Cisco HyperFlex Systems and Microsoft Exchange Server

Best practices for Microsoft Hyper-V

What you will learn

Your organization relies on email. Users expect fast, consistent, and reliable access to growing volumes of large and small messages alike, placing increasing demand on your Microsoft Exchange Server deployments. Achieving the right balance of computing, memory, network, and storage resources requires careful planning to help ensure that expected availability and performance characteristics are met. As a result, your administrators must find ways to replicate data, balance workloads, and keep email services running in the event of a hardware or software failure.

Cisco HyperFlex™ systems provide a new approach to IT infrastructure for Microsoft Exchange Server deployments. The systems unify server, storage, network, and management stacks into a highly available cluster without sacrificing performance. This document outlines best practices for deploying Microsoft Exchange Server on these innovative hyperconverged systems, and the reasoning behind the guidelines for implementation.

Cisco HyperFlex™ Systems with Intel® Xeon® Scalable processors

Contents

Cisco HyperFlex systems

Cisco HyperFlex™ systems use a software-defined infrastructure approach to unlock the full potential of hyperconvergence. They combine software-defined computing in the form of Cisco Unified Computing System™ (Cisco UCS®) servers, software-defined storage with the powerful Cisco HyperFlex HX Data Platform, and software-defined networking with the Cisco UCS fabric, which integrates with the Cisco® Application Centric Infrastructure (Cisco ACI™) solution.

With both hybrid and all-flash memory storage configurations and a choice of management tools, Cisco HyperFlex systems deliver a preintegrated cluster that is up and running in less than an hour and that scales resources independently to closely match your Microsoft Exchange Server requirements (Figure 1). For an in-depth look at the Cisco HyperFlex architecture, see the Cisco white paper Deliver Hyperconvergence with a Next-Generation Platform.

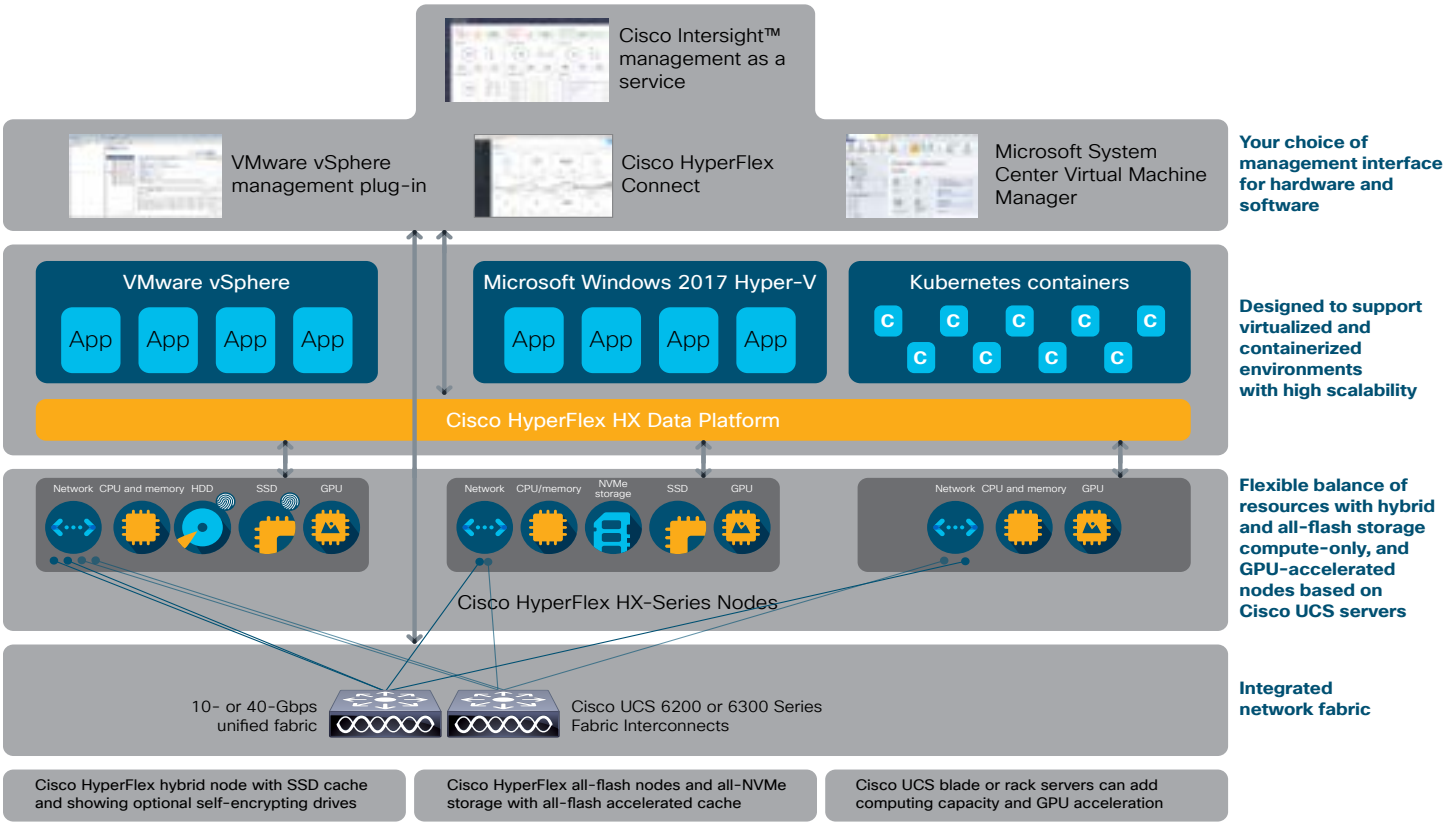

Figure 1 Cisco HyperFlex systems offer next-generation hyperconverged solutions with a unique set of features

Microsoft Exchange Server on Cisco HyperFlex systems

- **Full support.** Cisco HyperFlex HX-Series nodes fully support Microsoft Exchange Server.
- **SMB3 file shares.** Microsoft Exchange data that is stored in VHDX files can be placed on SMB3 file shares for the Mailbox Server role. SMB3 is used by compute and converged nodes to connect data stores on the HX Data Platform for Microsoft Hyper-V.
- **More performance, less compexity compared to converged solutions.** Unlike solutions that require the management of dozens to thousands of objects in many silos, HyperFlex systems use a single data store with no silos or performance hotspots.
- **Seamless scalability.** Global capacity, compute nodes, and converged nodes scale easily. Nodes and disks are simply added and automatically configured into the system. Data is rebalanced across all available devices in the cluster, increasing the cluster's aggregate capacity and performance.
- **Managed code.** Microsoft Exchange Server 2016 is the latest release from Microsoft and the second release in which the information store uses managed code.
- **New features.** Read the Technet article What's New in Exchange 2016.

Best practices for deploying Microsoft Hyper-V on Cisco HyperFlex systems running Microsoft Exchange Server

The following sections discuss major design guidelines and configurations for deploying Microsoft Exchange Server on Cisco HyperFlex systems. In addition, follow the best practices outlined by Microsoft.

High availability

High availability is important to messaging administrators. The HX Data Platform in Cisco HyperFlex systems builds in availability at the storage file system layer. All data is written in duplicate or triplicate (replication factor of 2 or 3). A copy of the data can be promoted to primary status if the primary storage controller that owns the primary copy is unavailable, without affecting virtual machines (VMs). In addition, you can configure VMs to be clustered resources in the Failover Cluster Manager or PowerShell software to protect against node or hypervisor failures.

```
PowerShell example:
Get-VM <VMName> | add-clustervirtualmachinerole
```

Microsoft Exchange Server includes high-availability options at the database layer when a database availability group (DAG) is deployed. Available features depend on the software edition deployed.

Microsoft Exchange Server database availability groups

DAGs provide an enterprise alternative to database mirroring that does not use shared storage. Initially implemented in Microsoft Exchange Server 2010, a DAG is a group of up to 16 Microsoft Exchange Servers that can each hold a copy of a Microsoft Exchange database. Three or four database copies are typical in production environments.

When deploying a highly available database, eliminating single points of failure helps increase availability in the event of a cluster failure due to power, network, or facility problems. A best practice is to stretch a DAG across multiple nearby clusters. As illustrated in Figure 2, only one copy of a database resides on a single node, and at least one copy of each database resides in a separate Cisco HyperFlex cluster. To help ensure that Microsoft Exchange Server VMs that house copies of the same database do not reside on the same physical host, see the failover clustering anti-affinity rules section of this document for information about increasing Microsoft Exchange Server availability.

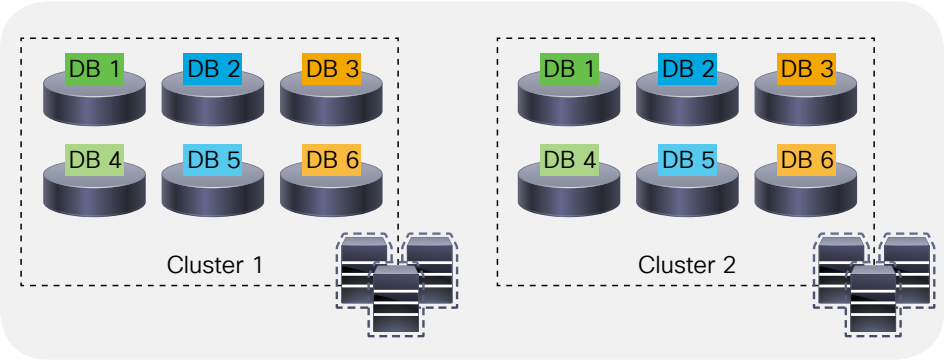

Figure 2 Isolate at least one database copy on a separate HX Data Platform cluster

Microsoft failover clustering

Microsoft Exchange Server 2016 can be installed on Microsoft Windows Server 2012, 2012 R2, and 2016. The failover clustering feature in the 2012 R2 release uses aggressive timeout values that can cause Microsoft Exchange nodes to consider another node as failed if a transient network issue lasts more than five seconds. A hotfix is available that increases the default timeout values to the same values used in Microsoft Windows Server 2016. For more information, see Tuning Failover Cluster Network Thresholds.

Sizing considerations

Sizing parameters and best practices include:

- **Microsoft Exchange Server estimation.** Use the Microsoft Exchange Server Role Requirements Calculator for Exchange 2013 and 2016 to estimate both the computing and storage requirements for your Microsoft Exchange deployment. The calculator estimates the number of servers, capacity, and I/O operations per second (IOPS) required.
- **Server selection.** With both Cisco HyperFlex hybrid and all-flash systems, capacity and computing resources can create a bottleneck and require careful planning. Recently announced support for 4-processor-socket Cisco UCS B460 blade server and Cisco UCS C460 rack server computing nodes can help reduce node counts.
- **Computing power.** A 4-node cluster supports more than 250,000 users from a storage IOPS perspective. However, the nodes may have sufficient computing resources for only a few thousand users depending on the core and Standard Performance Evaluation Corporation (SPEC) CINT2006 rate value of the CPUs in the system. Locate the information for your processor and enter it on the input tab of the Microsoft Exchange Server Role Requirements Calculator. Note that the calculator prefers symmetry, which can increase the value for the number of databases and servers required. This calculation must be balanced with proper planning to help ensure that two copies of the same database do not reside on the same physical server, using anti-affinity rules.
- **Cisco HyperFlex considerations.** Upload the completed Microsoft spreadsheet to the Cisco HyperFlex sizing tool to help size the computing and storage requirements of Microsoft Exchange and other workloads.

Planning IOPS

Five years ago, Microsoft Exchange users required 10 to 20 times more IOPS than the number required today. During that time, high-capacity hard-disk drive (HDD) storage area networks (SANs) achieved approximately 10 to 15 IOPS per nonparity SATA spindle. Even though today's clusters likely have many times the amount of I/O headroom that Microsoft Exchange requires, Microsoft Outlook online-mode users still experience a 2X IOPS penalty. Be sure to implement storage quality of service (QoS) so that neighboring VMs and backup jobs do not cause undue latency for Microsoft Exchange operations.

The next step is to consider storage capacity. You can add up to 23 devices to the persistent tier in Cisco HyperFlex HX240c nodes, and you can add converged nodes to help meet storage requirements. Consider the following:

- **Data deduplication.** Data deduplication is used on all storage in the cluster, including memory, solid-state disks (SSDs), and in the case of hybrid clusters, HDDs. By fingerprinting and indexing frequently used blocks, you can achieve high rates of deduplication using a small amount of memory, a high-value resource in cluster nodes. Deduplication rates vary with the content stored in user mailboxes. High rates of large deduplication is typical with operating system and application binary files, and higher deduplication rates are typical when there are many Microsoft Office document attachments. Environments with a large volume of attachments typically achieve deduplication rates of 10 to 35 percent, and smaller mailboxes without attachments typically achieve deduplication rates of 5 to 15 percent.
- **Inline compression.** The HX Data Platform uses high-performance inline compression on data sets to save storage capacity without negatively affecting performance. 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, which avoids typical read-modify-write penalties and significantly improves write performance. Overall, clusterwide compression rates for Microsoft Exchange Servers average 10 to 15 percent, although some options, such as in-guest encryption, significantly reduce achievable compression rates.
- **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 (data stores and virtual hard disk [VHDX] files) can present large amounts 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.

Sizing the database and log files

User database files and transaction logs grow as data is written to the database, and the behavior of transaction log files is influenced by the backup model. Typically, production databases are backed up, and at the completion of a full backup, transaction logs are truncated. When using a backup, the transaction log logical unit number (LUN) should be sized three times greater than the change rate of the backup window, with three days of change typical for deployments. With this approach, a backup that fails over a weekend will not cause a transaction log to run out of space.

With Exchange Server native data protection, which requires at least three nonlagged copies of the database, Microsoft Volume Shadow Copy Service (VSS) backups are not run, and circular logging is enabled on the database. In this configuration, the transaction logs do not require a buffer of additional space to protect against a backup failure.

Selecting the database size

At the smallest scale, a database has a single open transaction log file and Microsoft Exchange database (EDB) file. The size of the database directly affects the total number of databases in the configuration, and you must use care to help ensure that the database can be restored within the time specified in the service-level agreement (SLA). The use of fewer, larger databases can complicate the balancing of databases across nodes in the DAG, as copies of the same database must not reside on the same physical hosts.

Selecting the LUN layout

In a VM, the LUN, or logical disk, is a VHDX file stored in the Cisco HyperFlex data store. Usually after you add a disk to a Microsoft Windows VM, the disk is offline and must be brought online, initialized, and formatted prior to use.

Selecting the globally unique identifier partition table and file system

When New Technology File System (NTFS) partitions are initialized, the globally unique identifier (GUID) partition table (GPT) is preferred, as it has more file system redundancy in place and can be used for partitions larger than 2 TB. NTFS has been the recommended and default file system since Microsoft Exchange Server was launched in the 1990s. The new Resilient File System (ReFS) was given tentative support for Microsoft Exchange Server 2013, and now with Microsoft Exchange Server 2016, Microsoft recommends ReFS with integrity streams disabled in the Microsoft Exchange 2016 Preferred Architecture. If you use ReFS, declare it when you create the DAG.

```
PowerShell example for formatting a disk using ReFS:
Format-Volume –DriveLetter Z -FileSystem ReFS -AllocationUnitSize 65536
-SetIntegrityStreams $false

Create the DAG and specify the use of ReFS:
New-DatabaseAvailabilityGroup -Name MYDAG -FileSystem ReFS
```

Selecting the allocation unit size

For formatting, Microsoft recommends a 64-KB allocation unit size (AUS), sometimes referred to as the cluster size. If you use the default setting, the size will be 4 KB until you have very large partitions. A 64-KB AUS is recommended for disks that house user databases and user transaction logs.

```
Format a disk using ReFS and set the AUS to 64 KB:
Format-Volume –DriveLetter Z -FileSystem ReFS -AllocationUnitSize 65536
-SetIntegrityStreams $false
```

Isolating the page file, operating system, database, and log file

It is possible to run hundreds of databases in a single DAG, particularly when the databases require few IOPS and do not warrant careful isolation of every workload and file type. As databases scale, or for those databases that contain very heavy user workloads, isolation of everything can improve performance, or at least prevent file system fragmentation. Although Microsoft Exchange supports easy file migration with PowerShell, migration causes a service interruption when the database is dismounted before it is moved.

```
Move a database:
Move-DatabasePath -Identity DB01 -EdbFilePath C:\NewFolder\DB01.edb
```

If there must be no service interruption, a new database can be created at the new location, and the users can be moved to the new database.

```
Move users to the new database:
New-MoveRequest -Identity 'rquimbey@cisco.com' -TargetDatabase "DB01"

Remove the old database after the move:
Remove-MailboxDatabase -Identity DB01
```

In an optimal deployment, everything is isolated in separate VHDXs. The recommended approach is to set the page file size to 1 x RAM + 257 MB on an isolated VHDX on which a complete memory dump can be saved. For some environments, space for a kernel memory dump is sufficient. For kernel dump sizing, see the Microsoft Page File—The Definitive Guide blog.

Microsoft Exchange Server licenses

Microsoft licenses several versions of Exchange Server, each with multiple editions. Be sure to check with Microsoft for the exact terms of your license. Two editions are available: a Standard edition (5-mailbox database maximum) and an Enterprise edition (100-mailbox database maximum).

In addition to physical server and virtual machine licensing costs for Exchange Server, Microsoft requires either a Standard or Enterprise client access license (CAL) for each user. The Enterprise CAL is an add-on, applied to a user of a Standard CAL, that provides access to the following features:

- Journaling per user and digital library plus journal decryption
- Unified messaging
- In-place archive and hold capabilities
- Data-loss prevention (DLP)
- Information protection and control (IPC)

Data protection

Cisco HyperFlex systems support:

- Microsoft Windows 2016 production checkpoints for quick backups
- Hyper-V replicas for replication
- Ready clones for fast and efficient deployment of testing and development environments

Snapshots

Many basic backup applications read entire data sets or all changed blocks since the last backup, with data transfers completing as fast as the storage system and operating system can process them. Because Cisco HyperFlex systems are built on Cisco UCS, with fast 10 Gigabit Ethernet ports on each host, this process can result in multiple gigabytes per second of backup throughput with just a few simultaneous backup jobs. These basic backup applications, such as Microsoft Windows Server Backup, should be scheduled during off-peak hours. This is particularly important for the initial backup operation if the application uses some form of change-block tracking, such as Microsoft Resilient Change Tracking.

Clones

In the HX Data Platform, clones are writable snapshots that can be used to rapidly provision copies of the Microsoft Exchange Server infrastructure for development and testing environments. These fast, space-efficient clones rapidly replicate storage volumes so that VMs can be replicated through simple metadata operations. Disk space is consumed in the clones only when data is written or changed in the VM. With this approach, hundreds of clones can be created and deleted in minutes. It also saves a significant amount of time, increases agility, and improves productivity compared to full-copy methods. As a result, your Microsoft Exchange Server administrators can easily clone a production database and test for compatibility with a patch or update to Microsoft Exchange Server, Microsoft Windows, or a custom front-end application, and then easily remove the clones after testing is complete.

Virtual machine configuration

This section presents best practices for configuring VMs.

VM computing and memory resources

Microsoft's previously published guidance on the recommended maximum number of cores and memory size for Microsoft Exchange 2013 also applies to Microsoft Exchange 2016.

High-performance power plan

Cisco HyperFlex systems set the Microfot Windows power policy to high performance. This parameter should remain set to high performance for storage controller performance.

That guidance recommends:

- **Memory.** A maximum of 192 GB of memory.
- **Cores.** A maximum of 24 cores per VM.
- **Virtual CPUs.** Microsoft recommends a 1:1 ratio of virtual CPUs (vCPU) to physical CPU resources. As a result, most nodes support one to three Microsoft Exchange Server VMs before consuming all computing resources. Fortunately, Cisco HyperFlex systems allow you to add compute-only nodes and use all storage devices in the cluster within a distributed data store, reducing the storage cost per Microsoft Exchange user.

VHDX layout

Small Microsoft Exchange Server deployments run efficiently with a single VHDX, the C: drive, that contains everything. As the number of database IOPS scales up, isolating each workload onto its own VHDX can increase performance and is a best practice. Create a separate VHDX for each of the following: the operating system, paging file, databases, and transaction logs. Be sure to follow LUN layout guidance and consider the number of user database files, both active and passive, to be deployed.

Virtual CPU non-uniform memory access

Virtual CPU non-uniform memory access (vNUMA) exposes the NUMA topology to the guest operating system. In multisocket motherboards, memory DIMMs are assigned to a socket. As a result, processes running on a given CPU experience a performance penalty when they access memory that is assigned to the other socket.

A simple guideline is to size the VM with a number of CPU cores that is less than or equal to the number of CPU cores that are on one NUMA node on the physical CPU. In most cases, that is equal to the number of cores in the processor. If more cores are required, use a multiple of the number of cores on the NUMA node. Microsoft Exchange Serveris not NUMA aware; however, Microsoft Windows Server 2012 R2 and 2016 are NUMA aware.

CPU and memory reservation

Microsoft Exchange Server is a compute-intensive application. Use care when assigning resources to help ensure that the hypervisor, operating system, and Microsoft Exchange Server do not compete for computing and memory resources. Taking the precaution of setting memory and CPU reservation on important Microsoft Exchange Server VMs can protect the VMs from unanticipated resource consumption by smaller, overprovisioned VMs.

When performance is the primary goal for a VM, set the memory reservation equal to the provisioned memory, and reserve at least one CPU core's worth of megacycles in the VM's resource allocation settings. Microsoft does not support the overcommitment of memory resources for Microsoft Exchange Server, even though the CPU can be overcommitted to a ratio of up to 2:1.

Memory oversubscription and dynamic memory allocation

Memory oversubscription is the allocation of more memory to a VM than exists in the Microsoft Hyper-V host. Dynamic memory allocation is the hot addition of memory to an actively running VM. Although neither process is harmful, neither is supported by Microsoft for Microsoft Exchange Server VMs. This lack of support is partly the result of the way the software allocates memory at service startup. Rather than using a dynamic process, all memory allocation is static and set at service startup. The supported way to modify the amount of memory allocated to a Microsoft Exchange Server VM is to power off the VM and make the change by editing the VM's settings.

Hyper-V configuration

This section presents best practices for configuring Microsoft Hyper-V on Cisco HyperFlex systems running Microsoft Exchange Server.

Anti-affinity

Many enterprises that deploy Microsoft Exchange Server have more Microsoft Exchange Server VMs than they have nodes in their Microsoft Hyper-V cluster. As a result, a simple anti-affinity policy that includes all Microsoft Exchange Server deployments is not possible as a host will be required to run multiple Microsoft Exchange Server VMs. Careful planning and database placement can support the creation of anti-affinity sets and can prevent all copies of a database from residing on the same host. Rather than a random database placement, in which any two physical hosts will contain at least one copy of the same database, consider a planned database placement that avoids that scenario.

Table 1 describes a deployment in which no anti-affinity rules can be set, resulting in any two servers having at least one database that is the same.

Table 1 Deployment scenario that cannot support anti-affinity rules

HXHV1	HXHV2	HXHV3	HXHV4
DB1	DB2	DB3	DB4
DB5	DB6	DB7	DB8
DB4	DB1	DB6	DB3
DB7	DB5	DB8	DB2

Table 2 describes a deployment in which database placement allows two anti-affinity groups to be created: group 1 (HXHV1 and HXHV3) and group 2 (HXHV2 and HXHV4). When deciding to place Microsoft Exchange Server VMs on the same physical server, you could pair HXHV1 with either HXHV2 or HXHV4.

Table 2 Deployment scenario that supports two anti-affinity rules

HXHV1	HXHV2	HXHV3	HXHV4
DB1	DB2	DB1	DB2
DB3	DB4	DB3	DB4
DB5	DB6	DB5	DB6
DB7	DB8	DB7	DB8

Anti-affinity groups are created using the AntiAffinityClassNames property.

1. Create an anti-affinity group. The example below creates an anti-affinity group named ExchGroup1.

```
$ExchGroup1 = New-Object System.Collections.Specialized.StringCollection

$ExchGroup1.Add("Exchange Group 1")
```

2. Add the VMs to the anti-affinity group.

```
(Get-ClusterGroup –Name jetstress1).AntiAffinityClassNames = $ExchGroup1

(Get-ClusterGroup –Name jetstress2).AntiAffinityClassNames = $ExchGroup1
```

3. Finally, look at the list of VMs to see AntiAffinityClassGroup assignments.

```
Get-ClusterGroup |Select-Object -Property name,AntiAffinityClassNames
PS C:\> Get-ClusterGroup|Select-Object -Propertyname,
AntiAffinityClassNames
Name              AntiAffinityClassNames
----              ----------------------
Available Storage {}
Cluster Group     {}
jetstress1        {Exchange Group 1}
jetstress2        {Exchange Group 1}
jetstress3        {}
jetstress4        {}
```

Microsoft tools make it easy to create deployments in which VMs in an anti-affinity set are not placed on the same host with Microsoft System Center Virtual Machine Manager (VMM) availability sets. For more information, read "Configure Virtual Machine Settings in the VMM Computing Fabric".

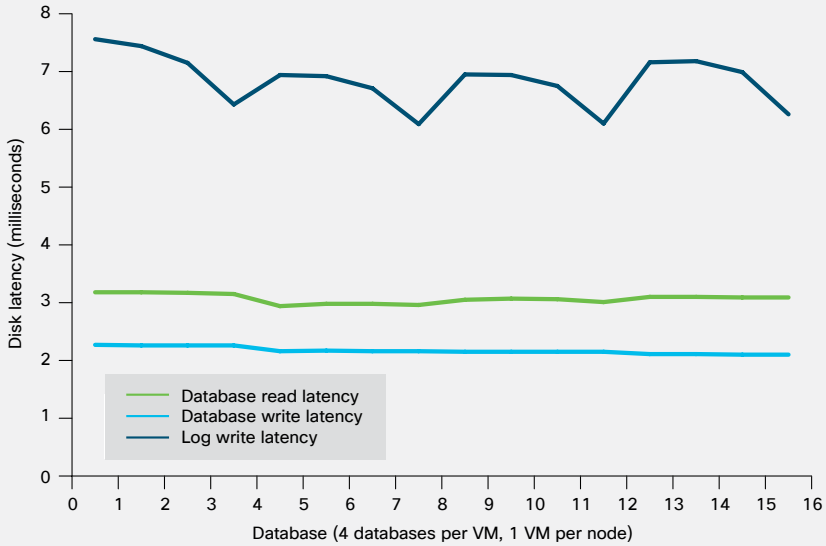

Figure 3 Jetstress latency results

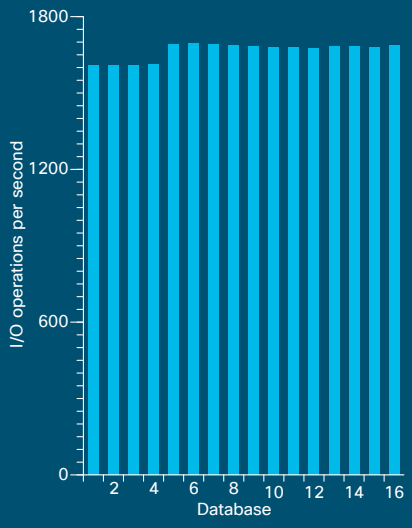

Figure 4 Database IOPS results

Jetstress configuration

The Microsoft Jetstress tool makes it easy to test the storage used for Microsoft Exchange data files. Jetstress simulates transaction log and database I/O, including additional database copies and database maintenance. You can optionally set hundreds of Jetstress settings, and storage best practices are often required to achieve good results. Out of the box, Cisco HyperFlex systems deliver stunning, low-latency results.

Optimization efforts are best spent on tuning computing parameters due to growing Microsoft Exchange Server 2016 memory and computing core requirements. Figure 3 depicts latency results for 16 x 650-GB databases with 4 databases on each VM and 1 VM per node in a 4-node all-flash cluster using a replication factor of 2. This configuration simulates 10 TB of databases running Jetstress with 16 threads.

I/O consistency is an important metric to consider, and Cisco HyperFlex systems perform very well with similar amounts of I/O at similar latency (Figure 4). When a storage array with similar outstanding threads, data set sizes, and I/O returns varying IOPS and latency, it is difficult to size correctly. It may also require significant work to protect VMs from unpredictable and noisy neighbors. Testing resulted in a database read average of 2.17 ms, a database write average of 6.85 ms, and a transaction log write average of 3.07 ms.

Doubling the threads to 32 increases the achievable IOPS by 46 percent with a penalty of approximately 0.9ms read latency. Increasing the number of threads to 60 improves IOPS by another 24 percent (an 80 percent improvement over 16 threads) and increases read latency to

For more information

- Cisco HyperFlex systems
- Microsoft Exchange Server 2016 licensing website
- Deliver Hyperconvergence with a Next-Generation Data Platform white paper
- "Ask the Perf Guy: Sizing Exchange 2016 Deployments" blog
- Microsoft Exchange Server Role Requirements Calculator

only approximately 4.5 ms. Available I/O capacity can be used by placing additional workloads on the cluster.

Microsoft Exchange Server is often hampered by a lack of computing resources. Cisco HyperFlex systems are uniquely positioned to deliver needed performance. Several technical advancements make this possible:

- Built-in IO Visor
- Capability to add Cisco UCS blade and rack servers to the HyperFlex cluster as compute nodes
- Independent scaling of storage and computing resources

For example, 26658.74 IOPS provides sufficient throughput to support more than 250,000 Microsoft Exchange users, while providing sufficient computing power with a 4-node cluster to power enough Microsoft Exchange Server VMs to support less than a fifth of those users.

Conclusion

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 provides 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 gives you the hyperconverged infrastructure you need to deliver adaptive IT infrastructure.

Cisco HyperFlex systems in hybrid and all-flash configurations lower both operating expenses (OpEx) and capital expenditures (CapEx) by allowing you to scale as you grow. They also simplify the convergence of computing, storage, and network resources. Size and acquire what you need now, and easily scale your storage with automated rebalancing after you add disks to the converged nodes or add more converged nodes. If more computing resources are required, use both Cisco UCS approved rack and blade servers, adding them to the cluster as compute-only nodes.

The Microsoft Exchange Server best practices discussed in this document are guidelines for high-performance (greater than 500 IOPS) VMs. Most VMs containing Microsoft Exchange Server work fine without configuring a multitude of settings at the storage, hypervisor, VM, or Microsoft Exchange Server layers. Unlike many traditional storage systems, you can easily increase the size of the VHDX or your data store. Native cloning is space efficient and fast and provides your Microsoft Exchange Server administrators with quick access to production data for testing and optimization without requiring additional storage capacity.

LE-68001-00 09/18