White Paper

Cisco HyperFlex 2.0 and Microsoft Exchange Server Best Practices

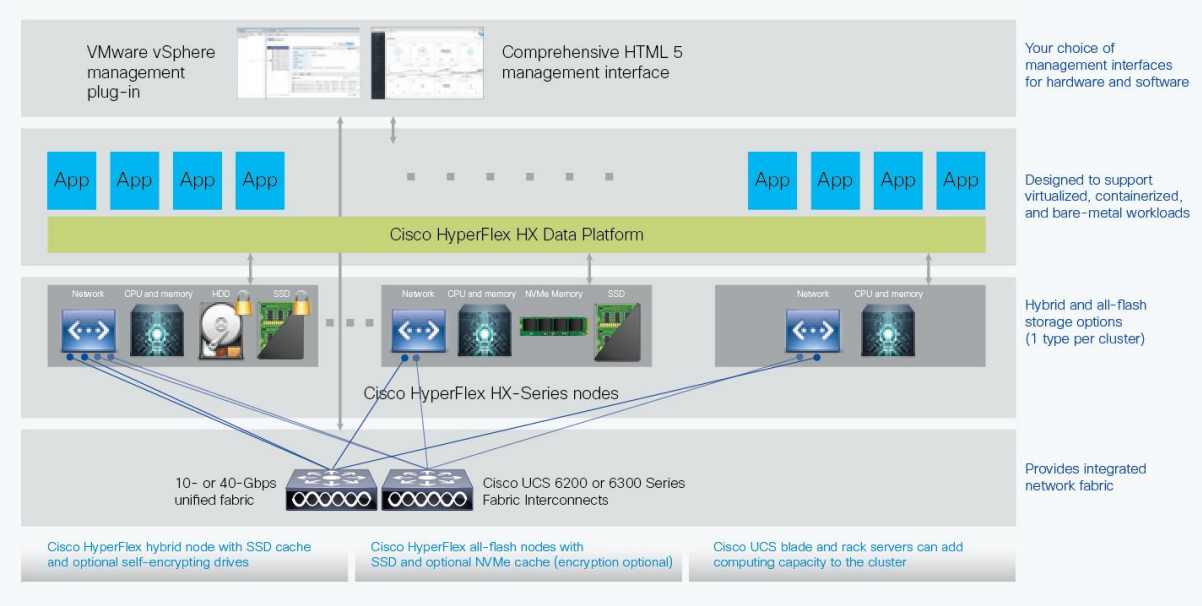
Author: Robert Quimbey

Cisco HyperFlex™ systems unlock the full 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™ (Cisco UCS®) servers; software-defined storage with the powerful Cisco HyperFlex HX Data Platform, and software-defined networking with the Cisco UCS fabric that integrates smoothly with the Cisco® Application Centric Infrastructure (Cisco ACI™) solution. Together with a single point of connectivity and hardware management, these technologies deliver a preintegrated and adaptable cluster that is ready to provide a unified pool of resources to power applications as your business needs dictate.

Cisco HyperFlex HX Data Platform 2.0

Cisco HyperFlex systems are designed with an end-to-end software-defined infrastructure that eliminates the compromises found in first-generation products. 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 set of features that only Cisco can deliver



Microsoft Exchange Server on Cisco HyperFlex systems

Cisco HyperFlex HX-Series nodes fully support Exchange Server, however Microsoft does not currently support NFS data stores for the Mailbox Server Role. NFS, in a proprietary implementation which is similar to pNFS in the HyperFlex distributed filesystem, is used by both compute and converged nodes to connect to data stores on the HX Data Platform and has been thoroughly tested by Cisco. Cisco HyperFlex HX-Series nodes increase performance and drastically reduce complexity in comparison to converged solutions. The tradeoffs to converged alternatives, is dozens to thousands of additional objects to manage in many silos of performance that scales with care and difficulty, rather than with HyperFlex, where there is only a single datastore, with no silos or performance hotspots. Global capacity and both compute and converged nodes can be scaled seamlessly with no additional administrative overhead other than that needed to simply add nodes or plug in more physical disks. The Cisco HyperFlex system automatically self-heals and rebalances the data across all available devices in the cluster, increasing the aggregate capacity and performance of the entire cluster as devices and nodes are added.

Microsoft Exchange Server 2016

Microsoft Exchange Server 2016 is the latest release from Microsoft and the second release, following Microsoft Exchange Server 2013, that uses managed code. For information about the new features, see the Technet article What's New in Exchange 2016.

Microsoft Exchange Server on Cisco HyperFlex systems: VMware best practices

When you implement Exchange Server on a Cisco HyperFlex system, you should follow the best practices proposed by VMware at https://blogs.vmware.com/apps/files/2016/07/Microsoft_Exchange_Server_2016_on_VMware_vSphere_Best_Practices_Guide.pdf.

The following sections present the major design guidelines and configurations for a successful Exchange Server implementation.

High availability

High availability is critically important to messaging administrators, and they have many tools at their disposal to meet this challenge. The HX Data Platform has availability built in at the storage file system layer, with all data written in duplicate or triplicate (replication factor of 2 [RF2] or replication factor of 3 [RF3]). The HX Data Platform can promote a copy of the data to primary status if the primary storage controller that owns the primary copy is unavailable, without affecting the virtual machines. To protect against a node or hypervisor failure, you can configure VMware High Availability (HA) to bring up the virtual machines on other nodes in the cluster.

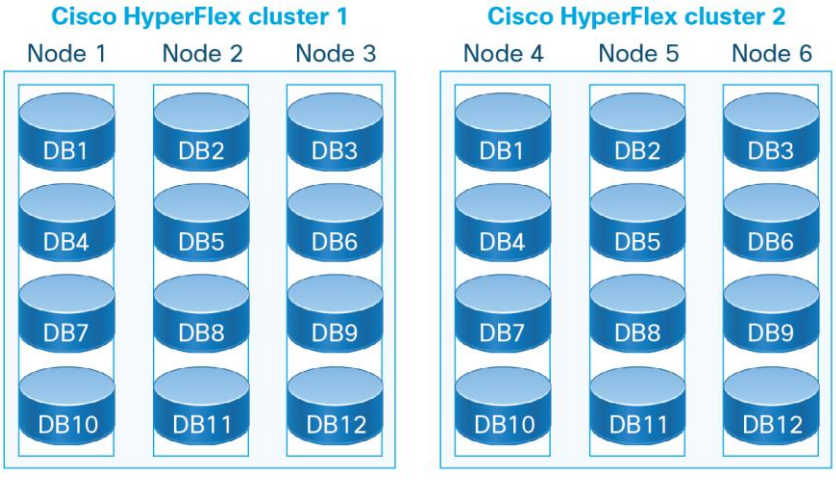
Exchange Server itself has high-availability options built in at the database layer when an Exchange Server Database Availability Group (DAG) is deployed. The features that are available depend on the edition of Exchange Server deployed. These high-availability features are described in the following sections.

Microsoft Exchange Server Database Availability Groups

DAGs provide an enterprise alternative to database mirroring that does not use shared storage. DAGs were first implemented in Exchange Server 2010. A DAG is a group of up to 16 Exchange servers that can each hold a copy of an Exchange database. In production environments, 3 or 4 database copies are most common.

When you are deploying a highly available database, eliminating single points of failure can increase availability in the event of a cluster failure due to power, network, or building problems. A best practice is to stretch a DAG across multiple clusters when they are nearby, as shown in Figure 2. In the figure, no more than 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 Exchange Server virtual machines that house copies of the same database do not reside on the same physical host, see the VMware Distributed Resource Scheduler anti-affinity rules section of this document for information about increasing Exchange Server availability.

Figure 2. Microsoft Exchange Server DAG best practice is to isolate at least one database copy on a separate HX Data Platform cluster



Microsoft failover clustering

The failover clustering feature in Microsoft Windows 2012 R2 uses very aggressive timeout values that can cause nodes to consider another node as failed if a transient network issue lasts longer than just 5 seconds. A hotfix is available that increases the 2012 R2 default timeout values to the same values used in Windows 2016. For more information, see Elden Christensen's Windows blog Tuning Failover Cluster Network Thresholds.

VMware also recommends changing these default values. For detailed information, see the VMware Exchange 2016 best practices guide.

Sizing considerations

Several factors can influence size requirements. You need to use the Exchange Server Role Requirements Calculator for Exchange 2013 and Exchange 2016 to accurately estimate both the computing and storage requirements for Exchange. The calculator estimates the number of servers, capacity, and I/O operations per second (IOPS) required. Five years ago, Exchange users required 10 to 20 times more IOPS than the number required today, and the higher-capacity hard-disk drive (HDD) SANs from the same time period could achieve approximately 10 to 15 IOPS per nonparity SATA spindle. Even though the cluster likely will have many times the amount of I/O headroom that Exchange requires, online-mode users of Outlook still experience a 2X IOPS penalty, and you should implement VMware Storage I/O Control so that a noisy neighbor virtual machine or backup job does not cause undue latency in Exchange.

With both Cisco HyperFlex hybrid and all-flash systems, capacity and computing resources can be a bottleneck by several orders of magnitude over IOPS and thus require careful planning. The 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. A 4-node cluster can support 150 to 200 thousand users from a storage IOPS perspective. However, the nodes may have only enough computing resources for 1 to 2 thousand users depending on the core and SPECint2006 rate value of the particular CPUs used. This amount can be determined by looking up the processor and entering its value on the input tab of the Exchange Server Role Requirements Calculator. Be aware that the calculator prefers perfect symmetry, which can bloat 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.

From a capacity perspective, you can add up to 23 devices to the persistent tier in the Cisco HyperFlex HX240c node and more converged nodes to meet that requirement. Note the potential capacity savings from both inline compression and deduplication, discussed in the following sections.

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 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. Deduplication rates vary with the content stored in user mailboxes, but many customers experience high rates of large deduplication with operating system and application binary files, and higher deduplication rates with heavy use of Microsoft Office document attachments. Environments with heavy use 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 Exchange Servers average 10 to 15 percent, although some options, such as in-guest encryption, significantly reduce the 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 Machine Disk [VMDK] 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 way that the transaction log files behave 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 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 being common. With this approach, if backup fails over a weekend, the transaction log will not run out of space.

With Exchange 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 Exchange database (EDB) file. The database size will directly affect 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 make balancing the databases across the nodes in the DAG difficult, because it is important that copies of the same database do not reside on the same physical host.

Selecting the logical unit number layout

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

Selecting the globally unique identifier partition table and ReFS

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

To format a disk using ReFS, enter the following in Microsoft PowerShell:

```
Format-Volume –DriveLetter Z -FileSystem ReFS -AllocationUnitSize 65536 -
SetIntegrityStreams $false
```

To declare ReFS during DAG creation so that reseeding will use the same file system, enter the following:

```
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 leave the setting at the default, the size will be 4 KB until you have very large partitions. The 64-KB AUS is recommended for any disk that will house both user databases and user transaction logs.

To format a disk using ReFS and setting the AUS to 64 KB, enter the following in PowerShell:

```
Format-Volume –DriveLetter Z -FileSystem ReFS -AllocationUnitSize 65536 -
SetIntegrityStreams $false
```

Isolating the page file, OS, database, and log

Many customers run dozens to hundreds of databases in a single DAG. Most of these databases require few IOPS and do not warrant careful isolation of every workload and file type. As databases scale, or for those few databases that contain very heavy user workloads, isolation of everything can improve performance, or at least prevent file-system fragmentation. Exchange enables easy file migration with PowerShell, but migration will cause a service interruption because the database must be dismounted if it is to be moved. The move command is as follows:

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

If no service interruption is required, a new database can be created at the new location, and the users can be moved to the new database with the following command:

```
New-MoveRequest -Identity 'rquimbey@cisco.com' -TargetDatabase "DB01"
```

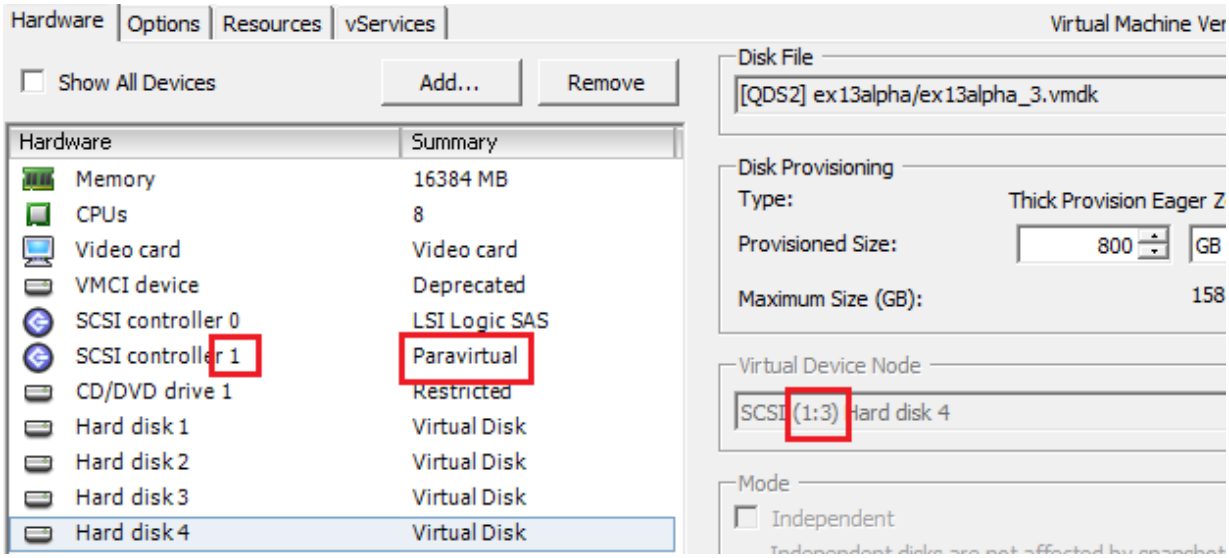
Use the following command to remove the old database after the move:

```
Remove-MailboxDatabase -Identity DB01
```

Optimally, you should isolate everything into separate VMDKs that are spread across up to four VMware Paravirtual Small Computer System Interface (PVSCSI) controllers. The recommended approach is to set the page file size to 1 x RAM + 257 MB on an isolated VMDK on which a complete memory dump can be saved. For some customers, enough space for a kernel memory dump is sufficient. For kernel dump sizing, see the Microsoft blog Page File—The Definitive Guide.

You should spread the database data file and log files across multiple PVSCSI controllers for optimal performance. Each controller will start with the number 0 and increment up to 3. The second number after the colon is the LUN or disk number. For example, Figure 3 shows 1:3, indicating controller 1 on disk 3.

Figure 3. Four PVSCSI controllers with each VMDK configured with round-robin load balancing across the controllers



Microsoft Exchange Server licensing

Microsoft licenses several versions of Exchange Server, each with a few different editions, in a variety of ways, so always check with Microsoft for the exact terms of the license. Two editions of Exchange Server are available: the Standard edition, which has a 5-mailbox database maximum, and the Enterprise edition, which has a 100-mailbox database maximum. In addition to the physical server or virtual machine licensing cost to run Exchange Server, Microsoft requires a client access license (CAL) for each user. The CAL is licensed as either Standard or Enterprise.

The Enterprise CAL is an add-on, applied to a user of a Standard CAL. It enables the following features:

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

Data protection

The Cisco HyperFlex system supports snapshots for quick backup and replication. It supports clones for fast and efficient testing and development.

Snapshots

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 the need to worry about consumption of physical storage capacity. The snapshot data can be moved to a backup repository, or if the snapshots are retained, they can be restored instantaneously.

- Fast snapshot updates: When a snapshot contains modified data, 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, rather than performing a long consolidation process as needed by solutions that use a delta-disk technique.

Many basic backup applications read the entire data set or all the changed blocks since the last backup at a rate that is usually as fast as the storage can provide or the operating system can handle. This process can have performance implications because the Cisco HyperFlex system is built on Cisco UCS with very fast 10 Gigabit Ethernet on each host, which could 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, particularly the initial backup operation if the application uses some form of change-block tracking.

Full-featured backup applications, such as Veeam Backup and Replication Version 9.5 (v9.5), can limit the amount of throughput that the backup application can consume, which can protect latency-sensitive applications during the production hours. With the release of Veeam v9.5 Update 2, Veeam is the first partner to integrate HX Data Platform native snapshots into its product. HX Data Platform native snapshots do not experience the performance penalty of delta-disk snapshots and do not require intensive disk I/O operations during snapshot consolidation when snapshots are deleted.

Particularly important for Exchange Server administrators is the capability to take an agentless VSS quiesced snapshot that is application aware. Veeam Explorer for Exchange Server can provide recovery of specific individual items and e-discovery options. Veeam Explorer for Exchange Server can restore Exchange Server databases from the backup restore point, allow roll-forward recovery, and even export items to a personal storage table (PST)—all without taking the virtual machine or Exchange Server services offline. The Veeam Explorer for Exchange is an object level explorer which opens the Exchange database inside of the Veeam Backup File instantly. This enables the Exchange Administrator to recover at the individual item level while lowering the recovery time objective.

Clones

In the HX Data Platform, clones are writable snapshots that can be used to rapidly provision copies of the Exchange Server infrastructure 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. Disk space is consumed in the clones only when data is written or changed in the virtual machine. 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. Exchange Server administrators can easily clone a production database and test for compatibility with a patch or update to Exchange Server, Windows, or a custom front-end application. They can then easily remove the clones after testing is complete.

Storage configuration: Configuring additional data stores

For most deployments, a single HX Data Platform data store is sufficient, resulting in fewer objects to manage. HX Data Platform is a distributed file system that is not vulnerable to many of the problems that face traditional systems that require data locality. A VMDK does not have to fit within the available storage of the physical node. If the cluster has enough space to hold the configured number of copies of the data, the VMDK will fit. Similarly, moving a virtual machine to a different node in the cluster is a host migration; the data itself is not moved.

In some cases, however, additional data stores may be beneficial. For example, an administrator may want to create an additional HX Data Platform data store to separate Exchange from other workloads. Because performance metrics can be filtered to the data store level, isolation of workloads or virtual machines may be desired. It is important that all VMDKs within a virtual machine reside on the same data store. The data store is always thinly provisioned on the cluster. However, the maximum data store size is set during data-store creation and can be used to keep a workload, a set of virtual machines, or end users from running out of disk space on the entire cluster and thus affecting other virtual machines.

Another good use for additional data stores is to assist in throughput and latency on high-performance Exchange Servers. If the cumulative IOPS of all the virtual machines on a VMware ESX host surpasses 10,000 IOPS, the system may begin to reach that queue depth. In ESXTOP, you should monitor the Active Commands and Commands counters, under Physical Disk NFS Volume. Dividing the virtual machines into multiple data stores while keeping all VMDKs from any one virtual machine in a single data store, or increasing the ESX queue limit of 256, can relieve the bottleneck.

If data store queues are causing the performance issues (which can be identified from ESXTOP reports: for instance, if reports show higher guest latencies than kernel latencies), follow these procedures:

- Increase the data-store queue depth from 256 (default) to 1024 using the following command at each node's VMware ESXi shell prompt:

```
esxcfg-advcfg –s 1024 /NFS/MaxQueueDepth
```

You can query to verify the change with the following command:

```
esxcfg-advcfg -g /NFS/MaxQueueDepth
```

The value of MaxQueueDepth is 1024.

- Evaluate the feasibility of deploying Exchange virtual machines with lower I/O demand in one data store and Exchange virtual machines with higher I/O demand in a different data store. This use of multiple data stores allows greater queue depth per data store based on the performance requirements.

Virtual machine configuration

This section presents some best practices for configuring virtual machines.

Virtual machine computing and memory resources

Microsoft's previously published guidance on the recommended maximum number of cores and memory size for Exchange 2013 currently also applies to Exchange 2016. That guidance recommends the use of a maximum of 96 GB of memory, and a maximum of 24 cores per virtual machine. Because Microsoft recommends a 1:1 ratio of virtual CPU (vCPU) to physical CPU, this configuration means that most nodes will hold only one to three Exchange Server virtual machines before consuming all the computing resources. The capability to add computing-only nodes using all the storage devices in the entire cluster in the distributed data store is a real differentiator and helps reduce the storage cost per user in Exchange.

SCSI controller

Disk I/O is queued at many levels in the stack, and understanding where bottlenecks can occur can help you make design decisions. In a virtual machine, the factor with the biggest impact is the queue depth set on the SCSI controller. By default, a virtual machine will use the LSI Logic SAS SCSI controller, which has an unchangeable queue depth of 32. VMware instead recommends the PVSCSI controller, which has a default queue depth of 64. A

virtual machine that is running Exchange Server and that requires fewer than 500 IOPS will probably be fine with the default settings, which simplifies hyperconvergence. Experienced Exchange Server administrators, however, are more cautious and are used to isolating everything to separate disks for performance or backup purposes. The Cisco HyperFlex best practice for Exchange Server is to use PVSCSI controllers.

Traditionally the PVSCSI controller was not supported as a boot device, and because a virtual machine can have a maximum of four SCSI controllers, one was used for boot, with up to three additional PVSCSI controllers for databases and transaction logs. With recent versions of Windows Server, you can change the original controller (SCSI controller 0) to PVSCSI after verifying that the driver is properly installed in Windows.

PVSCSI queue depth

For many of customers who deploy Exchange Server, the default settings will be sufficient. After an Exchange virtual machine nears the 500 IOPS threshold, which is unlikely in newer versions of Exchange, consider increasing the PVSCSI queue depth from the default of 64 to 254, as noted in the VMware knowledgebase. You should increase the RequestRingPages and MaxQueueDepth values up to 32 and 254 respectively. Because the queue-depth setting is configured per SCSI controller, consider adding PVSCI controllers to increase the total number of outstanding IOPS that the virtual machine can sustain.

A good indicator that you do not have enough queue depth is latency that is more than 10 percent higher in the guest system than the amount visible in the Cisco HyperFlex performance chart available in the Cisco HyperFlex user interface or ESXTOP. To verify this value in an existing live environment, check Windows Performance Monitor to see whether the cumulative active queue depth of all the VMDKs on the controller is sustained at greater than the queue depth of the controller during intensive I/O processing. For example, if two database files in separate VMDKs each experience sustained spikes of 80 in the queue while you are using the LSI SAS controller, you can switch to PVSCSI to double the controller queue depth (from 32 to 64). Placing each VMDK on a separate PVSCSI controller would again double the available maximum queue depth (from 64 total to 64 each, or 128). Changing the registry setting for the PVSCSI queue from 64 to 254 would change the maximum queue depth available to the database from 128 to 508 in this example.

VMDK layout

Small Exchange Servers can run efficiently with a single VMDK, the C: drive, that contains everything. As the number of IOPS of a database scales up, isolating different workloads on their own VMDKs can increase performance. The best practice is to isolate all workloads on their own VMDKs. Create separate VMDKs for the operating system, paging file, databases, and transaction logs. Be sure to follow the LUN layout guidance and to consider the number of user database files, both active and passive, to be deployed.

VMXNET3 network interface card

When a virtual machine starts to approach a gigabit per second in bandwidth, consider switching to VMXNET3 network interface cards (NICs) instead of the default VMware E1000 network card. VMXNET3 is designed for the best performance and requires VMware Tools to be installed. Enable receive-side scaling (RSS) on VMXNET3 NICs, which allows the network drivers to spread the incoming TCP traffic across multiple CPUs, in the guest virtual machine on the VMXNET3 adapter.

Virtual CPU non-uniform memory access

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

guideline is to size the virtual machine with CPU cores with a quantity 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 the number of cores on the processor, but not always. If more cores are required, use a multiple of the number of cores on the NUMA node. Although Exchange is not NUMA aware, Windows 2012R2 and 2016 are, though performance testing has shown no benefit to enabling vNUMA. Because Exchange does not benefit from vNUMA or CPU hot addition, VMware recommends against enabling CPU hot-add for Exchange Server virtual machines.

Virtual cores per virtual socket

When configuring a virtual machine, you can set both the number of virtual sockets (vSockets) and the number of virtual cores (vCores) per socket. VMware recommends setting the number of cores per socket to 1 and increasing the number of virtual sockets when allocating CPU resources for Exchange Server virtual machines.

High-performance VMware ESX policy

The Cisco HyperFlex system sets the ESX power policy to High Performance. For storage controller performance, this setting should remain set to High Performance.

CPU and memory reservation

Exchange Server is computation intensive, so you should use care to help ensure that the hypervisor, operating system, and Exchange Server are not constantly battling over computing and memory resources. Taking the precaution of setting memory and CPU reservation on important Exchange Server virtual machines can protect these virtual machines from unanticipated resource consumption by smaller overprovisioned virtual machines. When performance is the primary goal for a virtual machine, set memory reservation equal to the provisioned memory, and reserve at least one CPU core's worth of megacycles in the virtual machine's resource allocation settings. Microsoft does not support the overcommitting of memory resources for Exchange. Although CPU can be overcommitted to a maximum ratio of 2:1, Microsoft's best practice is to not overcommit CPU.

Memory oversubscription and dynamic memory allocation

Memory oversubscription is the allocation of more memory to a virtual machine than exists in the ESX host. Dynamic memory allocation is the hot addition of memory to a virtual machine while it is still running. Although neither process is harmful, neither is supported by Microsoft for Exchange Server virtual machines. This lack of support is partly the result of the way that Exchange allocates memory at service startup, which changed in Exchange 2013, when the store process was rewritten in C#. Rather than using a dynamic process, all memory allocation is static and set at service startup.

The supported way to increase or reduce the amount of memory allocated to an Exchange virtual machine is to power off the virtual machine and make the change by editing the virtual machine settings directly.

VMware ESX configuration

This section presents some best practices for configuring VMware ESX.

VMware vSphere Storage I/O Control

You can use VMware vSphere Storage I/O Control (SIOC) to prevent a noisy neighbor from consuming all the storage I/O space and starving other virtual machines in the cluster, which may require a fraction of the I/O space as a percentage of the total cluster IOPS. Unfortunately, SIOC does not differentiate based on I/O size. However, it can be enabled per data store and configured with a latency threshold that, when reached, applies a simple quality-of-service (QoS) policy across the data store I/O processing.

VMware ESX memory sharing across virtual machines

By default, inter–virtual machine page sharing is disabled due to security concerns, even though it can provide a greater benefit than just intra–virtual machine page sharing. You can think of this setting as memory deduplication to free memory space. This setting is discussed in great detail in the VMware vSphere Resource Management white paper in the section “Sharing memory across virtual machines.”

VMware ESX reqCallThreshold value

By default, in ESX the reqCallThreshold value is 8, so that the I/O in the virtual host bus adapter (vHBA) queue won’t flush to a lower layer until the threshold is reached. Some latency-sensitive databases with VMware Version 11 hardware experience improved latency when this value is lowered. This value can be lowered in the VMware VMX file of the virtual machine or globally in ESX. For exact settings, see the VMware white paper Performance Best Practices for VMware vSphere 6.0.

VMware Distributed Resource Scheduler anti-affinity rules

When using high-availability features such as DAGs in Exchange Server, be sure to configure anti-affinity rules in VMware Distributed Resource Scheduler (DRS) to indicate that virtual machines containing copies of the same database should be kept apart on different physical hosts. If both virtual machines in a two-copy DAG are homed to the same physical server, if that server fails all copies of the data will be moved with VMware HA to a new host, but an Exchange Server database outage will occur. With anti-affinity rules, VMware will work to isolate the virtual machines to separate hosts, and a host outage will not affect service availability because Exchange Server will activate the passive copy of the databases.

Anti-affinity rules can also help balance a cluster when you are deploying larger virtual machines that consume a large percentage of the physical server space. For example, if you are deploying three large Exchange Server virtual machines that each consume 40 percent of the memory or computing cycles on an ESX host, you can use anti-affinity rules to help ensure that, in most cases, no more than one of these virtual machines is ever on a single host. This configuration enables, for example, an SLA that requires that less than 75 percent of the host resources be used in a production environment so there will be enough resources available to handle a host failure and the subsequent failover of VMs to surviving hosts, without any impact on important services.

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 the 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 computing-only nodes.

The Exchange Server best practices discussed in this document are guidelines for high-performance (greater than 500 IOPS) virtual machines. Most virtual machines containing Exchange Server will work fine without the need for you to worry about a multitude of settings at the storage, ESX, virtual machine, or Exchange Server layers.

Unlike with many traditional storage systems, you can easily increase the VMDK size or even the data store. Veeam integration with Cisco HyperFlex systems reduces the impact on the cluster by eliminating the need for delta-disk consolidation and enables additional recovery functions with VSS application-consistent snapshots. Native cloning is space efficient and fast and provides Exchange Server administrators with quick access to production data for testing and for optimization without requiring additional storage capacity.

For more information

- VMware support for Microsoft Cluster Service and Windows Server failover clustering with shared storage: https://kb.vmware.com/kb/2147661
- Microsoft Exchange Server licensing: https://products.office.com/en-us/exchange/microsoft-exchange-server-licensing-licensing-overview
- VMware vSphere 6.0 performance best practices: http://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/vmware-perfbest-practices-vsphere6-0-white-paper.pdf
- VMware and Microsoft Exchange Server best practices: https://blogs.vmware.com/apps/files/2016/07/Microsoft_Exchange_Server_2016_on_VMware_vSphere_Best_Practices_Guide.pdf
- Cisco HyperFlex white paper: http://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/white-paper-c11-736814.pdf
- Microsoft Exchange 2016 core and memory guidance: https://blogs.technet.microsoft.com/exchange/2015/10/15/ask-the-perf-guy-sizing-exchange-2016-deployments/
- VMware vSphere resource management white paper: http://pubs.vmware.com/vsphere-60/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-601-resource-management-guide.pdf
- Microsoft Exchange Server Role Requirements Calculator: http://aka.ms/E2016Calc

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

Printed in USA C11-739273-00 07/17