integrates a low-latency; lossless 10 or 40 Gigabit Ethernet unified network fabric with enterprise-class, x86-architecture servers. It is an integrated, scalable, multi-chassis platform in which all resources participate in a unified management domain.

The Cisco Unified Computing System consists of the following components:

- Compute - The system is based on an entirely new class of computing system that incorporates rack mount and blade servers based on Intel® Xeon® scalable processors product family.
- Network - The system is integrated onto a low-latency, lossless, 40-Gbps unified network fabric. This network foundation consolidates Local Area Networks (LAN's), Storage Area Networks (SANs), and high-performance computing networks which are separate networks today. The unified fabric lowers costs by reducing the number of network adapters, switches, and cables, and by decreasing the power and cooling requirements.
- Virtualization - The system unleashes the full potential of virtualization by enhancing the scalability, performance, and operational control of virtual environments. Cisco security, policy enforcement, and diagnostic features are now extended into virtualized environments to better support changing business and IT requirements.
- Storage access - The system provides consolidated access to both SAN storage and Network Attached Storage (NAS) over the unified fabric. It is also an ideal system for Software Defined Storage (SDS). Combining the benefits of single framework to manage both the compute and Storage servers in a single pane, Quality of Service (QOS) can be implemented if needed to inject IO throttling in the system. In addition, the server administrators can pre-assign storage-access policies to storage resources, for simplified storage connectivity and management leading to increased productivity. In addition to external storage, both rack and blade servers have internal storage which can be accessed through built-in hardware RAID controllers. With storage profile and disk configuration policy configured in Cisco UCS Manager, storage needs for the host OS and application data gets fulfilled by user defined RAID groups for high availability and better performance.
- Management - the system uniquely integrates all system components to enable the entire solution to be managed as a single entity by the Cisco UCS Manager. The Cisco UCS Manager has an intuitive graphical user interface (GUI), a command-line interface (CLI), and a powerful scripting library module for Microsoft PowerShell built on a robust application programming interface (API) to manage all system configuration and operations.

The Cisco Unified Computing System is designed to deliver:

- Reduced Total Cost of Ownership and increased business agility.
- Increased IT staff productivity through just-in-time provisioning and mobility support.
- A cohesive, integrated system which unifies the technology in the data center. The system is managed, services and tested as a whole.
- Scalability through a design for hundreds of discrete servers and thousands of virtual machines and the capability to scale I/O bandwidth to match the demand.
- Industry standard supported by a partner ecosystem of industry leaders.

Cisco UCS Fabric Interconnect

The Cisco UCS Fabric Interconnect (FI) is a core part of the Cisco Unified Computing System, providing both network connectivity and management capabilities for the system. Depending on the model chosen, the Cisco UCS Fabric Interconnect offers line-rate, low-latency, lossless 10 Gigabit or 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE) and Fibre Channel connectivity. Cisco UCS Fabric Interconnects provide the management and communication backbone for the Cisco UCS C-Series, S-Series and HX-Series Rack-Mount Servers, Cisco UCS B-Series Blade Servers and Cisco UCS 5100 Series Blade Server Chassis. All servers and chassis, and therefore all blades, attached to the Cisco UCS Fabric Interconnects become part of a single, highly available management domain. In addition, by supporting unified fabrics, the Cisco UCS Fabric Interconnects provide both the LAN and SAN connectivity for all servers within its domain.

From a networking perspective, the Cisco UCS 6200 Series uses a cut-through architecture, supporting deterministic, low latency, line rate 10 Gigabit Ethernet on all ports, up to 1.92 Tbps switching capacity and 160 Gbps bandwidth per chassis, independent of packet size and enabled services. The product family supports Cisco low - latency, lossless 10 Gigabit Ethernet unified network fabric capabilities, which increase the reliability, efficiency, and scalability of Ethernet networks. The Fabric Interconnect supports multiple traffic classes over the Ethernet fabric from the servers to the uplinks. Significant TCO savings come from an FCoE-optimized server design in which network interface cards (NICs), host bus adapters (HBAs), cables, and switches can be consolidated.

The Cisco UCS 6300 Series offers the same features while supporting even higher performance, low latency, lossless, line rate 40 Gigabit Ethernet, with up to 2.56 Tbps of switching capacity. Backward compatibility and scalability are assured with the ability to configure 40 Gbps quad SFP (QSFP) ports as breakout ports using 4x10GbE breakout cables. Existing Cisco UCS servers with 10GbE interfaces can be connected in this manner, although Cisco HyperFlex nodes must use a 40GbE VIC adapter in order to connect to a Cisco UCS 6300 Series Fabric Interconnect.

Below list provides the supported Cisco UCS Fabric Interconnects by the HyperFlex System. For more specifications on the below Fabric Interconnects, see: https://www.cisco.com/c/en/us/products/servers-unified-computing/fabric-interconnects.html#~stickynav=2

- Cisco UCS 6248UP
- Cisco UCS 6296UP
- Cisco UCS 6332
- Cisco UCS 6332-16UP

Cisco HyperFlex HX-Series Nodes

A HyperFlex cluster requires a minimum of three HX-Series **“converged”** nodes (with disk storage). Data is replicated across at least two of these nodes, and a third node is required for continuous operation in the event of a single-node failure. Each node that has disk storage is equipped with at least one high-performance SSD drive for data caching and rapid acknowledgment of write requests. Each node also is equipped with additional disks, up to **the platform’s physical** limit, for long term storage and capacity.

Variety of HX-Series converged All-Flash nodes are supported by HyperFlex All-Flash system. The list below provides the supported HX-Series All-Flash converged nodes. For specifications of below listed products,

see: https://www.cisco.com/c/en/us/products/hyperconverged-infrastructure/hyperflex-hx-series/index.html#models

- Cisco HyperFlex HXAF220c-M5SX All-Flash Node
- Cisco HyperFlex HXAF240c-M5SX All-Flash Node
- Cisco HyperFlex HXAF220c-M4S All-Flash Node
- Cisco HyperFlex HXAF240c-M4SX All-Flash Node

Cisco VIC 1227 and 1387 MLOM Interface Cards

The Cisco UCS Virtual Interface Card (VIC) 1227 is a dual-port Enhanced Small Form-Factor Pluggable (SFP+) 10-Gbps Ethernet and Fibre Channel over Ethernet (FCoE)-capable PCI Express (PCIe) modular LAN-on-motherboard (mLOM) adapter installed in the Cisco UCS HX-Series Rack Servers. The VIC 1227 is used in conjunction with the Cisco UCS 6248UP or 6296UP model Fabric Interconnects.

The Cisco UCS VIC 1387 Card is a dual-port Enhanced Quad Small Form-Factor Pluggable (QSFP+) 40-Gbps Ethernet and Fibre Channel over Ethernet (FCoE)-capable PCI Express (PCIe) modular LAN-on-motherboard (mLOM) adapter installed in the Cisco UCS HX-Series Rack Servers. The VIC 1387 is used in conjunction with the Cisco UCS 6332 or 6332-16UP model Fabric Interconnects.

The mLOM slot can be used to install a Cisco VIC without consuming a PCIe slot, which provides greater I/O expandability. It incorporates next-generation converged network adapter (CNA) technology from Cisco, providing investment protection for future feature releases. The card enables a policy-based, stateless, agile server infrastructure that can present up to 256 PCIe standards-compliant interfaces to the host, each dynamically configured as either a network interface card (NICs) or host bus adapter (HBA). The personality of the interfaces is set programmatically using the service profile associated with the server. The number, type (NIC or HBA), identity (MAC address and World Wide Name [WWN]), failover policy, adapter settings, bandwidth, and quality-of-service (QoS) policies of the PCIe interfaces are all specified using the service profile.

Hardware revision V03 or later of the Cisco VIC 1387 is required for the Cisco HyperFlex HX-series servers.

Cisco HyperFlex Compute-Only Nodes

All current model Cisco UCS M4 and M5 generation servers, except the C880 M4 and C880 M5, may be used as compute-only nodes connected to a Cisco HyperFlex cluster, along with a limited number of previous M3 generation servers. Any valid CPU and memory configuration is allowed in the compute-only nodes, and the servers can be configured to boot from SAN, local disks, or internal SD cards. The following servers may be used as compute-only nodes:

Cisco UCS B200 M3 Blade Server

Cisco UCS B200 M4 Blade Server

Cisco UCS B200 M5 Blade Server

Cisco UCS B260 M4 Blade Server

Cisco UCS B420 M4 Blade Server

Cisco UCS B460 M4 Blade Server

Cisco UCS B480 M5 Blade Server

Cisco UCS C220 M3 Rack Mount Servers

Cisco UCS C220 M4 Rack Mount Servers

Cisco UCS C220 M5 Rack Mount Servers

Cisco UCS C240 M3 Rack Mount Servers

Cisco UCS C240 M4 Rack Mount Servers

Cisco UCS C240 M5 Rack Mount Servers

Cisco UCS C460 M4 Rack Mount Servers

Cisco UCS C480 M5 Rack Mount Servers

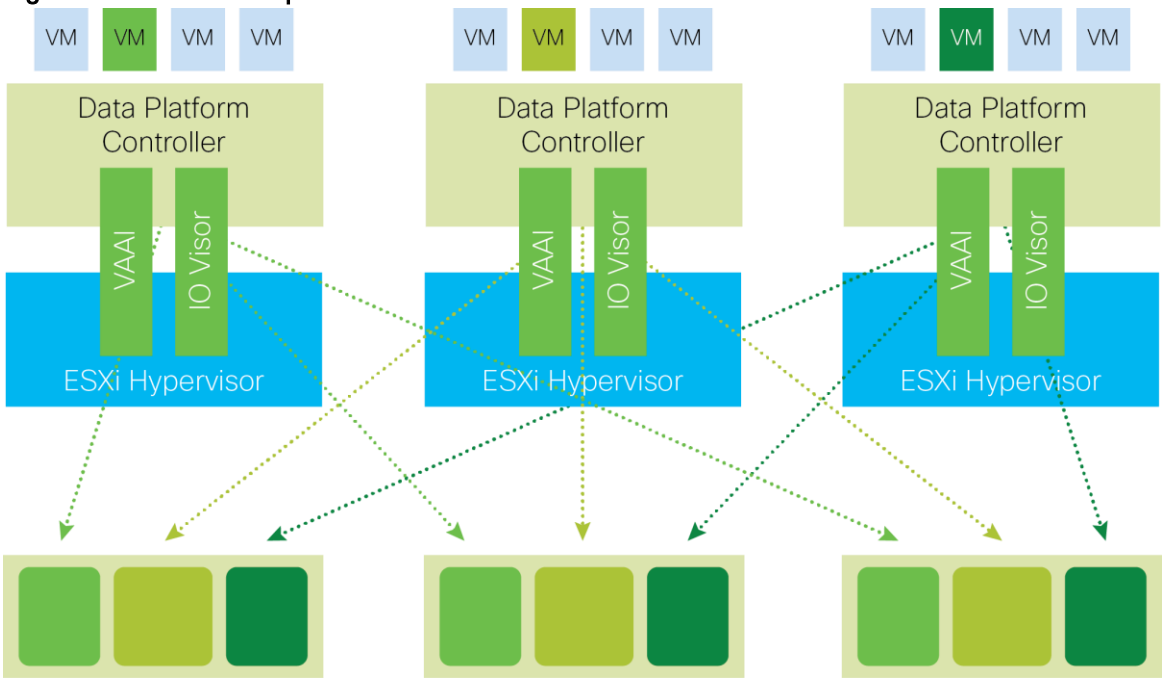
Cisco HyperFlex Systems Details

Engineered on the successful Cisco UCS platform, Cisco HyperFlex Systems deliver a hyperconverged solution that truly integrates all components in the data center infrastructure–compute, storage, and networking. The HX Data Platform starts with three or more nodes to form a highly available cluster. Each of these nodes has a software controller called the Cisco HyperFlex Controller. It takes control of the internal flash-based SSDs or a combination of flash-based SSDs and HDDs to store persistent data into a single distributed, multitier, object-based data store. The controllers communicate with each other over low-latency 10 or 40 Gigabit Ethernet fabric, to present a single pool of storage that spans across all the nodes in the cluster so that data availability is not affected if single or multiple components fail.

Data Distribution

Incoming data is distributed across all nodes in the cluster to optimize performance using the caching tier (Figure 3). Effective data distribution is achieved by mapping incoming data to stripe units that are stored evenly across all nodes, with the number of data replicas determined by the policies you set. When an application writes data, the data is sent to the appropriate node based on the stripe unit, which includes the relevant block of information. This data distribution approach in combination with the capability to have multiple streams writing at the same time avoids both network and storage hot spots, delivers the same I/O performance regardless of virtual machine location, and gives you more flexibility in workload placement. This contrasts with other architectures that use a data locality approach that does not fully use available networking and I/O resources and is vulnerable to hot spots.

Figure 3 Data is Striped Across Nodes in the Cluster



When moving a virtual machine to a new location using tools such as VMware Dynamic Resource Scheduling (DRS), the Cisco HyperFlex HX Data Platform does not require data to be moved. This approach significantly reduces the impact and cost of moving virtual machines among systems.

Data Operations

The data platform implements a distributed, log-structured file system that changes how it handles caching and storage capacity depending on the node configuration.

In the All-Flash-memory configuration, the data platform uses a caching layer in SSDs to accelerate write responses, and it implements the capacity layer in SSDs. Read requests are fulfilled directly from data obtained from the SSDs in the capacity layer. A dedicated read cache is not required to accelerate read operations.

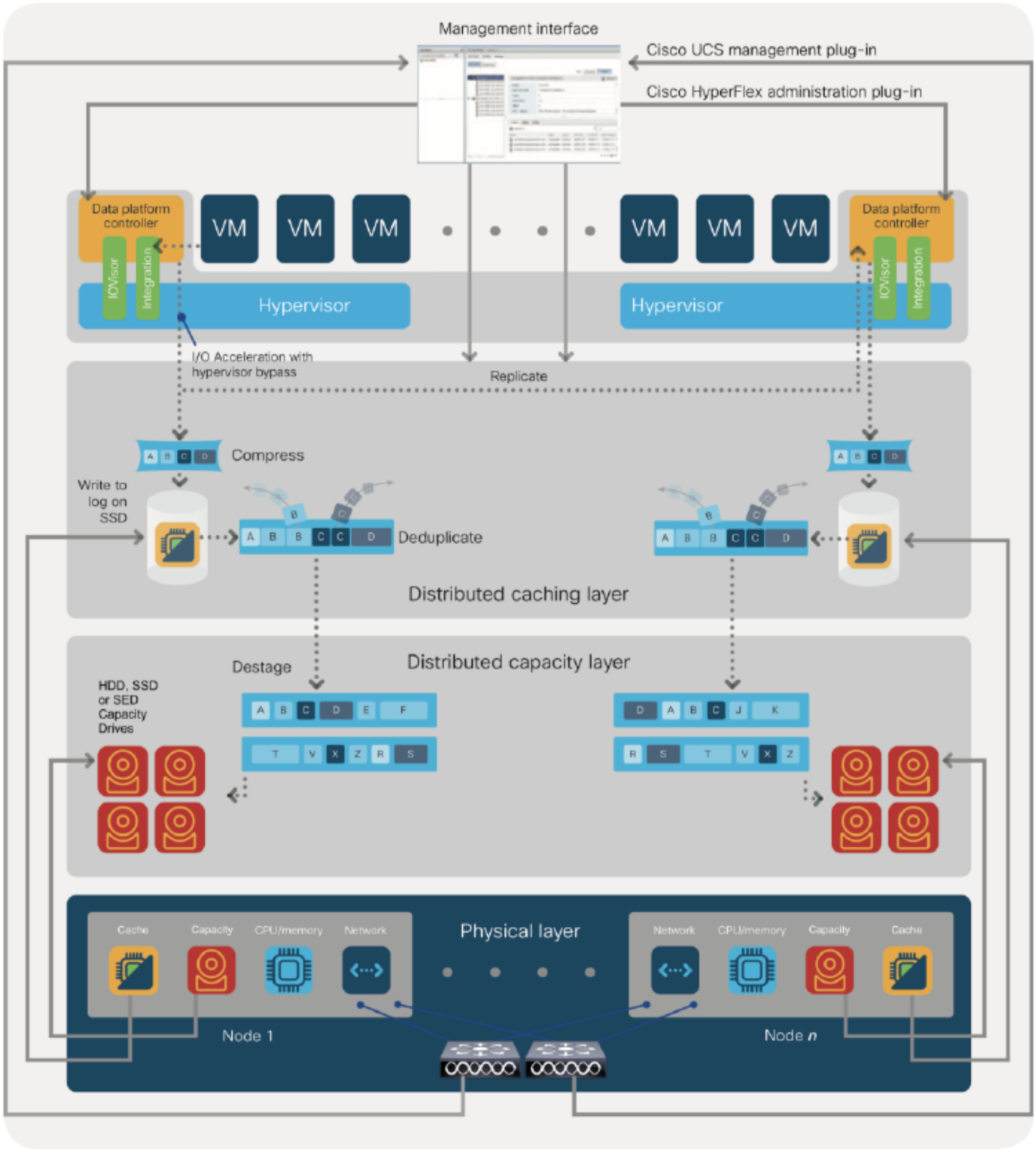
Incoming data is striped across the number of nodes required to satisfy availability requirements—usually two or three nodes. Based on policies you set, incoming write operations are acknowledged as persistent after they are replicated to the SSD drives in other nodes in the cluster. This approach reduces the likelihood of data loss due to SSD or node failures. The write operations are then de-staged to SSDs in the capacity layer in the All-Flash memory configuration for long-term storage.

The log-structured file system writes sequentially to one of two write logs (three in case of RF=3) until it is full. It then switches to the other write log while de-staging data from the first to the capacity tier. When existing data is (logically) overwritten, the log-structured approach simply appends a new block and updates the metadata. This layout benefits SSD configurations in which seek operations are not time consuming. It reduces the write amplification levels of SSDs and the total number of writes the flash media experiences due to incoming writes and random overwrite operations of the data.

When data is de-staged to the capacity tier in each node, the data is deduplicated and compressed. This process occurs after the write operation is acknowledged, so no performance penalty is incurred for these

operations. A small deduplication block size helps increase the deduplication rate. Compression further reduces the data footprint. Data is then moved to the capacity tier as write cache segments are released for reuse (Figure 4).

Figure 4 Data Write Operation Flow Through the Cisco HyperFlex HX Data Platform



Hot data sets–data that are frequently or recently read from the capacity tier–are cached in memory. All-Flash configurations, however, does not use an SSD read cache since there is no performance benefit of such a cache; the persistent data copy already resides on high-performance SSDs. In these configurations, a

read cache implemented with SSDs could become a bottleneck and prevent the system from using the aggregate bandwidth of the entire set of SSDs.

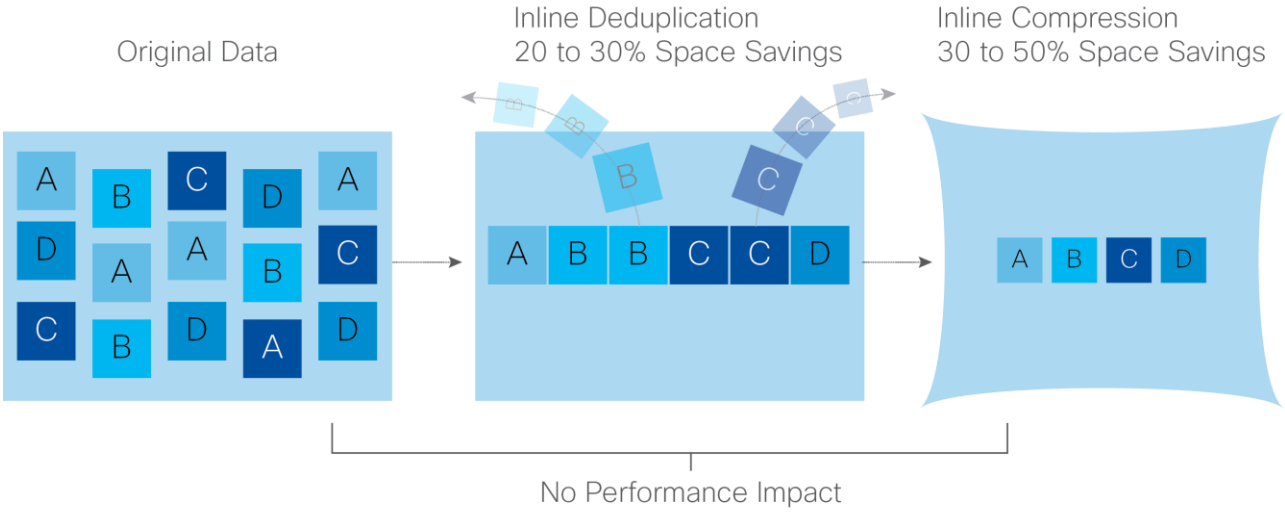
Data Optimization

The Cisco HyperFlex HX Data Platform provides finely detailed inline deduplication and variable block inline compression that is always on for objects in the cache (SSD and memory) and capacity (SSD or HDD) layers. Unlike other solutions, which require you to turn off these features to maintain performance, the deduplication and compression capabilities in the Cisco data platform are designed to sustain and enhance performance and significantly reduce physical storage capacity requirements.

Data Deduplication

Data deduplication is used on all storage in the cluster, including memory and SSD drives. Based on a patent-pending Top-K Majority algorithm, the platform uses conclusions from empirical research that show that most data, when sliced into small data blocks, has significant deduplication potential based on a minority of the data blocks. By fingerprinting and indexing just these frequently used blocks, high rates of deduplication can be achieved with only a small amount of memory, which is a high-value resource in cluster nodes (Figure 5).

Figure 5 Cisco HyperFlex HX Data Platform Optimizes Data Storage with No Performance Impact



Inline Compression

The Cisco HyperFlex HX Data Platform uses high-performance inline compression on data sets to save storage capacity. Although other products offer compression capabilities, many negatively affect performance. In contrast, the Cisco data platform uses CPU-offload instructions to reduce the performance impact of compression operations. In addition, the log-structured distributed-objects layer has no effect on modifications (write operations) to previously compressed data. Instead, incoming modifications are compressed and written to a new location, and the existing (old) data is marked for deletion, unless the data needs to be retained in a snapshot.

The data that is being modified does not need to be read prior to the write operation. This feature avoids typical read-modify-write penalties and significantly improves write performance.

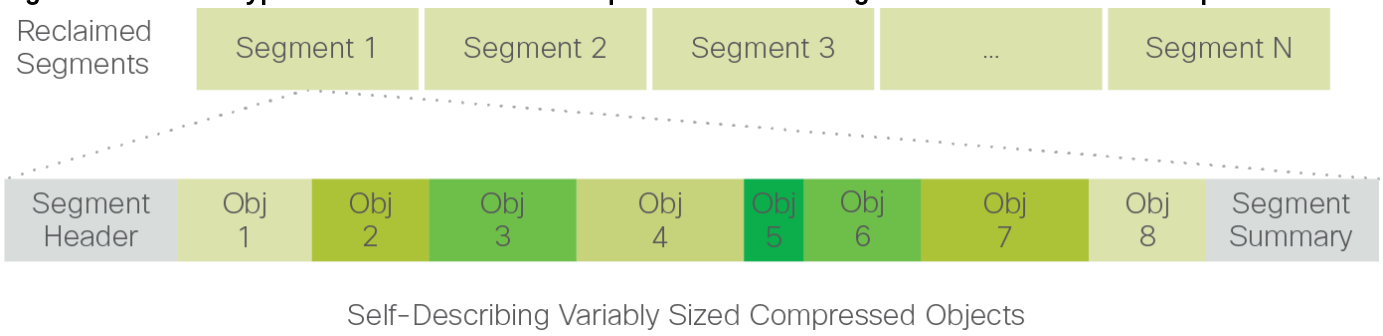
Log-Structured Distributed Objects

In the Cisco HyperFlex HX Data Platform, the log-structured distributed-object store layer groups and compresses data that filters through the deduplication engine into self-addressable objects. These objects are written to disk in a log-structured, sequential manner. All incoming I/O—including random I/O—is written sequentially to both the caching (SSD and memory) and persistent (SSD or HDD) tiers. The objects are distributed across all nodes in the cluster to make uniform use of storage capacity.

By using a sequential layout, the platform helps increase flash-memory endurance. Because read-modify-write operations are not used, there is little or no performance impact of compression, snapshot operations, and cloning on overall performance.

Data blocks are compressed into objects and sequentially laid out in fixed-size segments, which in turn are sequentially laid out in a log-structured manner (Figure 6). Each compressed object in the log-structured segment is uniquely addressable using a key, with each key fingerprinted and stored with a checksum to provide high levels of data integrity. In addition, the chronological writing of objects helps the platform quickly recover from media or node failures by rewriting only the data that came into the system after it was truncated due to a failure.

Figure 6 Cisco HyperFlex HX Data Platform Optimizes Data Storage with No Performance Impact



Encryption

Securely encrypted storage optionally encrypts both the caching and persistent layers of the data platform. Integrated with enterprise key management software, or with passphrase-protected keys, encrypting data at rest helps you comply with HIPAA, PCI-DSS, FISMA, and SOX regulations. The platform itself is hardened to Federal Information Processing Standard (FIPS) 140-1 and the encrypted drives with key management comply with the FIPS 140-2 standard.

Data Services

The Cisco HyperFlex HX Data Platform provides a scalable implementation of space-efficient data services, including thin provisioning, space reclamation, pointer-based snapshots, and clones—without affecting performance.

Thin Provisioning

The platform makes efficient use of storage by eliminating the need to forecast, purchase, and install disk capacity that may remain unused for a long time. Virtual data containers can present any amount of logical space to applications, whereas the amount of physical storage space that is needed is determined by the data that is written. You can expand storage on existing nodes and expand your cluster by adding more storage-intensive nodes as your business requirements dictate, eliminating the need to purchase large amounts of storage before you need it.

Snapshots

The Cisco HyperFlex HX Data Platform uses metadata-based, zero-copy snapshots to facilitate backup operations and remote replication: critical capabilities in enterprises that require always-on data availability. Space-efficient snapshots allow you to perform frequent online backups of data without needing to worry about the consumption of physical storage capacity. Data can be moved offline or restored from these snapshots instantaneously.

- Fast snapshot updates: When modified-data is contained in a snapshot, it is written to a new location, and the metadata is updated, without the need for read-modify-write operations.
- Rapid snapshot deletions: You can quickly delete snapshots. The platform simply deletes a small amount of metadata that is located on an SSD, rather than performing a long consolidation process as needed by solutions that use a delta-disk technique.
- Highly specific snapshots: With the Cisco HyperFlex HX Data Platform, you can take snapshots on an individual file basis. In virtual environments, these files map to drives in a virtual machine. This flexible specificity allows you to apply different snapshot policies on different virtual machines.

Many basic backup applications, read the entire dataset, or the changed blocks since the last backup at a rate that is usually as fast as the storage, or the operating system can handle. This can cause performance implications since HyperFlex is built on UCS with 10GbE which could result in multiple gigabytes per second of backup throughput. These basic backup applications, such as Windows Server Backup, should be scheduled during off-peak hours, particularly the initial backup if the application lacks some form of change block tracking.

Full featured backup applications, such as Veeam Backup and Replication v9.5, have the ability to limit the amount of throughput the backup application can consume which can protect latency sensitive applications during the production hours. With the release of v9.5 update 2, Veeam is the first partner to integrate HX native snapshots into the product. HX Native snapshots do not suffer the performance penalty of delta-disk snapshots, and do not require heavy disk IO impacting consolidation during snapshot deletion.

Particularly important for SQL administrators is the Veeam Explorer for SQL: https://www.veeam.com/microsoft-sql-server-explorer.html, which can provide transaction level recovery within the Microsoft VSS framework. The three ways Veeam Explorer for SQL Server works to restore SQL Server databases include; from the backup restore point, from a log replay to a point in time, and from a log replay to a specific transaction – all without taking the VM or SQL Server offline.

Fast, Space-Efficient Clones

In the Cisco HyperFlex HX Data Platform, clones are writable snapshots that can be used to rapidly provision items such as virtual desktops and applications for test and development environments. These fast, space-efficient clones rapidly replicate storage volumes so that virtual machines can be replicated through just metadata operations, with actual data copying performed only for write operations. With this approach, hundreds of clones can be created and deleted in minutes. Compared to full-copy methods, this approach can save a significant amount of time, increase IT agility, and improve IT productivity.

Clones are deduplicated when they are created. When clones start diverging from one another, data that is common between them is shared, with only unique data occupying new storage space. The deduplication engine eliminates data duplicates in the diverged clones to further reduce the clone's storage footprint.

Data Replication and Availability

In the Cisco HyperFlex HX Data Platform, the log-structured distributed-object layer replicates incoming data, improving data availability. Based on policies that you set, data that is written to the write cache is synchronously replicated to one or two other SSD drives located in different nodes before the write operation is acknowledged to the application. This approach allows incoming writes to be acknowledged quickly while protecting data from SSD or node failures. If an SSD or node fails, the replica is quickly re-created on other SSD drives or nodes using the available copies of the data.

The log-structured distributed-object layer also replicates data that is moved from the write cache to the capacity layer. This replicated data is likewise protected from SSD or node failures. With two replicas, or a total of three data copies, the cluster can survive uncorrelated failures of two SSD drives or two nodes without the risk of data loss. Uncorrelated failures are failures that occur on different physical nodes. Failures that occur on the same node affect the same copy of data and are treated as a single failure. For example, if one disk in a node fails and subsequently another disk on the same node fails, these correlated failures count as one failure in the system. In this case, the cluster could withstand another uncorrelated failure on a different node. See the Cisco HyperFlex HX Data Platform system administrator's guide for a complete list of fault-tolerant configurations and settings.

If a problem occurs in the Cisco HyperFlex HX controller software, data requests from the applications residing in that node are automatically routed to other controllers in the cluster. This same capability can be used to upgrade or perform maintenance on the controller software on a rolling basis without affecting the availability of the cluster or data. This self-healing capability is one of the reasons that the Cisco HyperFlex HX Data Platform is well suited for production applications.

In addition, native replication transfers consistent cluster data to local or remote clusters. With native replication, you can snapshot and store point-in-time copies of your environment in local or remote environments for backup and disaster recovery purposes.

HyperFlex VM Replication

HyperFlex Replication copies the virtual machine's snapshots from one Cisco HyperFlex cluster to another Cisco HyperFlex cluster to facilitate recovery of protected virtual machines from a cluster or site failure, via failover to the secondary site.

HyperFlex Stretched Clusters

Stretched Cluster allows nodes to be evenly split between two physical locations, keeping a duplicate copy of all the data in both locations, thereby providing protection in case of an entire site failure.

Data Rebalancing

A distributed file system requires a robust data rebalancing capability. In the Cisco HyperFlex HX Data Platform, no overhead is associated with metadata access, and rebalancing is extremely efficient. Rebalancing is a non-disruptive online process that occurs in both the caching and persistent layers, and data is moved at a fine level of specificity to improve the use of storage capacity. The platform automatically rebalances existing data when nodes and drives are added or removed or when they fail. When a new node is added to the cluster, its capacity and performance is made available to new and existing data. The rebalancing engine distributes existing data to the new node and helps ensure that all nodes in the cluster are used uniformly from capacity and performance perspectives. If a node fails or is removed from the cluster, the rebalancing engine rebuilds and distributes copies of the data from the failed or removed node to available nodes in the clusters.

Online Upgrades

Cisco HyperFlex HX-Series systems and the HX Data Platform support online upgrades so that you can expand and update your environment without business disruption. You can easily expand your physical resources; add processing capacity; and download and install BIOS, driver, hypervisor, firmware, and Cisco UCS Manager updates, enhancements, and bug fixes.

Why to use HyperFlex All-Flash systems for Database Deployments

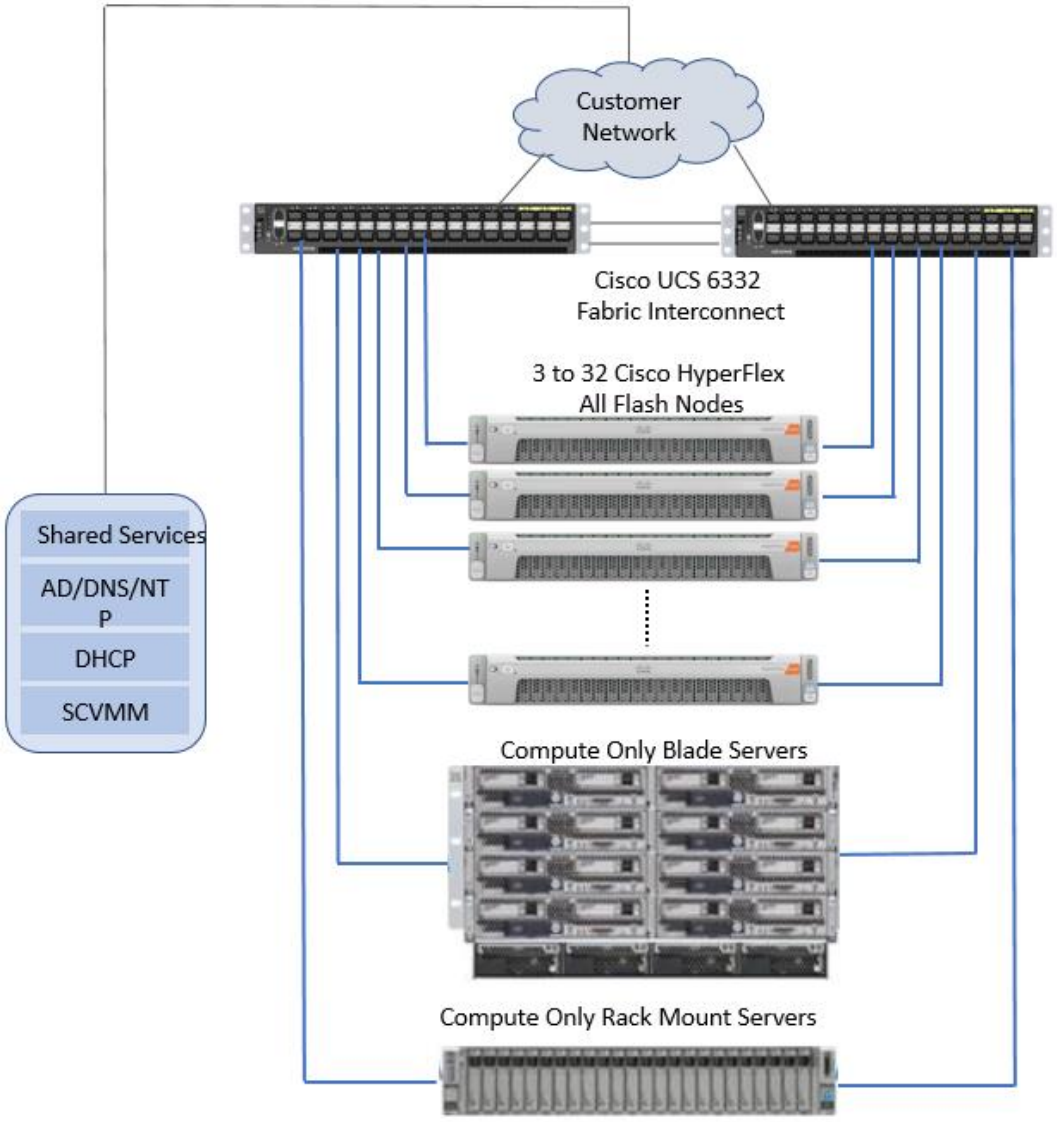
SQL server database systems act as the backend to many critical and performance hungry applications. It is very important to ensure that it delivers consistent performance with predictable latency throughout. Below are some of the major advantages of Cisco HyperFlex All-Flash hyperconverged systems which makes it ideally suited for SQL Server database implementations.

- Low latency with consistent performance: Cisco HyperFlex All-Flash nodes provides excellent platform for critical database deployment by offering low latency, consistent performance and exceeds most of the database service level agreements.
- Data protection (fast clones and snapshots, replication factor, VM replication and Stretched Cluster): The HyperFlex systems are engineered with robust data protection techniques that enable quick backup and recovery of the applications in case of any failures.
- Storage optimization: All the data that comes in the HyperFlex systems are by default optimized using inline deduplication and data compression techniques. Additionally, the HX Data Platform's log-structured file system ensures data blocks are written to flash devices in a sequential manner thereby increasing flash-memory endurance. HX System makes efficient use of flash storage by using Thin Provisioning storage optimization technique.
- Performance and Capacity Online Scalability: The flexible and independent scalability of the capacity and compute tiers of HyperFlex systems provide immense opportunities to adapt to the growing performance demands without any application disruption.
- No Performance Hotspots: The distributed architecture of HyperFlex Data Platform ensures that every VM is able to leverage the storage IOPS and capacity of the entire cluster, irrespective of the physical node it is residing on. This is especially important for SQL Server VMs as they frequently need higher performance to handle bursts of application or user activity.
- Non-disruptive System maintenance: Cisco HyperFlex Systems enables distributed computing and storage environment which enables the administrators to perform system maintenance tasks without disruption.

Solution Design

This section details the architectural components of Cisco HyperFlex, a hyperconverged system to host Microsoft SQL Server databases in a virtual environment. Figure 7 depicts a sample Cisco HyperFlex hyperconverged reference architecture comprising HX-Series rack mount servers.

Figure 7 Cisco HyperFlex Reference Architecture using All-Flash Nodes



Cisco HyperFlex is composed of a pair of Cisco UCS Fabric Interconnects along with up to thirty-two HX-Series rack mount servers per cluster. Up to thirty-two compute-only servers can also be added per HyperFlex cluster. Adding Cisco UCS rack mount servers and/or Cisco UCS 5108 Blade chassis, which house Cisco UCS blade servers allows for additional compute resources in an extended cluster design. Up to eight separate HX clusters can be installed under a single pair of Fabric Interconnects. The two Fabric

Interconnects connect to every HX-Series rack mount server, and connect to every Cisco UCS 5108 blade chassis, and Cisco UCS rack mount server. Upstream network connections, also referred as "north bound" network, are made from the Fabric Interconnects to the customer datacenter network at the time of installation. For more details on physical connectivity of HX-Series services, compute-only servers, Fabric Interconnect to the north bound network, please refer VSI CVD: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hyperflex_30_vsi_esxi.html#_Toc514225498

Infrastructure services such as Active Directory, DNS, NTP and VMWare vCenter are typically installed outside the HyperFlex cluster. Customers can leverage these existing services deploying and managing the HyperFlex cluster.

The HyperFlex storage solution has several data protection techniques, as explained in detail in the Technology overview section, one of which is data replication which needs to be configured on HyperFlex cluster creation. Based on the specific performance and data protection requirements, customer can choose either a replication factor of two (RF2) or three (RF3). For the solution validation (described in the "Solution Testing and Validation" later in this document), we had configured the test HyperFlex cluster to be of replication factor 3 (RF3).

As described in the earlier Technology Overview section, Cisco HyperFlex distributed file system software runs inside a controller VM, which gets installed on each cluster node. These controller VMs pool and manage all the storage devices and exposes the underlying storage as NFS mount points to the VMware ESXi hypervisors. The ESXi hypervisors exposes these NFS mount points as datastores to the guest virtual machines to store their data.

For this document, validation is done only on HXAF240c-M5SX All-Flash converged nodes, which act as both compute and storage nodes.

Logical Network design

In the Cisco HyperFlex All-Flash system, Cisco VIC 1387 is used to provide the required logical network interfaces on each host in the cluster. The communication pathways in Cisco HyperFlex system can be categorized in to four different traffic zones as described below.

Management Zone: This zone comprises the connections needed to manage the physical hardware, the hypervisor hosts, and the storage platform controller virtual machines (SCVM). These interfaces and IP addresses need to be available to all staff who will administer the HX system, throughout the LAN/WAN. This zone must provide access to Domain Name System (DNS) and Network Time Protocol (NTP) services and allow Secure Shell (SSH) communication. In this zone are multiple physical and virtual components:

- Fabric Interconnect management ports.
- Cisco UCS external management interfaces used by the servers, which answer via the FI management ports.
- ESXi host management interfaces.
- Storage Controller VM management interfaces.
- A roaming HX cluster management interface.

- Storage Controller VM Management interfaces.

VM Zone: This zone comprises the connections needed to service network IO to the guest VMs that will run inside the HyperFlex hyperconverged system. This zone typically contains multiple VLANs that are trunked to the Cisco UCS Fabric Interconnects via the network uplinks and tagged with 802.1Q VLAN IDs. These interfaces and IP addresses need to be available to all staff and other computer endpoints which need to communicate with the guest VMs in the HX system, throughout the LAN/WAN.

Storage Zone: This zone comprises the connections used by the Cisco HX Data Platform software, ESXi hosts, and the storage controller VMs to service the HX Distributed Data Filesystem. These interfaces and IP addresses need to be able to communicate with each other at all times for proper operation. During normal operation, this traffic all occurs within the Cisco UCS domain, however there are hardware failure scenarios where this traffic would need to traverse the network northbound of the Cisco UCS domain. For that reason, the VLAN used for HX storage traffic must be able to traverse the network uplinks from the Cisco UCS domain, reaching FI A from FI B, and vice-versa. This zone is primarily jumbo frame traffic therefore jumbo frames must be enabled on the Cisco UCS uplinks. In this zone are multiple components:

- A teamed interface is used for storage traffic on each Hyper-V host in the HX cluster.
- Storage Controller VM storage interfaces.
- A roaming HX cluster storage interface.

vMotion Zone: This zone comprises the connections used by the ESXi hosts to enable live migration of the guest VMs from host to host. During normal operation, this traffic all occurs within the Cisco UCS domain, however there are hardware failure scenarios where this traffic would need to traverse the network northbound of the Cisco UCS domain. For that reason, the VLAN used for HX live migration traffic must be able to traverse the network uplinks from the Cisco UCS domain, reaching FI A from FI B, and vice-versa.

By leveraging Cisco UCS vNIC templates, LAN connectivity policies and vNIC placement policies in service profile, eight vNICs are carved out from Cisco VIC 1387 on each HX-Series server for network traffic zones mentioned above. Every HX-Series server will detect the network interfaces in the same order and they will always be connected to the same VLANs via the same network fabrics. The following table outlines the vNICs and other configuration details being used in the solution.

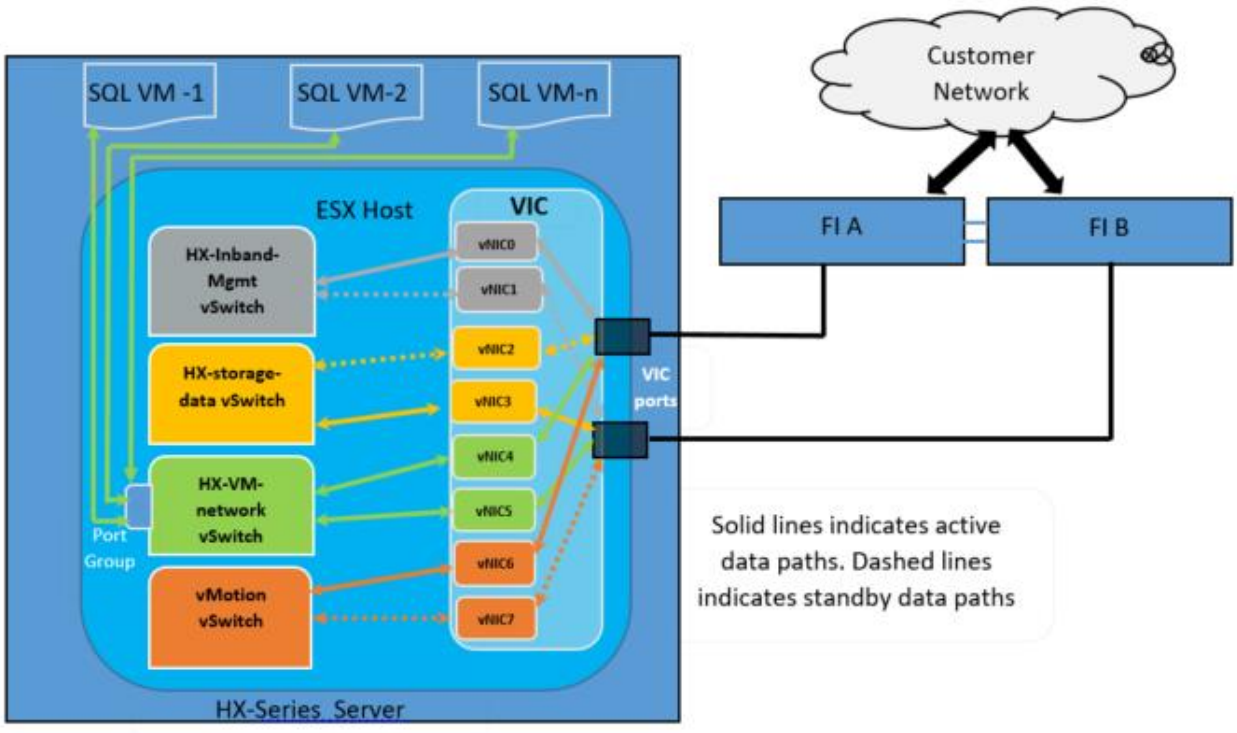
Table 1 Virtual Interface Order with in HX-Series server

vNIC Template Name:	hv-mgmt-a	hv-mgmt-b	storage-data-a	storage-data-b	vm-network-a	vm-network-b	hv-vmotion-a	hv-vmotion-b
Setting	Value	Value	Value	Value	Value	Value	Value	Value
Fabric ID	A	B	A	B	A	B	A	B
Fabric Failover	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled	Disabled
Target	Adapter	Adapter	Adapter	Adapter	Adapter	Adapter	Adapter	Adapter

Type	Updating Template	Updating Template	Updating Template	Updating Template	Updating Template	Updating Template	Updating Template	Updating Template
MTU	1500	1500	9000	9000	1500	1500	9000	9000
MAC Pool	hv-mgmt-a	hv-mgmt-b	storage-data-a	storage-data-b	vm-network-a	hv-network-b	hv-vmotion-a	hv-vmotion-b
QoS Policy	silver	silver	platinum	platinum	gold	gold	bronze	bronze
Network Control Policy	HyperFlex-infra	HyperFlex-infra	HyperFlex-infra	HyperFlex-infra	HyperFlex-vm	HyperFlex-vm	HyperFlex-infra	HyperFlex-infra
VLANs	<<hx-inband-mgmt>>	<<hx-inband-mgmt>>	<<hx-storage-data>>	<<hx-storage-data>>	<<hx-network>>	<<hx-network>>	<<vm-vmotion>>	<<vm-vmotion>>
Native VLAN	No	No	No	No	No	No	No	No

The following figure illustrates logical network design of a HX-Series server of HyperFlex cluster.

Figure 8 HX-Series Server Logical Network Diagram



As shown in the figure above, four virtual standard switches are configured for four traffic zones. Each virtual switch is configured with two vNICs and are connected to both the Fabric Interconnects. The vNICs are configured in active and standby fashion for Storage, Management and vMotion networks. However, for VM network virtual switch vNICs are configured in active and active fashion. This ensures that the data path for guest VMs traffic has aggregated bandwidth for the specific traffic type.

Jumbo frames are enabled for:

- Storage traffic: Enabling jumbo frames on the Storage traffic zone would benefit in the following SQL server database use case scenarios:
  - Heavy write SQL server guest VMs caused by the activities such as database restoring, rebuilding indexes, importing data etc.
  - Heavy read SQL server guest VMs caused by the typical maintenance activities such as backup database, export data, report queries, rebuilding indexes etc.
- vMotion traffic: Enabling jumbo frames on vMotion traffic zone help the system quickly failover the SQL VMs to other hosts; there by, reducing the overall database downtime.

Creating a separate logical network (using two dedicated vNICs) for guest VMs is beneficial with the following advantages:

- Isolating guest VM traffic from other traffic such as management, HX replication etc.
- A dedicated MAC pool can be assigned to each vNIC, which would simplify troubleshooting the connectivity issues.
- As shown in figure 8, the VM Network switch is configured with two vNICs in active and active fashion to provide two active data paths which will result in aggregated bandwidth.

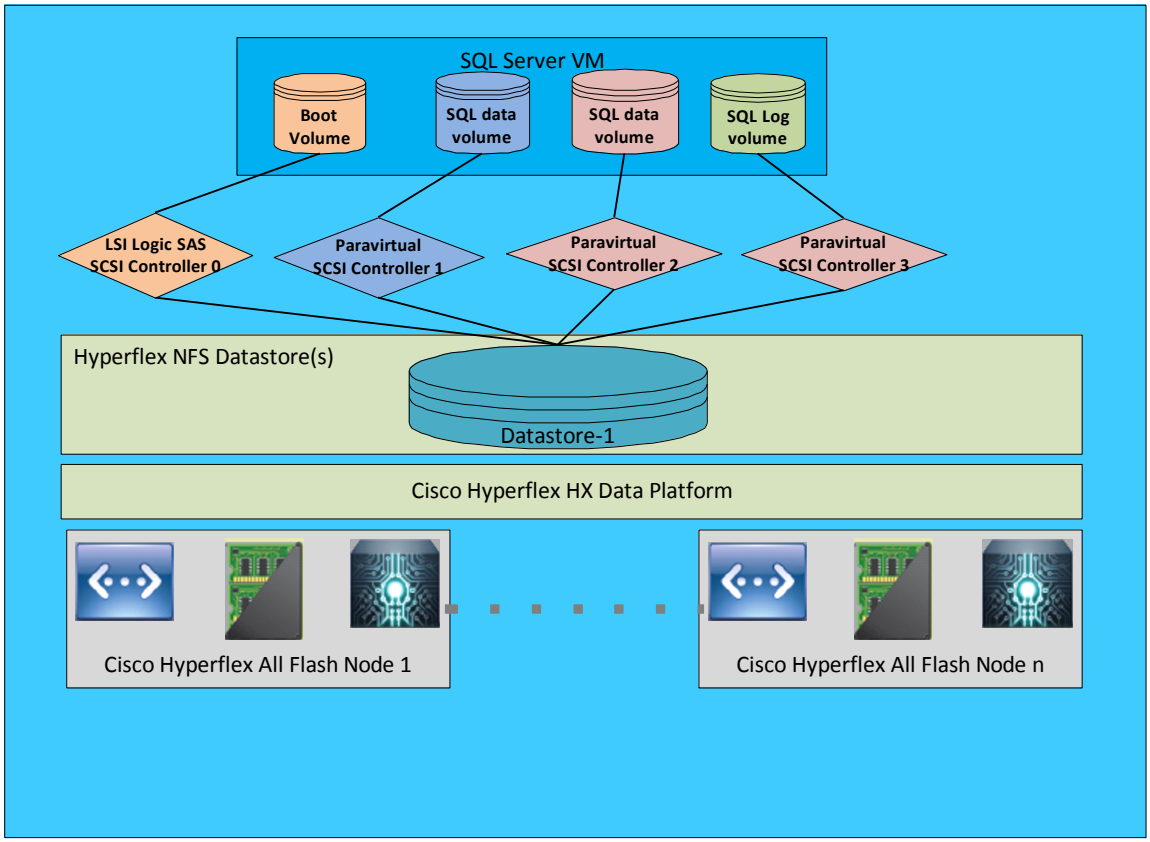
For more details on network configuration of HyperFlex HX-Server node, using Cisco UCS network policies, templates and service profiles, refer to the HyperFlex Network Design guidelines section of this of VSI CVD document.
https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/hyperflex_30_vsi_esxi.html#_Toc514225513

The following sections provide more details on configuration and deployment best practices for deploying SQL server databases on HyperFlex All-Flash nodes.

Storage Configuration for SQL Guest VMs

Figure 9 outlines the storage configuration recommendations for virtual machines running SQL server databases on HyperFlex All-Flash nodes. Single LSI Logic virtual SCSI controller is used to host the Guest OS. Separate Paravirtual SCSI (PVSCSI) controllers are configured to host SQL server data and log files. For large scale and high performing SQL deployments, it is recommended to spread the SQL data files across two or more different PVSCSI controllers for better performance as shown in the following figure, Additional performance guidelines are given in “Deployment Planning” section.

Figure 9 Storage Design for Microsoft SQL Server Database Deployment



Deployment Planning

It is crucial to follow and implement configuration best practices and recommendations in order to achieve best performance from any underlying system. This section details the major design and configuration best practices that should be followed when deploying SQL server databases on All-Flash HyperFlex systems.

Datastore recommendation

The following recommendations can be followed while deploying SQL server virtual machines on HyperFlex All-Flash Systems.

All the virtual machine's virtual disks comprising guest Operating System, SQL data, and transaction log files can be placed on a single datastore exposed as NFS file share to the ESXi hosts. Deploying multiple SQL virtual machines using single datastore simplifies the management tasks.

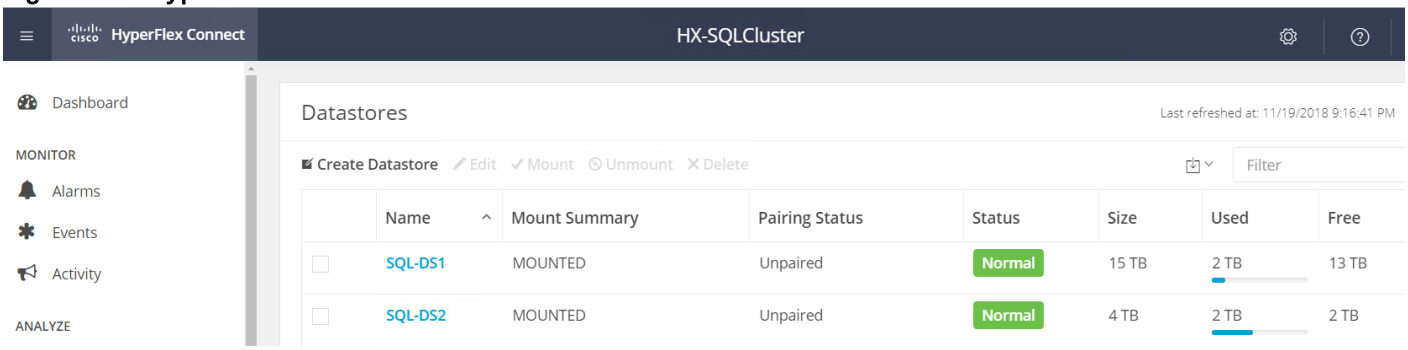
There is a maximum queue depth limit of 1024 for each NFS datastore per host, which is an optimum queue depth for most of the workloads. However, when consolidated IO requests from all the virtual machines deployed on the datastore exceeds 1024 (per host limit), then virtual machines might experience higher IO latencies. Symptoms of higher latencies can be identified using ESXTOP results.

In such cases, creating new datastore and deploying SQL virtual machines on the new datastore will help. The general recommendation is to deploy low IO demanding SQL virtual machines in one single datastore

until high guest latencies are noticed. Also, deploying a dedicated datastore for High IO demanding SQL VMs will allow dedicated queue and hence lesser latencies can be observed.

The following figure shows that two different datastores are used for deploying various SQL guest virtual machines. “SQL-DS1” is used to deploy multiple small to medium SQL virtual machines while “SQL-DS2” is dedicatedly used for deploying single large SQL virtual machine with high IO demanding performance requirements.

Figure 10 HyperFlex Datastores



SQL Virtual Machine configuration recommendation

While creating a VM for deploying SQL Server instance on a HyperFlex All-Flash system, the following recommendations should be followed for performance and better administration.

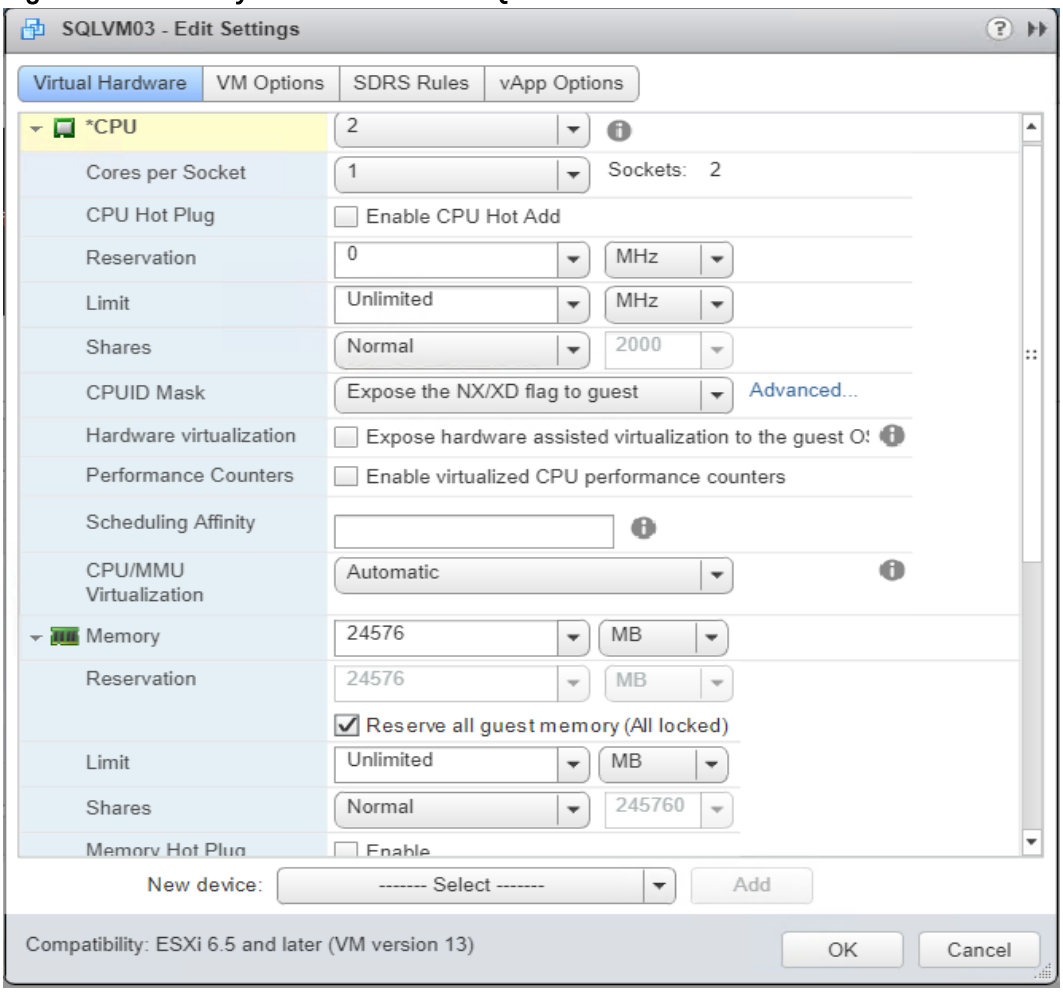
Cores per Socket

By default, when creating Virtual Machine, vSphere will create as many virtual sockets as you have requested vCPUs and the cores per socket is set to one. This will enable vNUMA to select and present the best virtual NUMA topology to the guest operating system, which will be optimal on the underlying physical topology. Typically, it is not recommended to change this setting, unless the changes made to this setting is thoroughly tested on the given environment.

Memory Reservation

SQL server database transactions are usually CPU and memory intensive. In a heavy OLTP database systems, it is recommended to reserve all the memory assigned to the SQL virtual machines. This ensures that the assigned memory to the SQL VM is committed and will eliminate the possibility of ballooning and swapping the memory out by the ESXi hypervisor. Memory reservations will have little overhead on the ESXi system. For more information about memory overhead, see Understanding Memory Overhead at: https://pubs.vmware.com/vsphere-51/index.jsp?topic=%2Fcom.vmware.vsphere.resmgmt.doc%2FGUID-4954A03F-E1F4-46C7-A3E7-947D30269E34.html

Figure 11 Memory Reservations for SQL Virtual Machine



Paravirtual SCSI adapters for large-scale high IO virtual machines

For virtual machines with high disk IO requirements, it is recommended to use Paravirtual SCSI (PVSCSI) adapters. PVSCSI controller is a virtualization aware, high-performance SCSI adapter that allows the lowest possible latency and highest throughput with the lowest CPU overhead. It also has higher queue depth limits compared to other legacy controllers. Legacy controllers (LSI Logic SAS, LSI Logic Parallel etc.) can cause bottleneck and impact database performance; hence not recommended for IO intensive database applications such as SQL server databases.

Queue Depth and SCSI controller recommendations

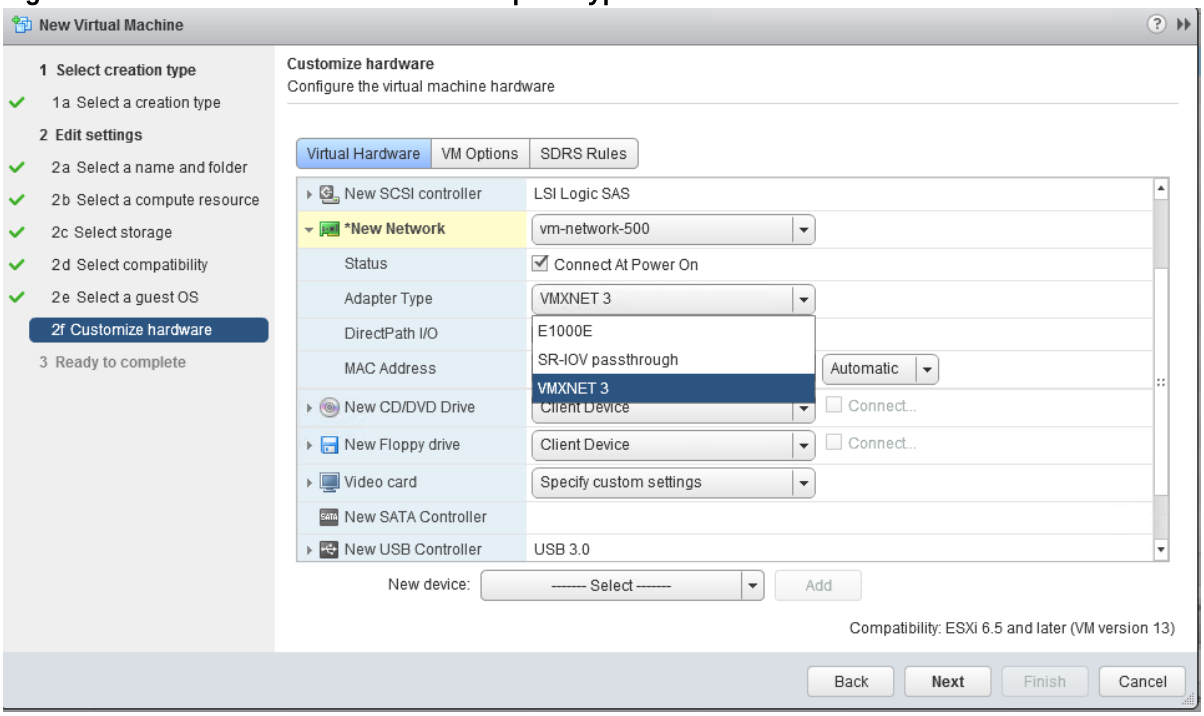
Many times, queue depth settings of virtual disks are overlooked, which can impact performance particularly in high IO workloads. Systems such as Microsoft SQL Server databases tend to issue a lot of simultaneous IOs resulting in an insufficient VM driver queue depth settings (default setting is 64 for PVSCSI) to sustain the heavy IOs. It is recommended to change the default queue depth setting to a higher value (up to 254) as suggested in this VMware KB article.
https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2053145

For large-scale and high IO databases, it is always recommended to use multiple virtual disks and have those virtual disks distributed across multiple SCSI controller adapters rather than assigning all of them to a single SCSI controller. This ensures that the guest VM will access multiple virtual SCSI controllers (four SCSI controllers maximum per guest VM), which in turn results in greater concurrency by utilizing the multiple queues available for the SCSI controllers.

Virtual Machine Network Adapter type

It is highly recommended to configure virtual machine network adapters with “VMXNET 3”. VMXNET 3 is the latest generation of Paravirtualized NICs designed for performance. It offers several advanced features including multi-queue support, receive side scaling, IPv4/IPv6 offloads, and MSI/MSI-X interrupt delivery. While creating a new virtual machine, choose “VMXNET 3” as the adapter type as shown below.

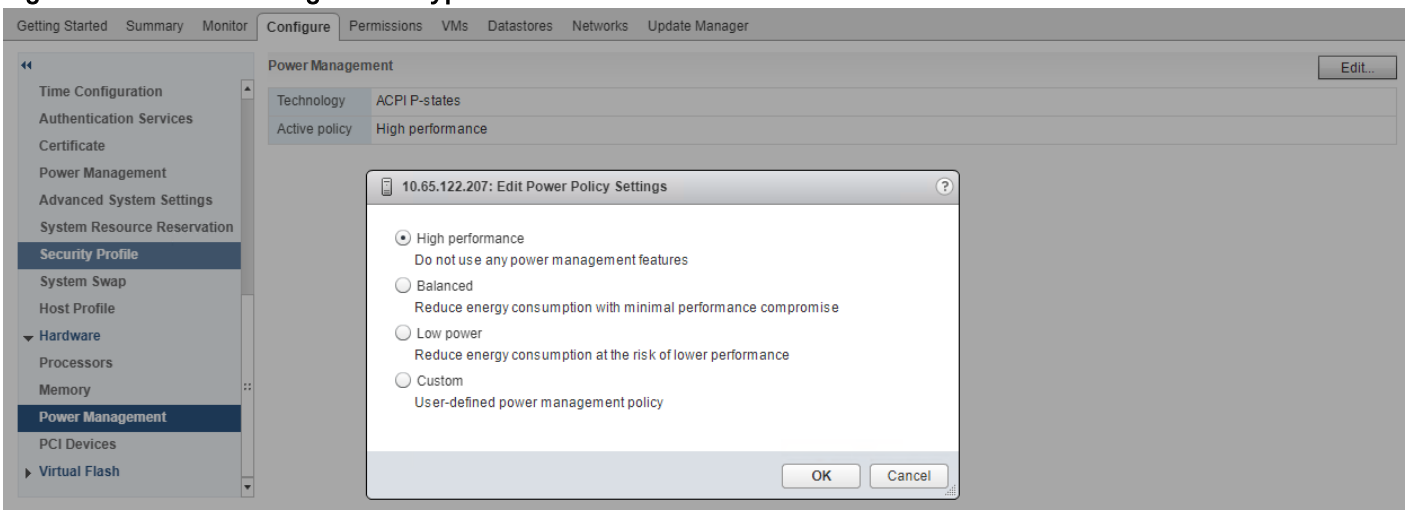
Figure 12 Virtual Machine Network Adapter Type



Guest Power Scheme Settings

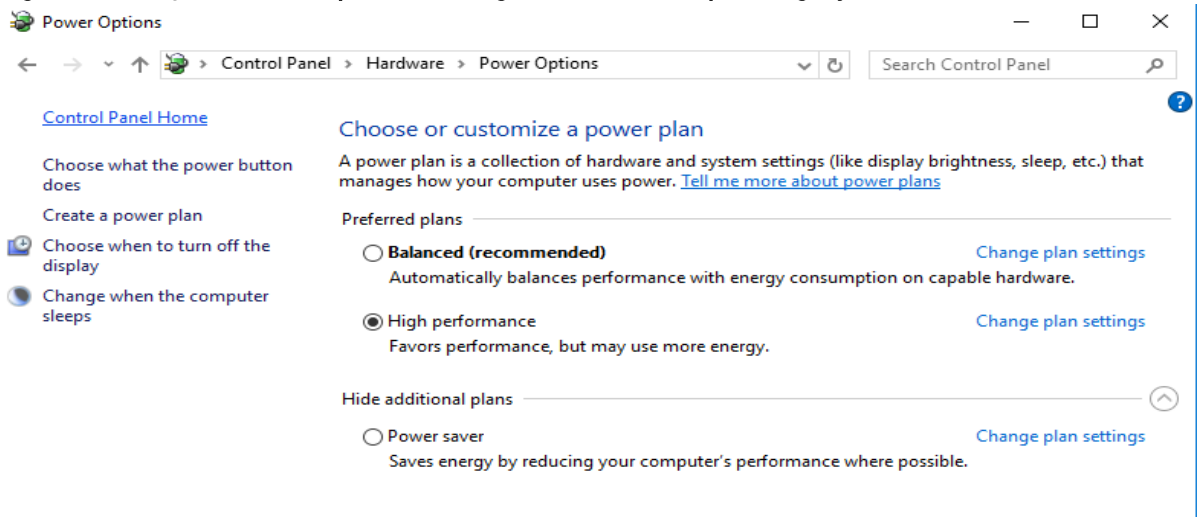
HX Servers are optimally configured, at factory installation time, with appropriate BIOS policy settings at the host level and hence does not require any changes. Similarly, ESXi power management option (at vCenter level) is set to “High performance” at the time of HX installation by installer as shown in the below figure.

Figure 13 Power setting on HX hypervisor node



Inside SQL server guest, it is recommended to set the power management option to “High Performance” for optimal database performance as shown in figure 14.

Figure 14 SQL Guest VM power settings in Windows Operating System



For other regular SQL server specific configuration recommendations on virtualized environments, see SQL Server best practices guide on VMware vSphere.
https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/solutions/sql-server-on-vmware-best-practices-guide.pdf

Achieving Database High Availability

Cisco HyperFlex storage systems incorporates efficient storage level availability techniques such as data mirroring (Replication Factor 2/3), native snapshot etc., to make sure continuous data access to the guest VMs hosted on the cluster. More details of the HX Data Platform Cluster Tolerated Failures are detailed in: https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/AdminGuide/3_5/b_HyperFlexSystems_AdministrationGuide_3_5/b_HyperFlexSystems_AdministrationGuide_3_5_chapter_00.html#id_13113.

This section discusses the high availability techniques that will be helpful in enhancing the availability of virtualized SQL server databases (apart from the storage level availability, which comes with HyperFlex solutions).

The availability of the individual SQL Server database instance and virtual machines can be enhanced using the technologies listed below.

- VMware HA: to achieve virtual machine availability
- Microsoft SQL Server AlwaysOn: To achieve database level high availability

Single VM / SQL Instance level high availability using VMware vSphere HA feature

Cisco HyperFlex solution leverages VMware clustering to provide availability to the hosted virtual machines. Since the exposed NFS storage is mounted on all the hosts in the cluster, they act as a shared storage environment to help migrate VMs between the hosts. This configuration helps migrate the VMs seamlessly in case of planned as well as unplanned outage. The vMotion vNIC need to be configured with Jumbo frames for faster guest VM migration.

You can find more information in the VMware document: https://docs.vmware.com/en/VMware-vSphere/6.5/vsphere-esxi-vcenter-server-65-availability-guide.pdf

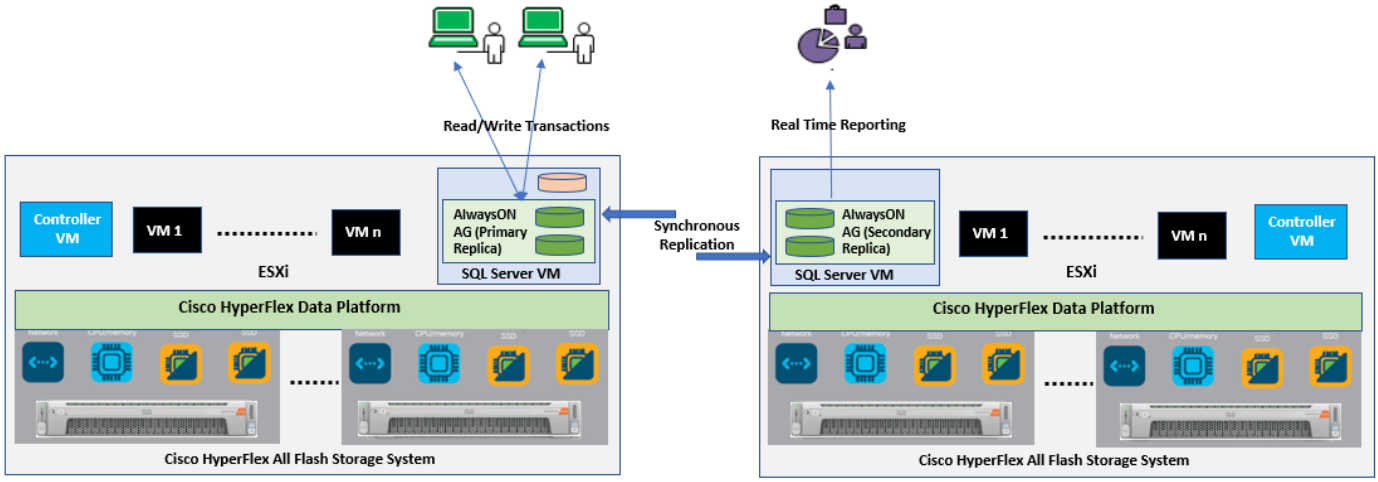
Database level high availability using SQL AlwaysOn Availability Group feature

HyperFlex architecture inherently uses NFS datastores. Microsoft SQL Server Failover Cluster Instance (FCI) needs shared storage which cannot be on NFS storage (unsupported by VMware ESXi). Hence FCI is not supported as high availability option, instead SQL Server AlwaysOn Availability Group feature can be used. Introduced in Microsoft SQL Server 2012, AlwaysOn Availability Groups maximizes the availability of a set of user databases for an enterprise. An availability group supports a failover environment for a discrete set of user databases, known as availability databases, that failover together. An availability group supports a set of read-write primary databases and one to eight sets of corresponding secondary databases. Optionally, secondary databases can be made available for read-only access and/or some backup operations. More information on this feature can be found at the Microsoft MSDN at: https://msdn.microsoft.com/en-us/library/hh510230.aspx.

Microsoft SQL Server AlwaysOn Availability Groups take advantage of Windows Server Failover Clustering (WSFC) as a platform technology. WSFC uses a quorum-based approach to monitor the overall cluster health and maximize node-level fault tolerance. The AlwaysOn Availability Groups will get configured as WSFC cluster resources and the availability of the same will depend on the underlying WSFC quorum modes and voting configuration explained at: https://docs.microsoft.com/en-us/sql/sql-server/failover-clusters/windows/wsfc-quorum-modes-and-voting-configuration-sql-server.

Using AlwaysOn Availability Groups with synchronous replication, supporting automatic failover capabilities, enterprises will be able to achieve seamless database availability across the database replicas configured. The following figure depicts the scenario where an AlwaysOn availability group is configured between the SQL server instances running on two separate HyperFlex Storage systems. To ensure that the involved databases provide guaranteed high performance and no data loss in the event of failure, proper planning need to be done to maintain a low latency replication network link between the clusters.

Figure 15 Synchronous AlwaysOn Configuration Across HyperFlex All-Flash Systems



Although there are no hard rules on the infrastructure used for hosting a secondary replica, following are some of the guidelines to be followed if planned to have a primary replica on the All-Flash High Performing cluster:

- In case of synchronous replication (no data loss)
  - The replicas need to be hosted on similar hardware configurations to ensure that the database performance is not compromised in waiting for the acknowledgment from the replicas.
  - A high speed low latency network connection between the replicas needs to be ensured.
- In case of asynchronous replication (may have data loss)
  - The performance of the primary replica does not depend on the secondary replica, so it can be hosted on low cost hardware solutions as well.
  - The amount to data loss will depend on the network characteristics and the performance of the replicas.

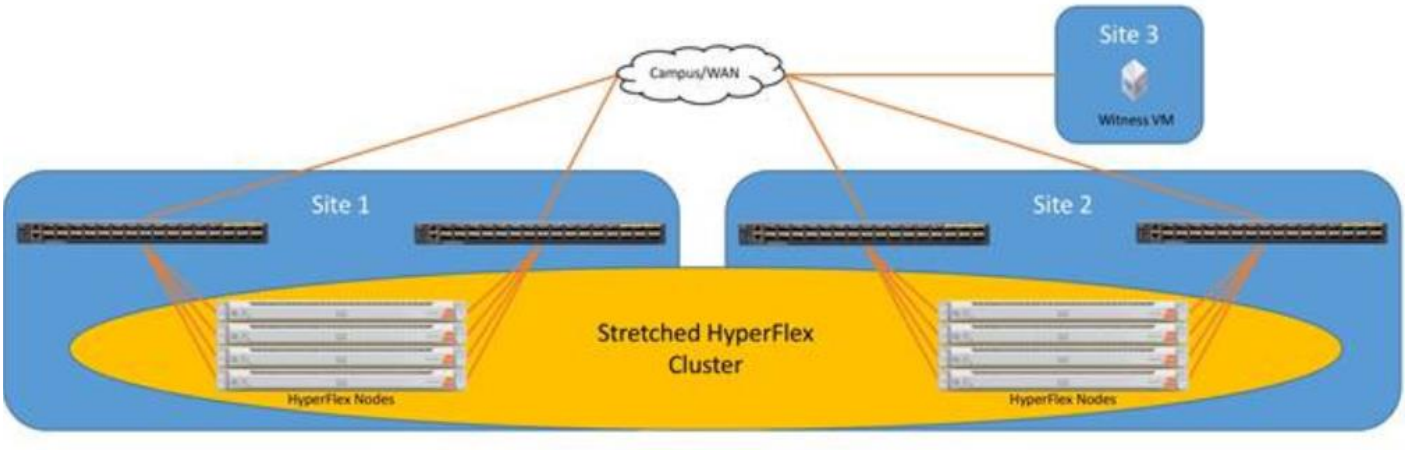
For customers who are willing to deploy AlwaysOn Availability Group within a single HyperFlex All-Flash cluster or AlwaysOn deployments, which involve more than 2 replicas, VMWare DRS anti-affinity rules must be used to ensure that each SQL VM replica is being placed on different VMware ESXi hosts in order to reduce database downtime. For more details on configuring VMware anti-affinity rules, refer the http://pubs.vmware.com/vsphere-60/index.jsp?topic=%2Fcom.vmware.vsphere.resmgmt.doc%2FGUID-7297C302-378F-4AF2-9BD6-6EDB1E0A850A.html.

Achieving Disaster Recovery for Databases using HyperFlex Stretched Cluster

HyperFlex Stretched Cluster provides customers with the ability to deploy a single HyperFlex Cluster that spread across two different datacenters. It is designed to offer business continuance in the event of significant disaster at a datacenter location. HyperFlex Stretched Cluster deploys half of the cluster resources in one physical datacenter location and the other half on a different distant datacenter. The data written to the HyperFlex Stretched Cluster is stored concurrently in both physical locations, therefore this system design provides additional data protection and availability because the data in the cluster remains available and online, even if an entire site experiences a failure. Since all data written to the cluster is written

simultaneously to the two sites, this design can be considered an active/active disaster recovery design. The recovery point objective (RPO, or the maximum amount of data that can potentially be lost prior to a failure) is essentially zero, due to all data being immediately and continuously available in both sites. The recovery time objective (RTO, or amount of time needed to return services to their functional state after a failure) is also near zero from an architectural standpoint, since all data and services are immediately available in the site, which remains online; however, actual recovery involves restarting the guest virtual machines and that does take some time. The below figure illustrates the HyperFlex Stretched Cluster.

Figure 16 HyperFlex Stretched Cluster



The cluster also requires a “tie breaker” or “witness” component, which should reside in a third, separate location. The witness runs in a virtual machine and act as a tie-breaker in the case of “split-brain” scenario. The HyperFlex Stretched Cluster is typically deployed in environments where the distance between the two clusters is limited, such as metropolitan or campus environments, and have low-latency and high bandwidth network available between the two locations in order to achieve better performance and quick recovery of the applications in event of failures.

For more details on the infrastructure requirements and limitations of the HyperFlex Stretched Cluster, see: https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_Stretched_Cluster/3_5/b_HyperFlex_Systems_Stretched_Cluster_Guide_3_5.pdf.

For complete details on installing HyperFlex Stretch Cluster, see: https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_Stretched_Cluster/3_5/b_HyperFlex_Systems_Stretched_Cluster_Guide_3_5/b_HyperFlex_Systems_Stretched_Cluster_Guide_3_5_chapter_011.html

Mission critical database deployments with zero data loss requirements can be seamlessly deployed on HyperFlex Stretched Cluster. It provides complete data protection for critical databases by replicating the two copies of the data in the local datacenter and two copies in the remote datacenter. Hence total of four data copies are maintained across the Stretched Cluster. Since the data copies are maintained at two different geographically separated datacenters, this design can tolerate complete datacenter failure with zero data loss, which is one of the most important requirements for critical database deployments.

Below list provides some of advantages that HyperFlex Stretched Cluster offers to the critical database deployments.

- High performance hyperconverged distributed cluster with Zero RPO and near zero RTO for running critical database applications.
- HyperFlex Stretched Cluster is Active-Active datacenter solution where each datacenter can be used as primary datacenter to host applications and at the same each of these datacenters are acting as backup for each other.
- Failures in the local site such as disk and node failures are limited to that datacenter itself. Meaning, the lost data can be regenerated using the local data copies only. Remote data copies are not used for regenerating the data. Hence it prevents data movement from site to site while healing from local host/datacenter failures.
- Since Stretched Cluster operates at infrastructure level, all the virtual machines deployed in the HyperFlex datastores are by default protected. It does not have any dependency on application edition like Enterprise, Standard etc. Thus, customers can deploy standard or any lower edition and can still achieve complete data protection to their databases. This option is financially optimal for customers who otherwise would have to procure enterprise application licenses in order to achieve same level of data protection that HyperFlex Stretched Cluster can offer.

Choosing between HyperFlex Stretched Cluster and SQL Server AlwaysOn Availability Groups for Data Protection

Microsoft SQL Server AlwaysOn Availability Group feature provides high availability and disaster recovery capabilities to the databases. In terms of data protection offering, there is an overlap between HyperFlex Stretched Cluster and SQL server AlwaysOn Availability Groups. Hence based on the business needs, appropriate solution should be chosen. The following table provide a few guidelines that would help customers to choose appropriate solution for protecting their SQL Server databases.

Table 2 HyperFlex Stretched Cluster and Microsoft SQL Server AlwaysOn Availability Groups

	HyperFlex Stretched Cluster	SQL Server AlwaysOn With Synchronous Replication (typically configured with in same datacenter)	SQL Server AlwaysOn With Asynchronous Replication
Protection against failures	Provides data protection against local hardware component failures as well complete datacenter failures. Example: disk failure of a host/ complete host failure in a site/ complete datacenter failure are covered Apart from HyperFlex Stretched Cluster software related failures, it does not protect other service failures (Applications and Operating System) running	Provides data protection from all types of local failures only; complete datacenter failures are not covered Example: disk failure of a host/ complete host failure of site/SQL or Windows failures are covered Application and Operating System Software failures are also protected.	Provides data protections against component failures as well as complete datacenter failures with some data loss.

	HyperFlex Stretched Cluster	SQL Server AlwaysOn With Synchronous Replication (typically configured with in same datacenter)	SQL Server AlwaysOn With Asynchronous Replication
	inside the virtual machine.		
Data protection	Full data protection (RPO=0) against above mentioned failures.	Full data protection (RPO=0) again above-mentioned failures.	May have data loss depending on network latency between the datacenters.
High Availability/Automatic failover	Automatic failover is supported. Recovery speed (RTO) depends on type of failure and how soon VMware HA failovers and restarts the VMs from failed node/datacenter to survival node/datacenter. Typical range is 30 seconds to 2 minutes.	Automatic failover is supported. Recovery speed (RTO) depends on type of failures and how soon SQL server exchanges the roles between primary and secondary replicas. Typical range is 5 to 30 seconds as exchanges of roles is much quicker.	Manual failover.
How many SQL VM/Databases/instances are protected	All the virtual machines deployed on HyperFlex Stretched Cluster are by default protected.	Only the SQL Databases/Instance/VMs that are part of AlwaysOn Availability Groups are protected.	Only the SQL Databases/Instance/VMs that are part of AlwaysOn Availability Groups are protected.
Suited for	Typically suited for metropolitan or campus environments where the two datacenters are separated with limited distance with low network latency between the datacenters.	Typically implemented within the same datacenter on two different racks.	Suited for long distant two different datacenters
Edition support	Supported in HyperFlex Enterprise Edition. Does not impose any constraints on SQL Server editions. It works fine with all the SQL editions.	Supported only in SQL server Enterprise Edition. Basic Availability Group (BAG) is also supported in SQL Server Standard Edition but with just one Read-Only replica and so many other limits.	SQL server Enterprise Edition only.

Given there would be multiple duplicate data copies, it is not required to implement both HyperFlex Stretched Cluster feature as well the Microsoft SQL Server AlwaysOn Availability Group for achieving database protection.

For database performance analysis on the stretched cluster, see “Solution Resiliency Testing and validation” section.

Microsoft SQL Server Deployment

Cisco HyperFlex 3.5 All-Flash System Installation and Deployment

This CVD focuses on Microsoft SQL Server virtual machine deployment and assumes the availability of an already running healthy All-Flash HyperFlex 3.5 cluster. For more information on deployment of Cisco HyperFlex 3.5 All-Flash System, see installation guide at: https://www.cisco.com/c/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/Installation_VMWare_ESXi/3_5/b_HyperFlexSystems_Installation_Guide_for_VMware_ESXi_3_5.html

Deployment Procedure

This section provides step by step deployment procedure of setting up a test Microsoft SQL server 2016 on Windows Server 2016 virtual machine on a Cisco HyperFlex All-Flash system. Cisco recommends following the guidelines mentioned in: http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/solutions/sql-server-on-vmware-best-practices-guide.pdf to have an optimally performing SQL server database configuration.

1. Before proceeding with creating guest VM and installing SQL server on the guest, it is important to gather the following information. It is assumed that information such as IP addresses, Server names, DNS/ NTP/ VLAN details of HyperFlex Systems are available before proceeding with SQL VM deployment on HX All-Flash System. An example of the database checklist is shown in the following table.

Table 3 Virtual Interface Order with in HX-Series server

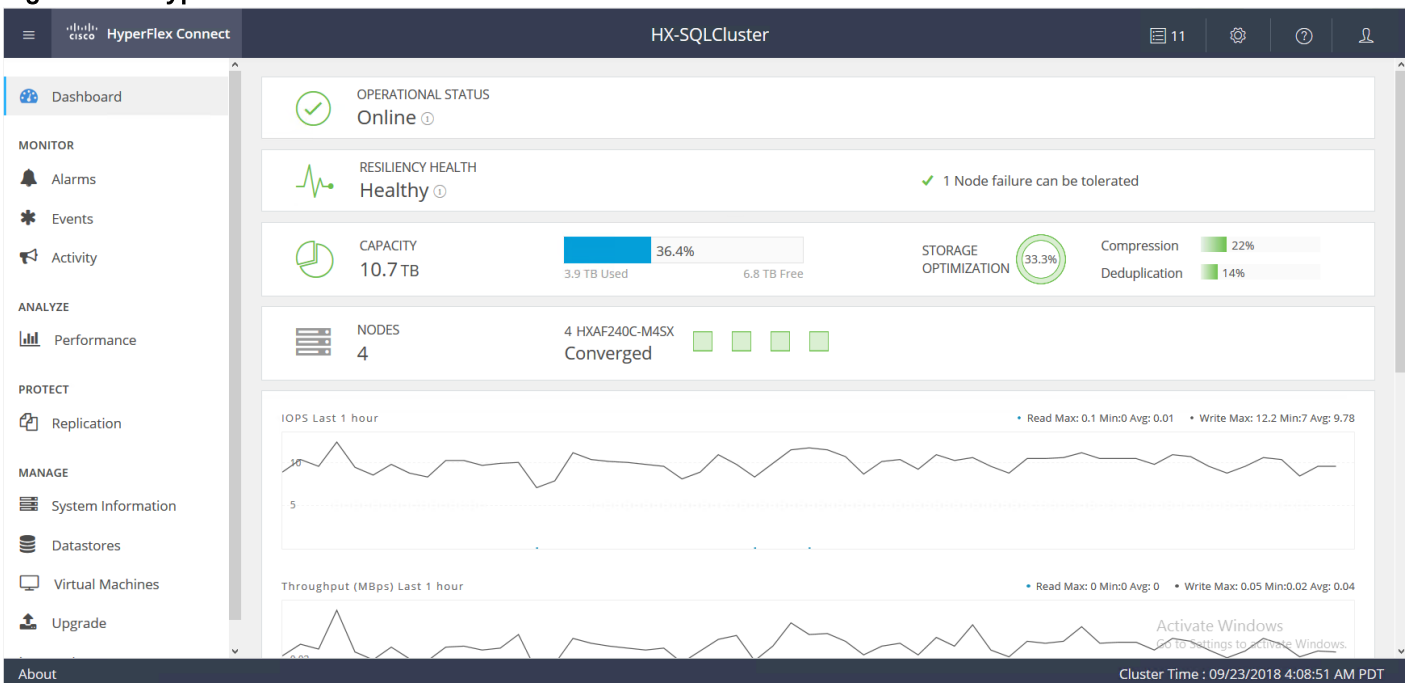
Component	Details
Cisco UCSM user name /password	admin / <<password>>
HyperFlex cluster credentials	Admin / <<password>>
VCenter Web client user name/password	administrator@vsphere.local / <<password>>
Datastores names and their sizes to be used for SQL VM deployments	SQL-DS1: 4TB
Windows and SQL server ISO location	\SQL-DS1\ISOs\
VM Configuration: vCPUs, memory, vmdk files and sizes	vCPUs: 8 Memory: 16GB OS: 40GB DATA volumes: SQL-DATA1: 350GB and SQL-DATA2: 350GB

Component	Details
	Log volume: SQL-Log: 150GB All these files to be stored in SQLDS-1 datastore.
Windows and SQL server License Keys	<<Client provided>>
Drive letters for OS, Swap, SQL data and Log files	OS: C:\ SQL-Data1: F:\ SQL-Data2: G:\ SQL-Log: H:\

2. Verify HyperFlex Cluster System is healthy and configured correctly. It is suggested to verify in the following ways:

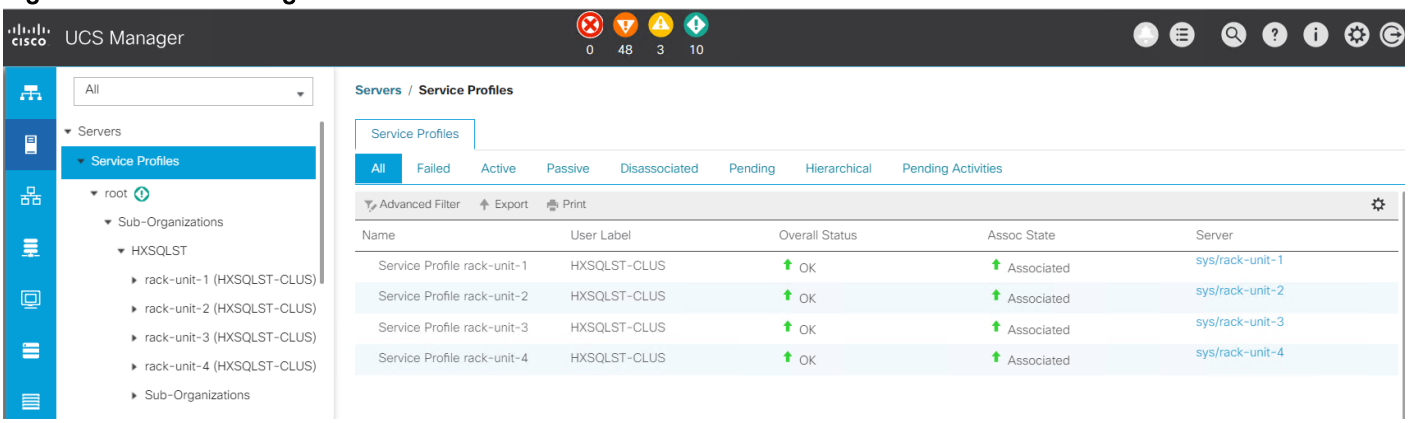
    a. Login to HX Connect dashboard using the HyperFlex Cluster IP address on a browser as shown below.

Figure 17 HyperFlex Cluster health status



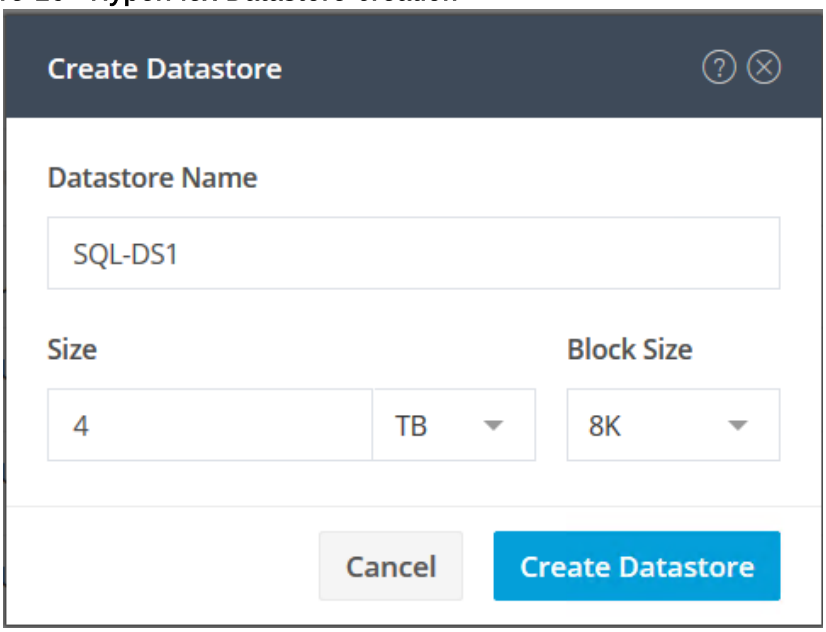
    b. Make sure that VMware ESXi Host service profiles in the Cisco UCS Manager are all healthy without any errors. Following is the service profile status summary screenshot from Cisco UCS Manager UI.

Figure 18 UCS Manager Service Profile



3. Create datastores for deploying SQL guest VMs and make sure the datastores are mounted on all the HX cluster nodes. The procedure for adding datastores to the HyperFlex system is given in the HX Administration guide. The following figure shows creation of a sample datastore. Block size of 8K is chosen for datastore creation as it is appropriate for SQL server database.

Figure 19 HyperFlex Datastore creation



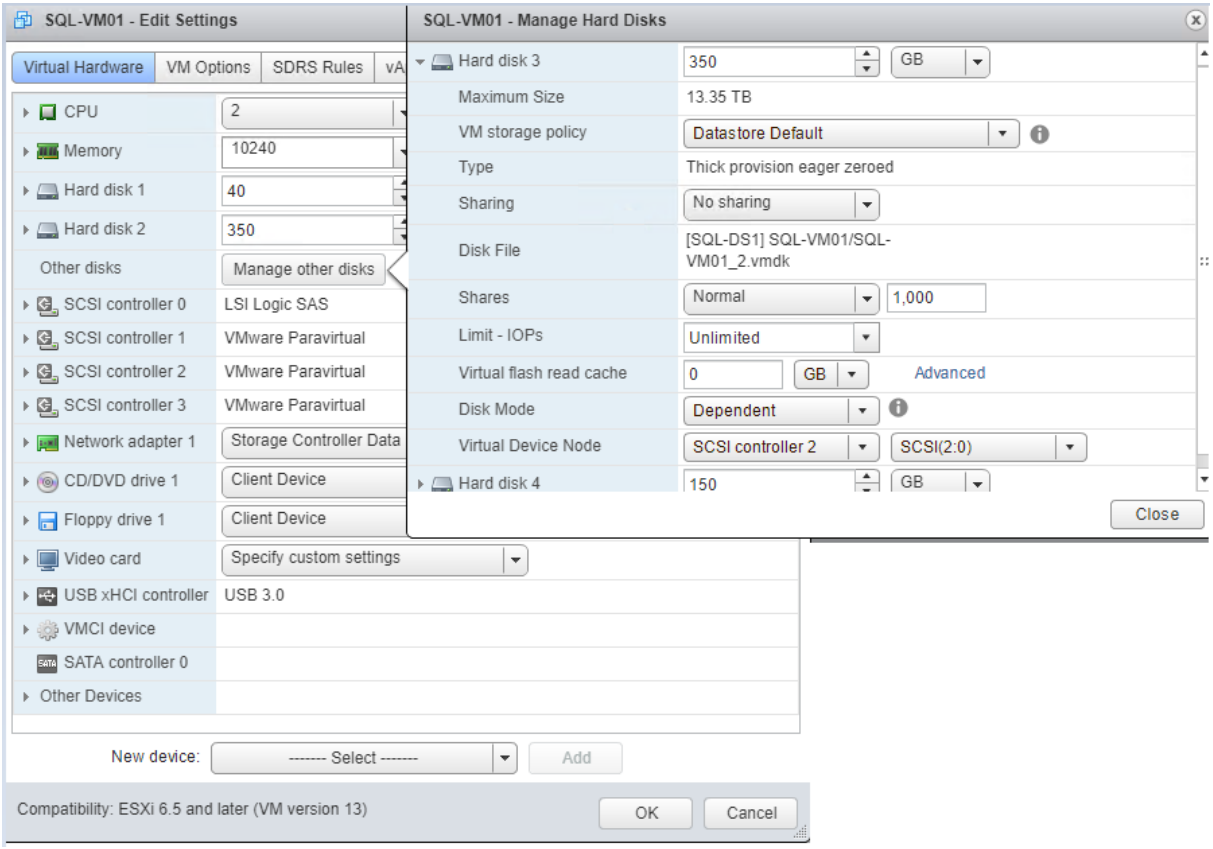
4. Install Windows Server 2016 virtual machine as per the instructions given in the VMware article http://partnerweb.vmware.com/GOSIG/Windows_Server_2016.html. As described before in the Deployment Procedure section of this guide, make sure that the OS, data and log files are segregated and balanced by configuring separate Paravirtual virtual SCSI controllers as shown in the figure 20 below. In VMware vCenter, go to Hosts and Clusters -> datacenter -> cluster -> VM-> VM properties -> Edit Settings to change the VM configuration as shown below figures 21.

Figure 20 Sample SQL Server Virtual Machine Disk Layout

Sample SQL VM Configuration	
vCPUs*	2
vRAM*	10GB

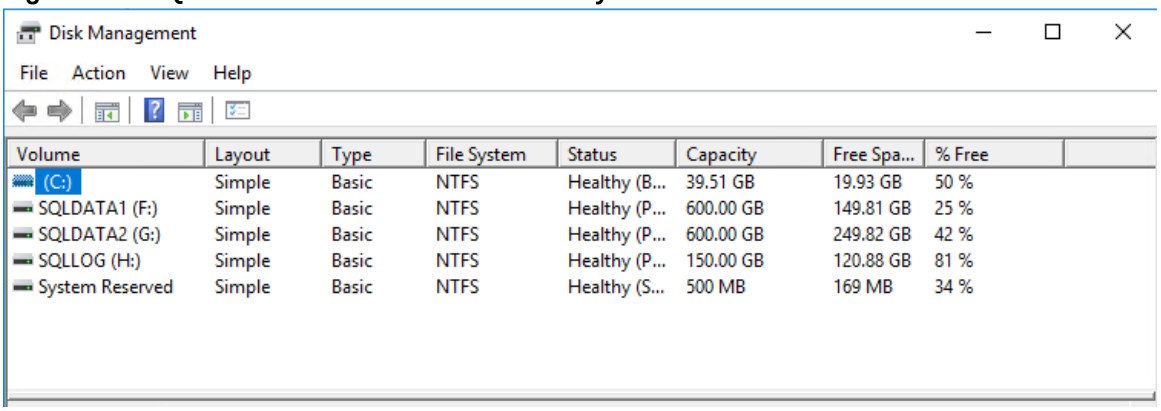
Disk Layout					
SCSI Controller	**Controller Type**	**Disks Purpose**	**Disk Size (GB)***	**Datastore**	**Provisioning Type**
SCSI Controller 0	LSI Logic SAS	OS Disk + SQL Server Binaries	40	SQL-DS1	Thick Provision Eager Zeroed
SCSI Controller 1	ParaVirtual	SQL Server Data Disk 1 (user database and TempDB data files)	350	SQL-DS1	Thick Provision Eager Zeroed
SCSI Controller 2	ParaVirtual	SQL Server Data Disk 2 (user database and TempDB data files)	350	SQL-DS1	Thick Provision Eager Zeroed
SCSI Controller 3	ParaVirtual	SQL Server Transaction Log Disk 1 (user database and TempDB Log files)	150	SQL-DS1	Thick Provision Eager Zeroed
* will change based on performane and capacity requirements					

Figure 21 SQL Server Virtual Machine Configuration



5. Initialize, format and label the volumes for Windows OS files, SQL server data and log files. Use 64K as allocation unit size when formatting the volumes. Following screenshot (disk management utility of Windows OS) shows a sample logical volume layout of our test virtual machine.

Figure 22 SQL Server Virtual Machine Disk Layout



6. Increase PVSCSI adapter's queue depth by adding a registry entry inside the guest VM as described in the VMware knowledgebase article: https://kb.vmware.com/s/article/2053145. Both RequestRingPages and MaxQueueDepth should be increased to 32 and 254 respectively. Since the queue depth setting is per SCSI controller, consider additional PVSCSI controllers to increase the total number of outstanding IOPS the VM can sustain.

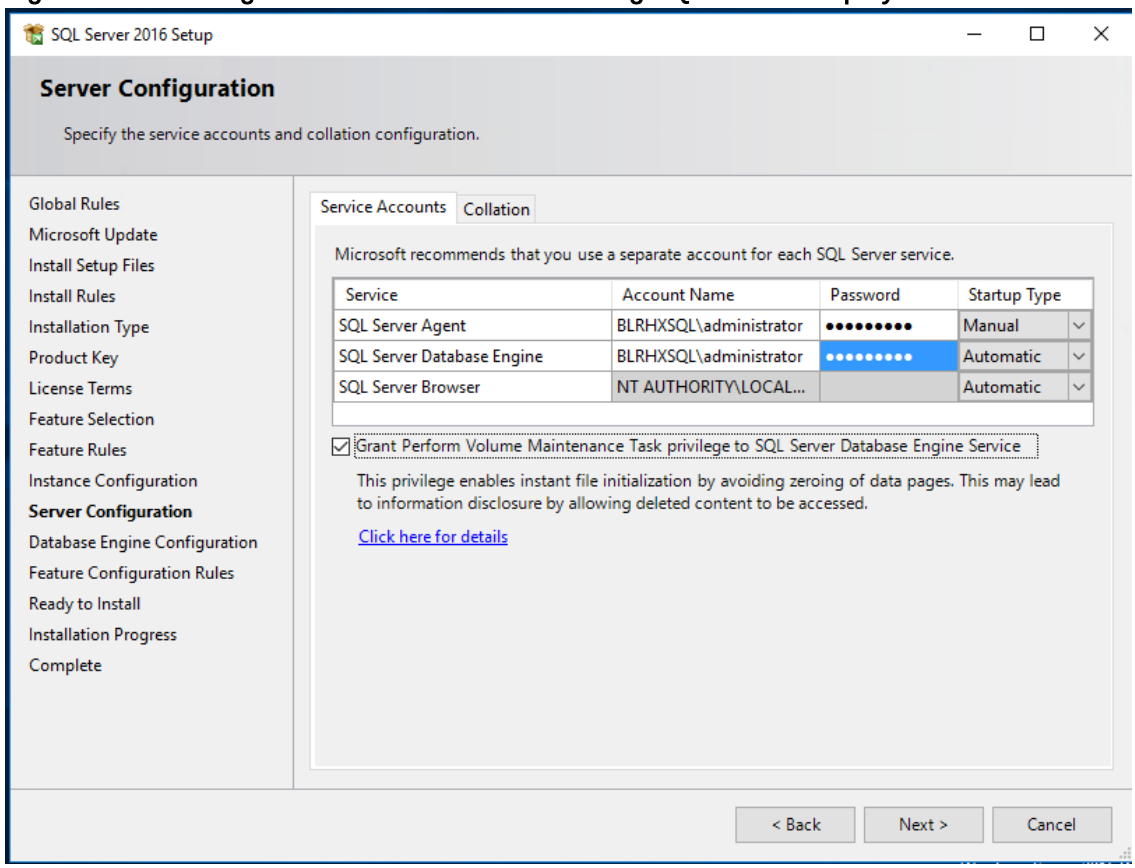
7. When the Windows Guest Operating System is installed in the virtual machine, it is highly recommended to install VMware tools as explained here. https://kb.vmware.com/s/article/1014294

8. Install Microsoft SQL Server 2016 SP1 on the Windows Server 2016 virtual machine. To install the database engine on the guest VM, following Microsoft documentation https://docs.microsoft.com/en-us/sql/database-engine/install-windows/install-sql-server-database-engine.

    a. Download and mount the required edition of Microsoft SQL Server 2016 SP1 ISO to virtual machine from the vCenter GUI. The choice of Standard or Enterprise edition of Microsoft SQL Server 2016 can be selected based on the application requirements.

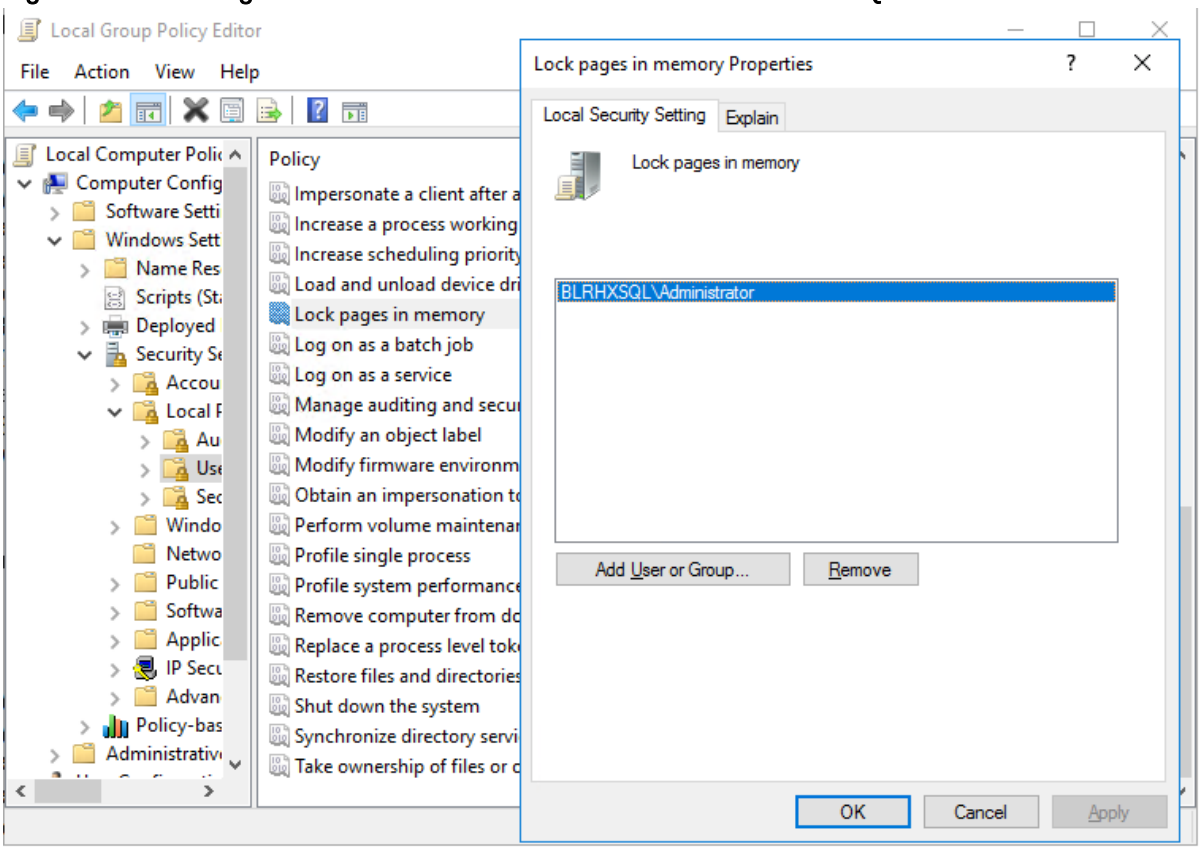
    b. On the Server Configuration window of SQL server installation, make sure that instant file initialization is enabled by enabling check box as shown in Figure 23. This enables the SQL server data files are instantly initialized avowing zeroing operations.

Figure 23 Enabling Instant File Initialization During SQL Server Deployment



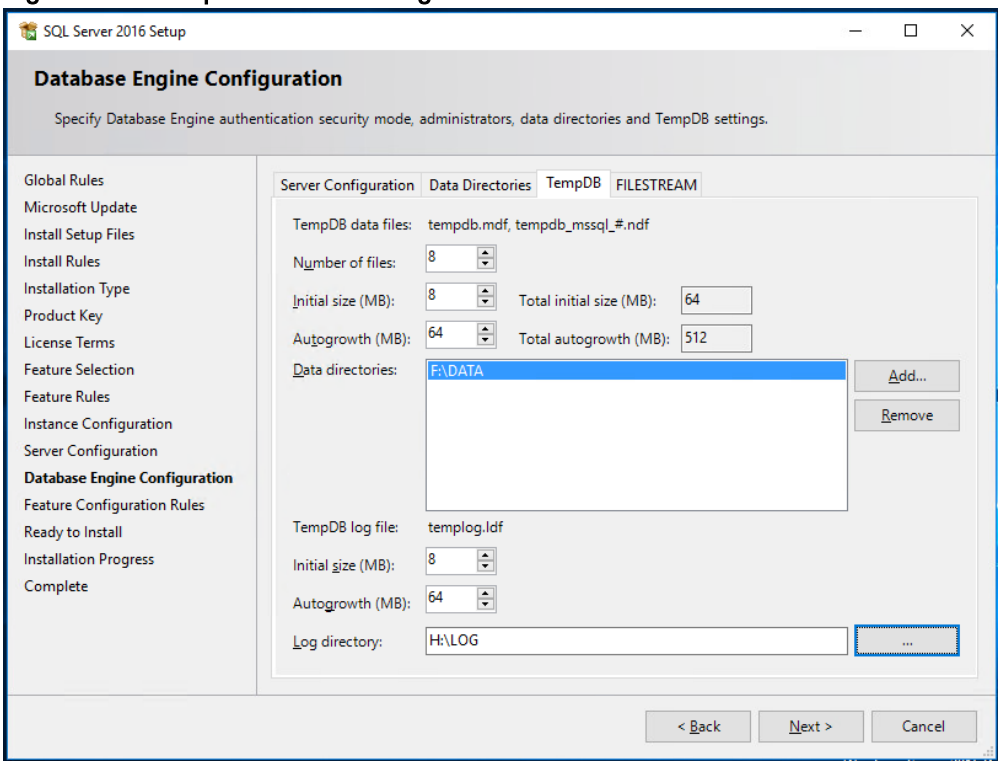
c. If the domain account which is used as SQL server service account is not member of local administrator group, then add SQL server service account to the “Perform volume maintenance tasks” policy using Local Security Policy editor as shown below.

Figure 24 Granting Volume Maintenance Task Permissions to the SQL Server Service Account



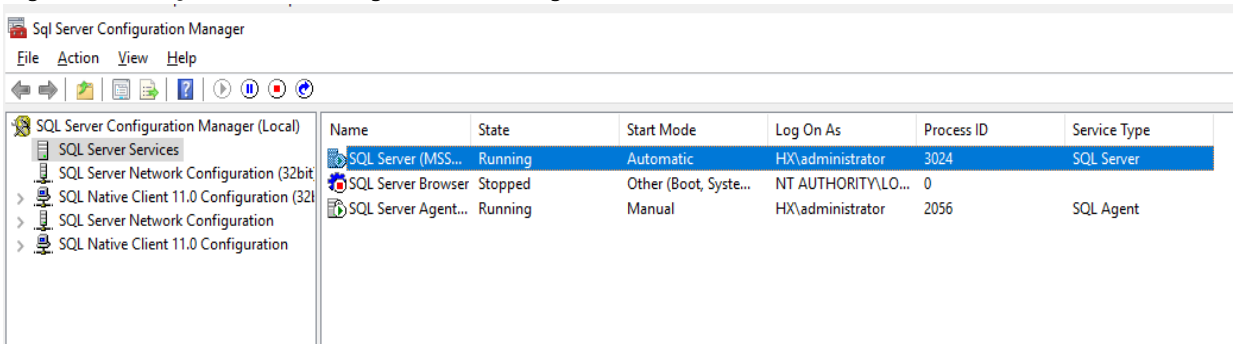
d. In the Database Engine Configuration window under the TempDB tab, make sure the number of TempDB data files are equal to 8 when the vCPUs or logical processors of the SQL VM is less than or equal to 8. If the number of logical processors is more than 8, start with 8 data files and try to add data files in the multiple of 4 when the contention is noticed on the TempDB resources (using SQL Dynamic Management Views). The following diagram shows that there are 8 TempDB files chosen for a SQL virtual machine which has 8 vCPUs. Also, as general best practice, keep the TempDB data and log files on two different volumes.

Figure 25 TempDB Data and Log files location



e. Once the SQL server is successfully installed, use SQL server Configuration manager to verify that the SQL server service is up and running as shown below.

Figure 26 SQL Server Configuration Manager



f. Create a user database using SQL Server Management studio or Transact-SQL so that the database logical file layout is in line with the desired volume layout. Detailed instructions are: https://docs.microsoft.com/en-us/sql/relational-databases/databases/create-a-database

Solution Resiliency Testing and Validation

This section discusses some of the tests conducted to validate the robustness of the solution. These tests were conducted on a HyperFlex cluster built with four HXAF240c M5 All-Flash nodes. The following table lists the component details of the test setup. Other failure scenarios (like failures of disk, network etc.) are out of the scope of this document. The test configuration used for validation is described below.

Table 4 Hardware and Software component details used in HyperFlex All-Flash Testing and Validation

Component	Details
Cisco HyperFlex HX data platform	Cisco HyperFlex HX Data Platform software version 3.5.1a Replication Factor: 3 Inline data dedupe/ compression: Enabled(default)
Fabric Interconnects	2x Cisco UCS 3rd Gen UCS 6332-16UP UCS Manager Firmware: 4.0(1b)
Servers	4x Cisco HyperFlex HXAF240c-M5SX All-Flash Nodes
Processors per node	2x Intel® Xeon® Gold 6140 CPUs @2.30GHz, 18 Cores each
Memory Per Node	768GB (32x 64GB) at 2666 MHz
Cache Drives Per Node	1x 400GB 2.5-inch Ent. Performance 12G SAS SSD (10X Endurance)
Capacity Drives Per Node	10x 960GB 2.5-inch Enterprise Value 6G SATA SSD
Hypervisor	VMware ESXi 6.5.0 build-8935087
Network switches (optional)	2x Cisco Nexus 9372 (9000 series)
Guest Operating System	Windows 2016 Standard Edition
Database	Microsoft SQL Server 2016 SP1
Database Workload	Online Transaction Processing With 70:30 Read Write Mix

The major tests conducted on the setup are as follows and will be described in detail in this section.

- Node failure Test
- Fabric Interconnect Failure Test
- Database Maintenance Tests

Note that in all the above tests, HammerDB tool (www.hammerdb.com) tool is used to generate required stress on the guest SQL VM. A separate client machine located outside the HyperFlex cluster is used to run the testing tool and generate the database workload.

Node Failure Test

The intention of this failure test is to analyze how the HyperFlex system behaves when failure is introduced into the cluster on an active node (running multiple guest VMs). The expectation is that the Cisco HyperFlex system should be able to detect the failure, initiate the VM migration from a failed node and retain the pre-failure state with an acceptable limit of performance degradation.

In our testing, node failure was introduced when the cluster is stressed with eight SQL VMs utilizing 35-40% of cluster storage capacity and 40% CPU utilization. When one node was powered off (unplug both power cables), all the SQL guest VMs running on the failed node successfully failed over to other node. No database consistency errors were reported in SQL server logs of the migrated VMs. After the VMs migrated to the other nodes, database workload is manually restarted. The impact observed on the overall performance (IOPS dip) because of failed node was around 5%. Later when the failed node was powered up, it rejoined the cluster automatically and started syncing up with the cluster. The cluster returned to the pre-failure performance within 5-10 minutes after the failed node was brought online (including the cluster sync-up time).

Fabric Interconnect Failure Test

The intention of this failure test is to analyze the performance impact in case of a fabric interconnect switch failure. Since Cisco UCS Fabric Interconnects (FIs) are always deployed in pairs and operating as a single cluster, failure of single Fabric Interconnect should not impact systems connected to them.

In our tests, we introduced the Fabric Interconnect failure when the cluster was stressed with SQL VMs utilizing 60-70% of cluster storage capacity and 70% CPU utilization. When one of the FIs was powered off (unplugged both power supplies), no impact was observed on the VMs running the workload. All the VMs remained fully functional and no VM migrations were observed in the cluster.

Database Maintenance Tests

In any hyperconverged systems, it is important to assess the impact caused by database maintenance tasks to the regular SQL workloads running on the same cluster. The maintenance activities typically have sequential IO pattern as opposed to random IO pattern of regular OLTP workloads. Usually in typical hyperconverged shared storage environments, caution must be exercised while running DB maintenance activities during the business hours as they may impact the regular operational workloads.

Following maintenance activities were carried out on a SQL guest VM deployed on an All-Flash cluster to assess the impact on the ongoing workload in the cluster. The cluster setup used is the same as detailed in table 4.

- Full database restore from an external backup source (backup source residing outside of HX cluster)
- Full database consistency check and complete backup to HX volume with in the virtual machine.
- Rebuilding indexes of two large tables (of size around 50GB)

- Exporting SQL data to flat files (located within HX cluster)
- Importing data from flat files into SQL (files located within HX Cluster)

Note that these maintenance activities are run on a separate SQL VM in parallel to the regular SQL VMs, which were used to exert stress on the cluster up to its 35-40% storage capacity utilization. As expected, full database restore caused 10 to 15% IOPS drop with IO which is understandable given the full database restore activity (100% sequential writes activity) done on a cluster which is already exercised to 70% resource usage by normal OLTP workload. Other activities (in the above list) had IOPS impact ranging from 3 to 5% with marginal increase in latencies.

The amount of impact caused by the maintenance activities would typically depend on the replication factor and percentage of cluster resource utilization in addition to factors such as back up settings etc. On system with appropriate resource headroom (which is done by right capacity planning), the impact would be much lower. Below figure (HX performance dashboard GUI) shows the cluster behavior when the maintenance activities such as restore, backup and index rebuild are performed on a VM and when an operational workload is running on the other VMs in the same cluster.

Figure 27 Performance Impact analysis of Typical Database Maintenance Activities on Ongoing Database Workload

Database Performance Testing

This section contains examples of different ways in which Microsoft SQL server workloads can take advantage of the HyperFlex Data Platform architecture and its performance scaling attributes.

Single Large VM Performance

As discussed earlier in the paper – HyperFlex Data Platform uses a distributed architecture. One of the main advantages of this approach is that the cluster resources form a single, seamless pool of storage capacity and performance resources. This allows any individual VM to take advantage of the overall cluster resources and is not limited to the resources on the local node that hosts the VM. This is a unique capability and significant architectural differentiator that Cisco HyperFlex Data Platform provides.

There are a couple of deployment scenarios that are common in data centers which benefit from Cisco HyperFlex Data Platform:

- VM Hotspot – rarely do all the VMs in any shared virtual infrastructure show uniform utilization of resources. The capacity growth for the individual VMs usually are different and also their performance requirements are different at different points in time. With the HyperFlex distributed architecture, those hotspots are easily absorbed by the infrastructure without having capacity or performance hotspots in the infrastructure.
- Large VMs – Since the cluster presents a common pool of resources – it makes it possible to deploy large applications and VMs with performance and capacity requirements that exceed the capability of any single node in the cluster. The fact that Cisco HyperFlex supports compute only nodes to be a part of the cluster further strengthens this use case.

This section demonstrates the above-mentioned attribute of the architecture through a performance test with a large SQL VM. The cluster setup used is the same as detailed in table 4. The details of the VM configuration and the workload are provided in the below table. OLTP workload with 70:30 read write ratio was exerted on the guest VM. The workload stressed the VM up to 65-70% of CPU utilization which resulted in around 40% of ESXi host CPU utilization.

Table 5 VM and workload details used for Single Large VM test

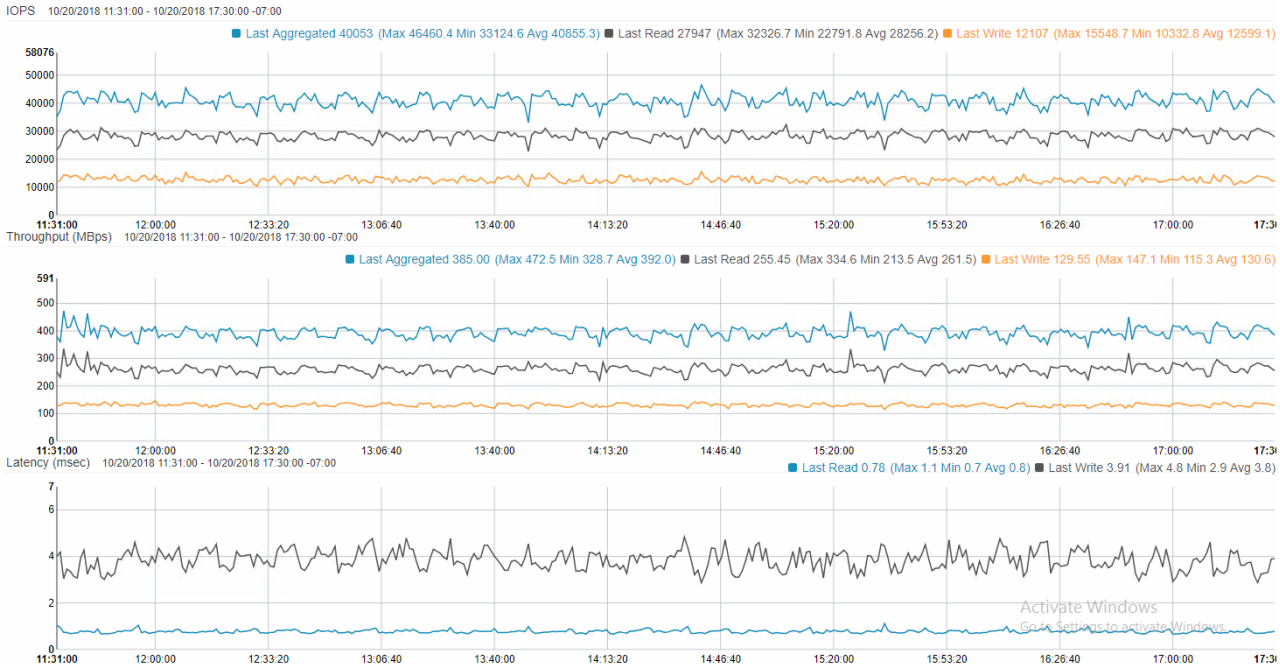
Configuration	Details
VM	8 vCPUs, 14GB Memory (16G assigned to sql) Two Data volumes & one Log volume (each with a dedicated SCSI controller) PVSCSI Max Queue Depth is set to 254
Workload	Tool Kit: HammerDB testing tool Users: 61 Data Warehouses: 8000 DB Size= 800GB

Configuration	Details
	RW Ratio: 70:30

The figure below shows the performance seen by the single VM running a large SQL workload for roughly 8 hours. There are a few noteworthy points listed as below:

- Large VM with a very large working set size can get sustained high IOPS leveraging resources (capacity and performance) from all 4 nodes in the cluster. Note that it is possible to scale to higher IOPS with an even larger VM size.
- Dedupe and Compression is turned on by default and the default setting is being used in this test as well.
- Delivers consistent performance throughout the test without any drop in the performance.

Figure 28 Single Large Working SQL Server Database Workload on HyperFlex All-Flash Cluster



This test demonstrates the ability that HyperFlex has leveraged the resources from all nodes in the cluster to satisfy the performance (and capacity) needs of any given VM.

Performance Scaling with Multiple VMs

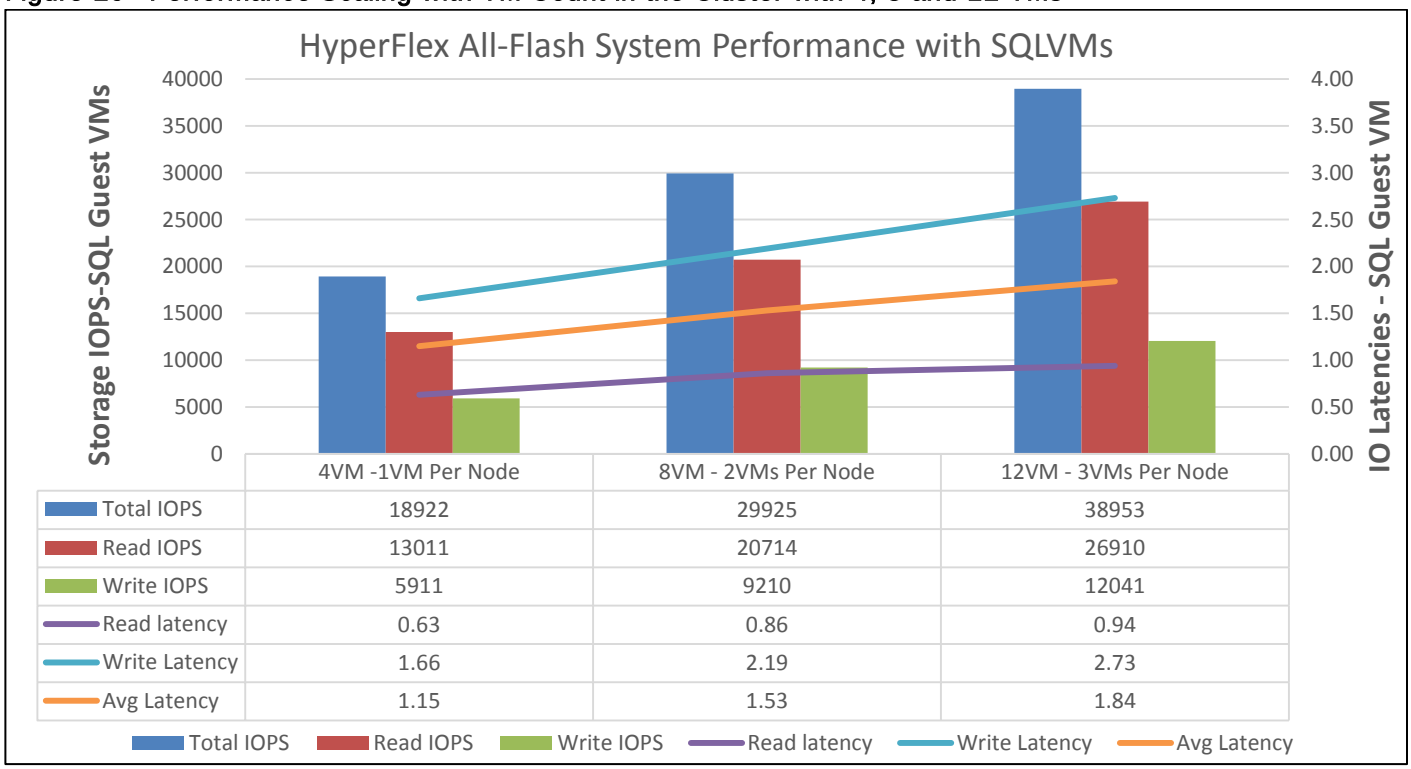
The seamless pool of capacity and performance presented by the HyperFlex cluster can also be accessed by multiple smaller VMs. In this section we are using multiple VMs that are running a SQL instance with HammerDB. The cluster setup used is the same as detailed in table 4. The details of VM configuration and workload are given in the below table. OLTP workload with 70:30 read write ratio was exerted on each guest VM. The workload stressed each VM with up to 35-40% of guest CPU utilization.

Table 6 VM and workload details used for Single Large VM test

Configuration	Details
VM	2 vCPUs, 10GB Memory (8G is assigned to SQL) Two Data LUNs and one Log LUN VM count scales in units of 4 (1 VM per node).
Workload	Tool Kit: HammerDB Users: 5 Data Warehouses: 1000 DB Size= ~100GB RW Ratio: 70:30

The below figure shows performance scaling seen by scaling the number of VMs in the cluster from 4 (1 VM per node), 8 (2 VMs per node) and finally 12 (3 VMs per node). The performance data shown in the below graphs is captured using Windows Perfmon tool. Note that these are sustained level of performance. Also, deduplication and compression is turned on by default and the default setting is being used in this test as well. If one or more of these VMs needed additional IOPS / throughout – there will be an increase in storage performance provided the VM itself is not CPU or memory bottlenecked and there is additional performance headroom available in the cluster.

Figure 29 Performance Scaling with VM Count in the Cluster with 4, 8 and 12 VMs

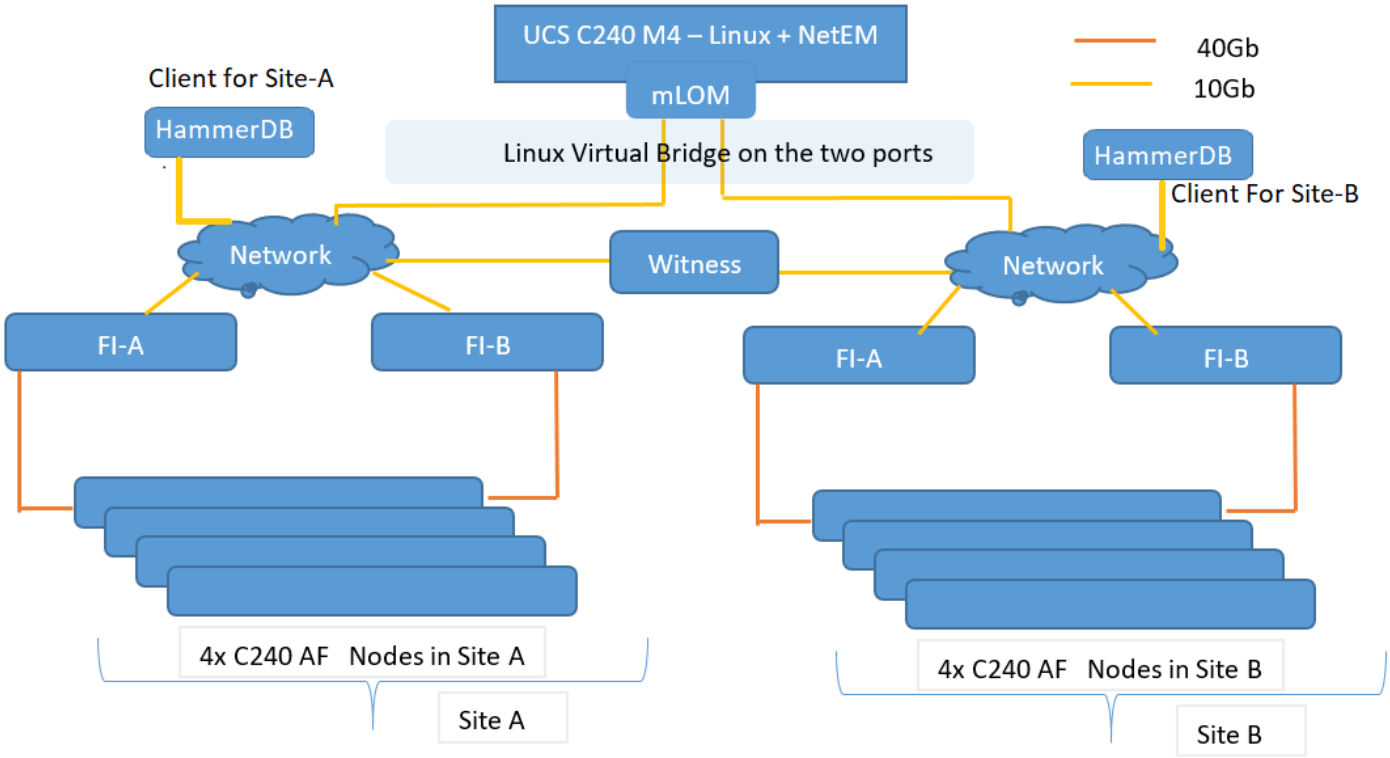


This test demonstrates HyperFlex Data Platform's ability to scale the cluster performance with a large number of VMs.

Performance testing with HyperFlex Stretched Cluster

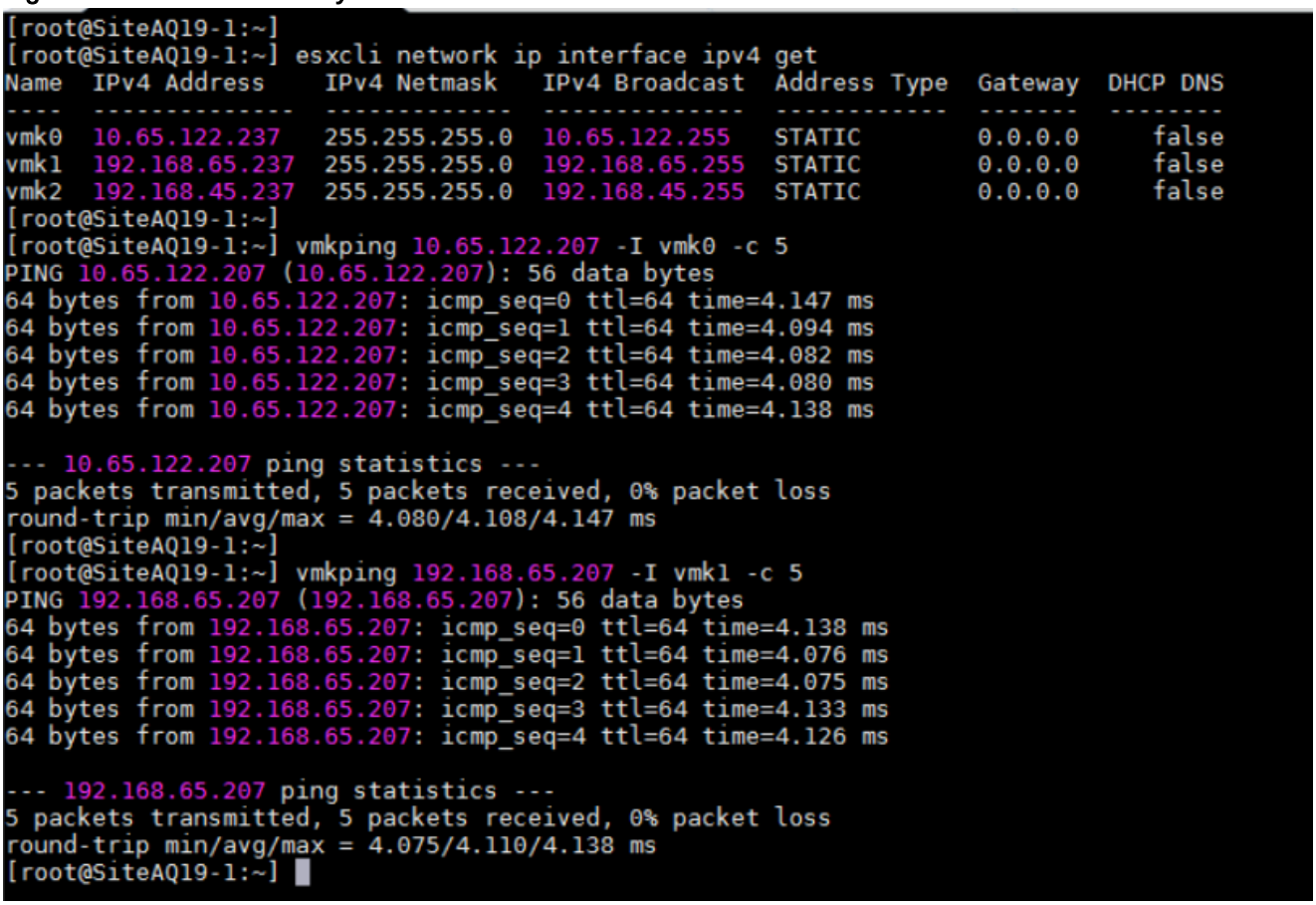
This section demonstrates the HyperFlex Stretched Cluster performance for SQL server database workloads to validate and assess HyperFlex Stretched Cluster functionality and performance. The below figure illustrates the pictorial view of eight node HyperFlex Cluster deployed in our labs.

Figure 30 HyperFlex Stretched Cluster used for SQL Server database workloads



The inter-site latency which is typically seen in real deployments of HyperFlex Stretched Cluster is simulated using a network latency emulator tool called NetEM: https://wiki.linuxfoundation.org/networking/netem. This Linux based emulator is connected to each site using a dedicated port. A virtual bridge, https://www.linuxjournal.com/article/8172, is configured on the two ports of the Linux machine to act as a simple bridge for passing all the traffic between the two sites. This allows to configure required network latency between the two sites. The below figure shows 4 milliseconds latency configured between site A and site B for management (10.x.x.x) and storage data (192.X.X.X) traffics.

Figure 31 Network latency between the two sites



The below table provides the software and hardware components used in the test bed.

Table 7 Hardware and Software component details used in HyperFlex All-Flash Stretched Cluster Testing and Validation

Component	Details
Cisco HyperFlex HX data platform	Cisco HyperFlex HX Data Platform software version 3.5(1a) Replication Factor: 2+2 Inline data dedupe/ compression: Enabled(default)
Fabric Interconnects on each site	2x Cisco UCS 3rd Gen UCS 6332-16UP UCS Manager Firmware: 4.0(1b)
Servers on each site	4 x Cisco HyperFlex HXAF240c-M5SX All-Flash Nodes
Processors per node on each site	2x Intel® Xeon® Gold 6140 CPUs @2.30GHz, 18 Cores each
Memory Per Node on each site	192GB (12x 16GB) at 2666 MHz

Component	Details
Cache Drives Per Node on each site	1x 400GB 2.5-inch Ent. Performance 12G SAS SSD (10X Endurance)
Capacity Drives Per Node on each site	10x 960GB 2.5-inch Enterprise Value 6G SATA SSD
Hypervisor	VMware ESXi 6.5.0 build-8935087
Network switches (optional) on each site	1x Cisco Nexus 9372 (9000 series)
Guest Operating System	Windows 2016 Standard Edition
Database	Microsoft SQL Server 2016 SP1
Database Workload	Online Transaction Processing with 70:30 Read Write Mix

On Site A, a datastore, DS1-SiteA, of 3TB size is created and affinity set to site-A. Similarly, another datastore, DS2-SiteB, of same size is created and set affinity too Site B. For each site, four virtual machines each running Microsoft SQL Server 2016 are deployed on the corresponding affinized datastore, thus making eight virtual machines running on this eight node Stretched Cluster. The below table describes the workload details and virtual machine configuration details used for testing.

Table 8 Virtual Machines and Database Workload details

Configuration	Details
VM	2 vCPUs, 10GB Memory (8G is assigned to SQL) Two Data LUNs and one Log LUN VM count scales in units of 4 (1 VM per node).
Workload	Tool Kit: HammerDB Users: 5 Data Warehouses: 1000 DB Size= ~100GB RW Ratio: 70:30

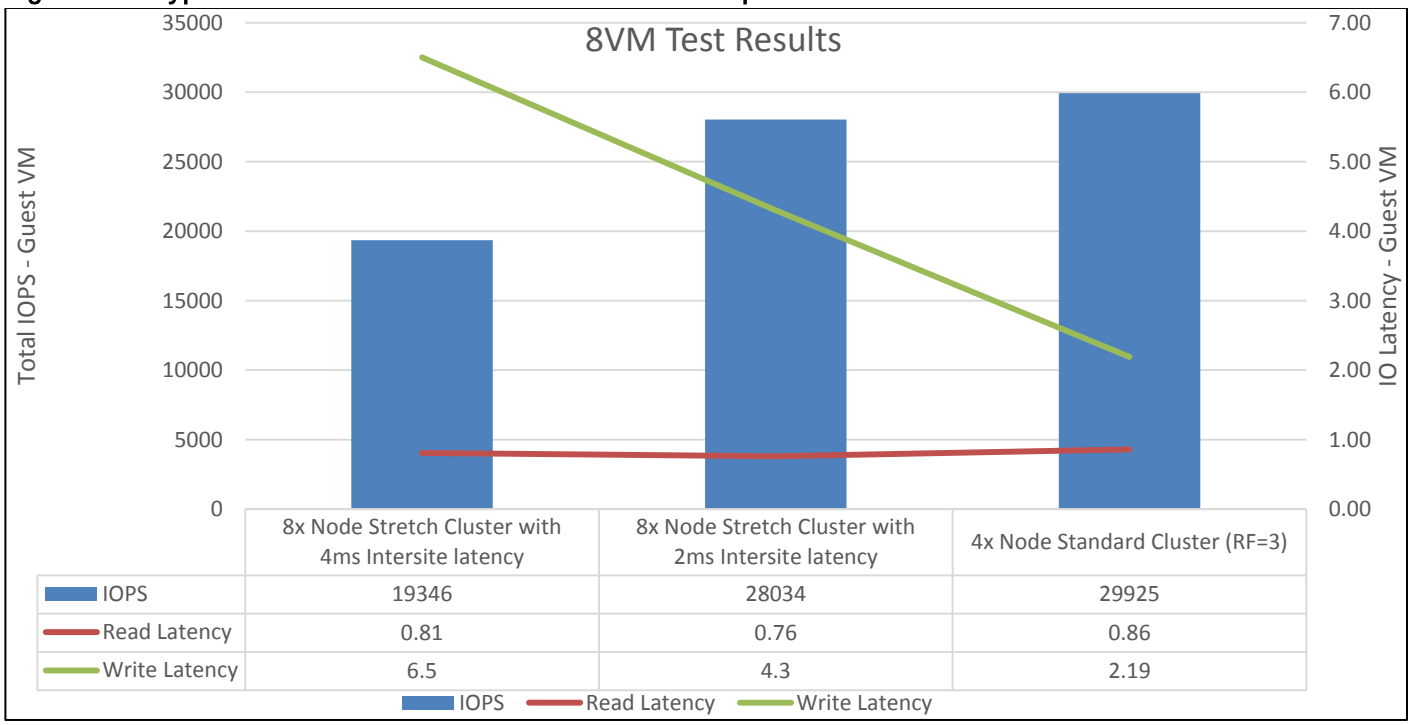
Two additional machines are used for hosting HammerDB tool (deployed one per site as show in figure 30), each running 4 client instances to run against the four SQL virtual machines deployed on the corresponding site.

Below section compare the performance test results for eight node Stretched Cluster with round trip inter-site latency of 2ms and 4ms against Four node All-Flash cluster (Standard, non-Stretched cluster setup)

The test bed configuration used for four node HyperFlex Standard All-Flash cluster is same as explained in the table 4 and the test bed configuration for HyperFlex Stretched Cluster is as mentioned in table 7.

The performance numbers shown in the below figure are captured using Windows Perfmon tool.

Figure 32 HyperFlex Stretched Cluster Performance comparison



From the above test results, the Stretched Cluster with 4ms inter-site latency delivered about 20,000 IOPS, and with 2ms inter-site latency it delivered about 28,000 and standard cluster delivered about 30,000 IOPS with write latencies of 6.5,4,3 and 2.1 milliseconds respectively.

As expected, the inter site network latencies does not have impact on read latencies as HyperFlex Stretched Cluster architecture allows reads are served locally. For better write performance, keep the inter-site performance as low as possible.

Additional validations for component failures such as disk, network, host, and complete site failure etc., see: https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/operating-hyperflex.pdf

Common Database Maintenance Scenarios

This section discusses common database maintenance activities and provides a few guidelines for planning database maintenance activities on the SQL VMs deployed on the All-Flash HyperFlex system.

The most common database maintenance activities include export, import, index rebuild, backup, restore and running database consistency checks on regular intervals. The IO pattern of these activities usually differs from business operational workloads hosted on the other VMs in the same cluster. The maintenance activities would typically generate sequential IO when compared to the business transactions, which generate random IO (in case of transactional workloads). When sequential IO pattern is introduced to the system alongside with random IO pattern, there is a possibility of impact on IO sensitive database applications. Hence caution must be exercised while sizing the environment or controlling the impact by running DB maintenance activities during the business hours in production environments. The following list provide some of the guidelines to run the management activities to avoid the impact on business operations.

- As a general best practice all the management activities such as export, import, backup, restore and DB consistency checks must be scheduled to run off business hours when no critical business transactions are running on the underlying HyperFlex system to avoid impact on the ongoing business operations. Another way of limiting the impact is to size the system with appropriate headroom.
- In case of any urgency to run the management activities in the business hours, administrators should know the IO limits of hyperconverged systems and plan to run accordingly.
- For clusters running at peak load or near saturation–when exporting a large volume of data from SQL database hosted on any hyperconverged system to any flat files, it should be ensured that the destination files are located outside of the HyperFlex cluster. This will avoid the impact on the other guest VMs running on the same cluster. For small data exports, the destination files can be on the same cluster.
- Most of the import data operations will be followed by recreation of index and statistics in order to update the database metadata pages. Usually Index recreation would cause lot of sequential read and writes hence it is recommended to schedule import data in off business hours.
- Database restore, backup, rebuilding indexes and running database consistency checks typically generate huge sequential IO. Therefore, these activities must be scheduled to run in the out of business hours.

In case of complete guest or database backups, it is not recommended to keep the backups in the same cluster as it would not protect against the scenario where the entire cluster is lost, for example, during a geographic failure, large scale power outage, etc. Data protection of the virtualized applications that are deployed on the hyperconverged systems are becoming one of the major challenges to the customers. Hence there is a need for most flexible, efficient and scalable data protection platform.

Cisco HyperFlex has integration with several backup solutions, for example, Cisco HyperFlex™ System's solution together with Veeam Availability Suite gives customers a flexible, agile, and scalable infrastructure that is protected and easy to deploy. More details on Veeam data protection platform is available at: https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/HX181_dataprotection_Veeam95.html

Workloads are rarely static in their performance needs. They tend to either grow or shrink over time. One of the key advantages of the Cisco HyperFlex architecture is the seamless scalability of the cluster. In scenarios where the existing workload needs to grow – Cisco HyperFlex can handle the scenario by growing the existing cluster's compute, storage capacity or storage performance capabilities depending on the resource requirement. This gives administrators enough flexibility to right size their environment based on today's needs without worrying about future growth.

Troubleshooting Performance

VMware discusses the common troubleshooting scenarios for the virtual machines hosted on VMware ESXi cluster at: https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2001003. This section discusses some of the commonly seen performance problems and the methodology followed to solve those.

The in-guest performance monitors (like windows perfmon) may not be able to collect the performance data which is based on time slices as the time period/interval is abstracted from the virtual machine by the VMware ESXi. Therefore, it is recommended to analyze the ESXTOP metrics for the performance troubleshooting of SQL server virtual machines. More details on interpreting ESXTOP statistics can be found at: https://communities.vmware.com/docs/DOC-9279.

Some of the commonly seen performance problems on virtualized/ hyperconverged systems are discussed as below.

 Tuning the SQL server transaction/ query performance is out scope of this document.

High SQL Guest CPU Utilization

When high CPU utilization with lower disk latencies on SQL guest VMs is observed and CPU utilization on ESXi hosts appears to be normal, then it might be the case that virtual machine is experiencing a CPU contention. The solution to overcome this is to add more vCPUs to the virtual machine as the workload is demanding more CPU resources.

When high CPU utilization is observed on both guest and Hosts, then one of the options to be looked at is upgrading to a higher performing processor. More options to solve this issue is mentioned in VMware vSphere documentation at: https://pubs.vmware.com/vsphere-60/index.jsp#com.vmware.vsphere.monitoring.doc/GUID-5F8147A1-6416-4D29-BA3D-E4CED3966016.html.

High Disk latency on SQL Guest

The following guidelines can be used to troubleshoot when higher disk latencies are observed on SQL guest VMs.

Use ESXTOP charts to identify the guest latencies versus kernel latencies and follow the options mentioned in the “Deployment Planning” Section.

In case of higher HX storage capacity utilization nearing expected thresholds (above 60% usage), SQL VMs might also experience IO latencies at both guest and kernel levels. In such case, it is recommended to scale up the cluster by adding new HX node to the cluster.

Summary

The Cisco HyperFlex HX Data Platform revolutionizes data storage for hyperconverged infrastructure deployments that support new IT consumption models. The platform's architecture and software-defined storage approach gives you a purpose-built high-performance distributed file system with a wide array of enterprise-class data management services. With innovations that redefine distributed storage technology, the data platform provides you the optimal hyperconverged infrastructure to deliver adaptive IT infrastructure. Cisco HyperFlex systems lower both operating expenses (OpEx) and capital expenditures (CapEx) by allowing you to scale as you grow. They also simplify the convergence of compute, storage, and network resources.

All-Flash configurations enhance the unique architectural capabilities of Cisco Hyperflex systems to cater to the high performance low latency enterprise application requirements. This makes it possible to utilize the entire cluster resources efficiently by the hosted virtual machines regardless the host. This enables the virtualized SQL server implementations as an excellent candidate for the high-performing Cisco HyperFlex All-Flash systems.

Lab solution resiliency tests detailed in this document show the robustness of the solution to host IO sensitive applications like Microsoft SQL Server. The system performance tuning guidelines described in this document addresses the platform specific tunings that will be beneficial for attaining the optimal performance for a SQL server virtual machine on Cisco HyperFlex All-Flash System. SQL server performance observed on Cisco HyperFlex systems focused in this document proves Cisco HyperFlex All-Flash system to be an ideal platform to host high performing low latency applications like Microsoft SQL Server database.

About the Authors

Gopu Narasimha Reddy, Cisco Systems Inc.

Gopu is Technical Marketing Engineer with Cisco UCS Data Center Solutions. He has over 10+ years of experience focusing on customer driven solutions for Microsoft SQL Server databases on various Operating systems and Hypervisors. His areas of interest include building and validating reference architectures, development of sizing tools in addition to assisting customers in SQL deployments.

Acknowledgements

- Babu Mahadevan Venkata Subramanian, Cisco Systems Inc.
- Sanjeev Naldurgkar, Cisco Systems Inc.
- Vadi Bhatt, Cisco Systems Inc.
- Rajesh Sundaram, Cisco Systems Inc.