Microsoft Exchange Server on VMware vSphere
Best Practices Guide

VMware, Inc.
3401 Hillview Ave
Palo Alto, CA 94304
www.vmware.com

Table of Contents

1.   Introduction

Microsoft Exchange Server is the dominant enterprise-class electronic messaging and collaboration applications in the industry today. Given the multitude of technical and operational enhancements in the latest released version of Microsoft Exchange Server (2016), customers are expected to continue using Exchange Server, which should retain its dominant position in the enterprise.

While VMware has historically supported Exchange Server virtualization, Exchange Server 2016 is the first version in which all the components and roles are officially supported for virtualization. See *Exchange Server 2016* (https://technet.microsoft.com/en-us/library/mt170645%28v=exchg.160%29.aspx).

Concurrent usage of the Exchange Server 2016 native high availability feature (Database Availability Group or DAG) with VMware vSphere® native high availability features is also supported. Because the vSphere hypervisor is part of the Microsoft SVVP program, virtualizing an Exchange Server 2016 instance on vSphere is fully supported (see https://www.windowsservercatalog.com/results.aspx?&bCatID=1521&cpID=11779&avc=0&ava=0&avg=0&OR=1&PGS=25&ready=0).

This document provides technical guidance for VMware customers who are considering virtualizing their Exchange Server on the vSphere virtualization platform.

Enterprise communication and collaboration is now so integral to an organization's operations that applications such as Exchange Server are now routinely classified as mission-critical. Organizations expect measurable and optimal performance, scalability, reliability, and recoverability from this class of applications. The main objective of this guide is to provide the information required to help a customer satisfy the operational requirements of running an Exchange Server 2016 on vSphere.

1.1   Purpose

This guide provides best practice guidelines for deploying Exchange Server 2016 on vSphere. The recommendations in this guide are not specific to any particular hardware nor to the size and scope of any particular Exchange implementation. The examples and considerations in this document provide guidance but do not represent strict design requirements, because the flexibility of Exchange Server 2016 on vSphere allows for a wide variety of valid configurations.

1.2   Target Audience

This guide assumes a basic knowledge and understanding of vSphere and Exchange Server 2016.

- Architectural staff can use this document to gain an understanding of how the system will work as a whole as they design and implement various components.

- Engineers and administrators can use this document as a catalog of technical capabilities.

- Messaging staff can use this document to gain an understanding of how Exchange might fit into a virtual infrastructure.

- Management staff and process owners can use this document to help model business processes to take advantage of the savings and operational efficiencies achieved with virtualization.

1.3 Scope

The scope of this document is limited to the following topics:

- VMware ESXi™ Host Best Practices for Exchange – Best practice guidelines for preparing the vSphere platform for running Exchange Server 2016. Guidance is included for CPU, memory, storage, and networking.

- Using VMware vSphere vMotion®, VMware vSphere Distributed Resource Scheduler™ (DRS), and VMware vSphere High Availability (HA) with Exchange Server 2016 – Overview of vSphere vMotion, vSphere HA, and DRS, and guidance for usage of these vSphere features with Exchange Server 2016 virtual machines.

- Exchange Performance on vSphere – Background information on Exchange Server performance in a virtual machine. This section also provides information on official VMware partner testing and guidelines for conducting and measuring internal performance tests.

- VMware Enhancements for Deployment and Operations – Brief look at vSphere features and add-ons that enhance deployment and management of Exchange Server 2016.

The following topics are out of scope for this document but are addressed in other documentation in the *Microsoft Exchange Server 2016 on VMware Solution Sales Enablement Toolkit*.

- Design and Sizing Guidance – This information is available in the *Microsoft Exchange Server 2016 on VMware Design and Sizing Guide.* This document details the capacity planning process and provides sizing examples for split-role, multi-role, and real-world customer configurations.

- Availability and Recovery Options – This document briefly covers VMware features that can enhance availability and recovery. An in-depth discussion of this subject is included in *Microsoft Exchange Server 2016 on VMware Availability and Recovery Options*.

This and other guides are limited in focus to deploying Exchange on vSphere. Exchange deployments cover a wide subject area, and Exchange-specific design principles should always follow Microsoft guidelines for best results.

1.4 External References

This document includes references to external links on third-party websites for the purposes of clarifying statements, where necessary. The statements are accurate as at the time of this writing. However, these third-party websites are not under the control of VMware, and the content available at those sites might change.

2.   ESXi Host Best Practices for Exchange

A well-designed VMware vSphere hypervisor platform is crucial to the successful implementation of virtualized enterprise applications such as Exchange Server. The following sections outline general best practices for designing vSphere for Exchange Server 2016.

2.1   CPU Configuration Guidelines

The latest release of vSphere (vSphere 6.0) has dramatically increased the scalability of virtual machines, enabling configurations of up to 128 virtual processors for a single virtual machine. With this increase, one option to increase performance is simply to create larger virtual machines. But additional factors are involved in deciding how much processing power goes into a virtual machine. This section reviews features that are available in vSphere with regard to virtualizing CPUs. Where relevant, this document discusses the impact of those features to Exchange Server 2016 and the recommended practices for using those features.

2.1.1 Physical and Virtual CPUs

VMware uses the terms virtual CPU (vCPU) and physical CPU (pCPU) to distinguish between the processors within the virtual machine and the underlying physical processor cores. Virtual machines with more than one virtual CPUs are also called symmetric multiprocessing (SMP) virtual machines. The virtual machine monitor (VMM) is responsible for virtualizing the CPUs. When a virtual machine begins running, control transfers to the VMM, which is responsible for virtualizing guest operating system instructions.

2.1.2 Architectural Limitations in Exchange Server

Microsoft provides guidelines to calculate the required compute resources for a single instance of Exchange Server (as an application) so that Exchange Servers do not experience unintended performance degradation due to incorrect sizing. These maximums are the same whether the Exchange Server is virtualized or installed on physical servers.

See the following table.

Table 1. Exchange Server Maximum Supported Compute Resource

Configuration Item	Maximum Supported
Memory Per Exchange Server Instance	96 GB
Number of CPUs per Exchange Server Instance	24

2.1.3 vSphere Virtual Symmetric Multiprocessing

VMware Virtual Symmetric Multiprocessing enhances virtual machine performance by enabling a single virtual machine to use multiple physical processor cores simultaneously. The most recent version of vSphere (version 6.1 as of the time of this writing) supports allocating up to 128 virtual CPUs per virtual machine. The biggest advantage of an SMP system is the ability to use multiple processors to execute multiple tasks concurrently, thereby increasing throughput (for example, the number of transactions per second). Only workloads that support parallelization (including multiple processes or multiple threads that can run in parallel) can benefit from SMP.

Be aware of the maximum 24-vCPU requirement for Exchange Server 2016 when making your sizing decision. The fact that you could allocate up to 128 vCPUs to a VM should of less importance in this

consideration. VMware strongly recommends that you allocate resources to a VM based on the actual needs of the applications hosted on the VM.

The ESXi scheduler uses a mechanism called *relaxed co-scheduling* to schedule processors. *Strict co-scheduling* requires all vCPUs to be scheduled on physical cores simultaneously, whereas relaxed co-scheduling monitors time skew between vCPUs to make scheduling or co-stopping decisions. A leading vCPU might decide to co-stop itself to allow for a lagging vCPU to catch up. Consider the following points when using multiple vCPUs:

- Virtual machines with multiple vCPUs perform well in the latest versions of vSphere, as compared with older versions where strict co-scheduling was used.

- Regardless of relaxed co-scheduling, the ESXi scheduler prefers to schedule vCPUs together, when possible, to keep them in synch. Deploying virtual machines with multiple vCPUs that are not used wastes resources and might result in reduced performance of other virtual machines.

  For detailed information regarding the CPU scheduler and considerations for optimal vCPU allocation, please see the section on ESXi CPU considerations in *Performance Best Practices for VMware vSphere 6.0 (*http://www.vmware.com/files/pdf/techpaper/VMware-PerfBest-Practices-vSphere6-0.pdf?vmw_so_vex_mande_12%20-%20G6.1006951).

- VMware recommends allocating multiple vCPUs to a virtual machine only if the anticipated Exchange workload can truly take advantage of all the vCPUs.

- Use the Microsoft-provided Exchange Server role calculator sizing tool to aid in your sizing exercise (http://blogs.technet.com/b/exchange/archive/2013/05/14/released-exchange-2013-server-role-requirements-calculator.aspx).

  Note the following:

  o The Exchange calculator is intentionally generous in its recommendation and limits. The recommendations might not be optimal for a virtualized workload.

  o The calculator does not factor in the Non-Uniform Memory Access (NUMA) topology of a given hardware when making compute resource recommendations. While Exchange Server (as an application) is unaware of NUMA optimization, VMware still recommends sizing a VM with the physical NUMA topology in mind. See Section 2.1.7, Non-Uniform Memory Access.

  o The calculator assumes a 10% hypervisor overhead in its computation. Although VMware testing indicates a variation of 3%-5% in a worst-case performance scenario, VMware recommends not changing this value in the calculator when modelling Exchange Server 2016 VMs for capacity. Given the relative age of Exchange Server 2016, the true impact of the hypervisor on Exchange Server 2016 is currently unknown. Leaving this value unchanged helps customers remain as compliant as possible with the Microsoft requirements.

- If the exact workload is not known, size the virtual machine with a smaller number of vCPUs initially, and increase the number later if necessary.

- Microsoft supports up to 2:1 virtual-to-physical CPU allocation for Exchange Server 2016 in a virtual environment. VMware recommends that, for the initial sizing of performance-critical Exchange virtual machines (production systems), the total number of vCPUs assigned to all the virtual machines be no more than the total number of *physical* cores on the ESXi host machine, not *hyper-threaded* cores. By following this guideline, you can gauge performance and utilization within the environment until you are able to identify potential excess capacity that could be used for additional workloads.

- Although larger virtual machines are possible in vSphere, VMware recommends reducing the number of virtual CPUs for a VM if monitoring of the actual workload shows that the Exchange application is not benefitting from the increased virtual CPUs.

2.1.4 CPU Reservations

Setting a CPU reservation sets a guaranteed CPU allocation for the virtual machine. This practice is generally not recommended because reserved resources are not available to other virtual machines, and flexibility is often required to manage changing workloads. However, SLAs and multitenancy might require a guaranteed amount of compute resource to be available. In these cases, reservations can be used to meet these requirements. VMware has conducted tests on virtual CPU over-commitment with SAP and SQL, showing that the performance degradation inside the virtual machines is linearly reciprocal to the over-commitment. Because the performance degradation is *graceful*, any virtual CPU over-commitment can be effectively managed by using DRS and vSphere vMotion to move virtual machines to other ESXi hosts to obtain more processing power.

2.1.5 Virtual Cores and Virtual Sockets

vSphere now supports configuration of the number of virtual cores per virtual socket in the VMware vSphere Web Client. This feature provides two functions:

- When used with virtual Non-Uniform Memory Access (vNUMA)-enabled virtual machines, this setting can be used to present specific NUMA topologies to the guest operating system.

- More commonly, this feature allows a guest operating system to utilize all of its assigned vCPUs in the case of an operating system that is limited to a certain number of CPUs.

On vSphere, virtual CPUs can be allocated to a virtual machine by socket, or by number of cores per socket. Controlling vCPUs by number of cores can help you overcome the limitations in previous versions of the Windows operating system. As an example, Windows Server 2008 R2 Standard Edition is limited to seeing four CPUs. In a physical environment, physical servers with more than four sockets utilize only the total cores in four of those sockets. In previous versions of vSphere, configuring a Windows Server 2008 R2 Standard virtual machine with more than four vCPUs resulted in the virtual machine seeing only four vCPUs. This was configurable within the virtual machine VMX file and is now configurable in the GUI. By configuring multiple cores per socket, the guest operating system can see and utilize all configured vCPUs.

While controlling vCPU allocation through cores per socket helps overcome this limitation, the configuration also impedes the efficiency of the vSphere CPU scheduler's resource optimization algorithms. The configuration can also present an incorrect NUMA topology to the underlying operating system, leading to performance penalty and impaired VM mobility within an ESXi cluster.

Because Exchange Server 2016 requires a minimum OS version of Windows Server 2012 and the operating system does not suffer from the same socket limitations of prior Windows versions, VMware recommends that customers allocated virtual CPUs to an Exchange Server 2016 VM using the "Socket" option. VMware recommends that customers leave the cores per socket count at one at all times.

The following figure shows the configuration interface.

Figure 1. New Virtual Machine CPU Configuration
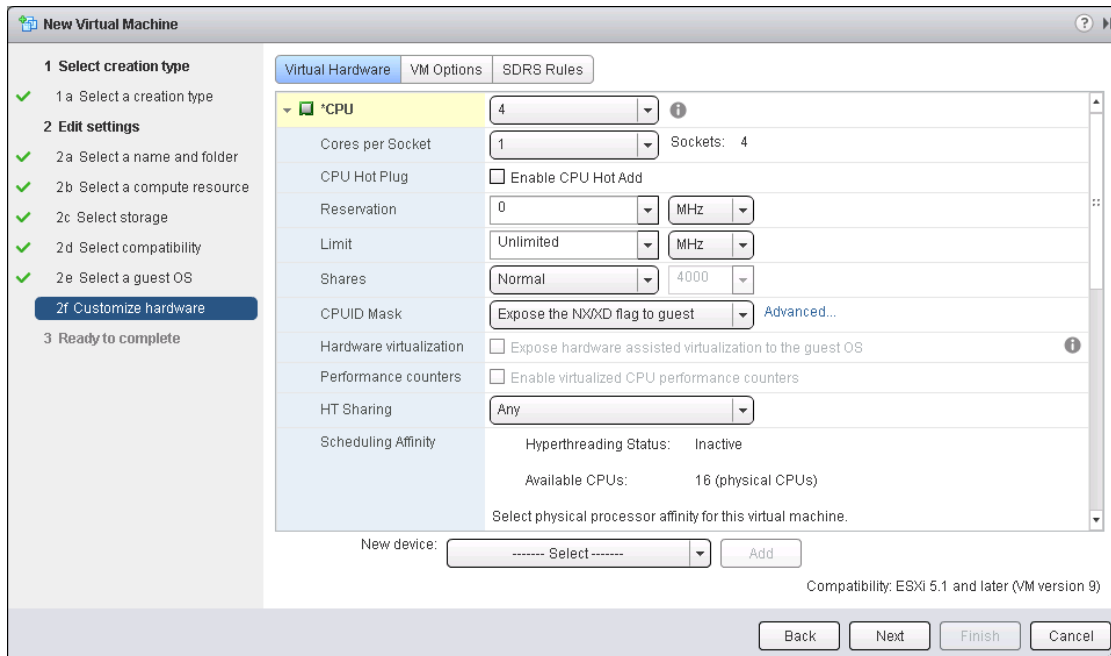
Virtual machines, including those running Exchange Server 2016, should be configured with multiple virtual sockets/CPUs. Use the **Cores per Socket** option only if the guest operating system requires the change to make all vCPUs visible. Virtual machines using vNUMA might benefit from this option, but the recommendation for these virtual machines is to use virtual sockets (CPUs in the web client). Although Exchange Server 2016 is not a NUMA-aware application and performance tests have not shown any significant performance improvements by enabling vNUMA, the Windows Server 2012 OS is NUMA-aware and Exchange Server 2016 (as an application) does not experience any performance, reliability, or stability issue attributable to vNUMA.

2.1.6 Hyper-threading

Hyper-threading technology (recent versions are called symmetric multithreading, or SMT) allows a single physical processor core to behave like two logical processors, so that two independent threads are able to run simultaneously. Unlike having twice as many processor cores that can roughly double performance, Hyper-threading can provide anywhere from a slight to a significant increase in system performance by keeping the processor pipeline busier. For example, an ESXi host system enabled for SMT on an 8-core server sees 16 threads that appear as 16 logical processors.

Previous guidance provided by Microsoft regarding Exchange sizing and the use of Hyper-threading led to some confusion among those looking at virtualizing Exchange Server. Microsoft has since updated all the applicable documents to clarify that statements relating to Hyper-threading and Exchange Server do not apply to the virtualization platforms. The Microsoft guidance in this respect is expected to complement Microsoft's "Preferred Architecture" design option, which does not incorporate any virtualization design choices, options or considerations. See *Ask The Perf Guy: What's The Story With Hyper-threading and Virtualization?* for the most recent guidance from Microsoft (http://blogs.technet.com/b/exchange/archive/2015/09/15/ask-the-perf-guy-what-s-the-story-with-Hyper-threading-and-virtualization.aspx).

vSphere uses hyper-threads to provide more scheduling choices for the hypervisor. Hyper-threads provide additional targets for *worlds*, a schedulable CPU context that can include a virtual CPU or

hypervisor management process. For workloads that are not CPU bound, scheduling multiple vCPUs onto a physical core's logical cores can provide increased throughput by increasing the work in the pipeline. The CPU scheduler schedules to a whole core over a hyper-thread, or partial core, if CPU time is lost due to hyper-thread contention. Consequently, VMware recommends enabling Hyper-threading on the ESXi host if the underlying hardware supports the configuration.

2.1.7 Non-Uniform Memory Access

In Non-Uniform Memory Access (NUMA) systems, a processor or set of processor cores have memory that they can access with very little latency. The memory and its associated processor or processor cores are referred to as a NUMA node. Operating systems and applications designed to be NUMA-aware can make decisions as to where a process might run relative to the NUMA architecture. This allows processes to access memory local to the NUMA node rather than having to traverse an interconnect, incurring additional latency. Exchange Server 2016 is not NUMA-aware, but both ESXi and Windows Server 2012 are NUMA-aware.

vSphere ESXi provides mechanisms for letting virtual machines take advantage of NUMA. The first mechanism is transparently managed by ESXi while it schedules a virtual machine's virtual CPUs on NUMA nodes. By attempting to keep all of a virtual machine's virtual CPUs scheduled on a single NUMA node, memory access can remain local. For this to work effectively, size the virtual machine to fit within a single NUMA node. This placement is not a guarantee because the scheduler migrates a virtual machine between NUMA nodes based on the demand.

The second mechanism for providing virtual machines with NUMA capabilities is vNUMA. When enabled for vNUMA, a virtual machine is presented with the NUMA architecture of the underlying hardware. This allows NUMA-aware operating systems and applications to make intelligent decisions based on the underlying host's capabilities. By default, vNUMA is automatically enabled for virtual machines with nine or more vCPUs on vSphere. Because Exchange Server 2016 is not NUMA aware, enabling vNUMA for an Exchange virtual machine does not provide any additional performance benefit nor does doing so induce any performance degradation.

Consider sizing Exchange Server 2016 virtual machines to fit within the size of the physical NUMA node for best performance. The following figure depicts an ESXi host with two NUMA nodes, each comprising four physical cores and 32GB of memory. The virtual machine allocated with four vCPUs and 24GB of memory can be scheduled by ESXi onto a single NUMA node.

Figure 2. NUMA Architecture Sizing
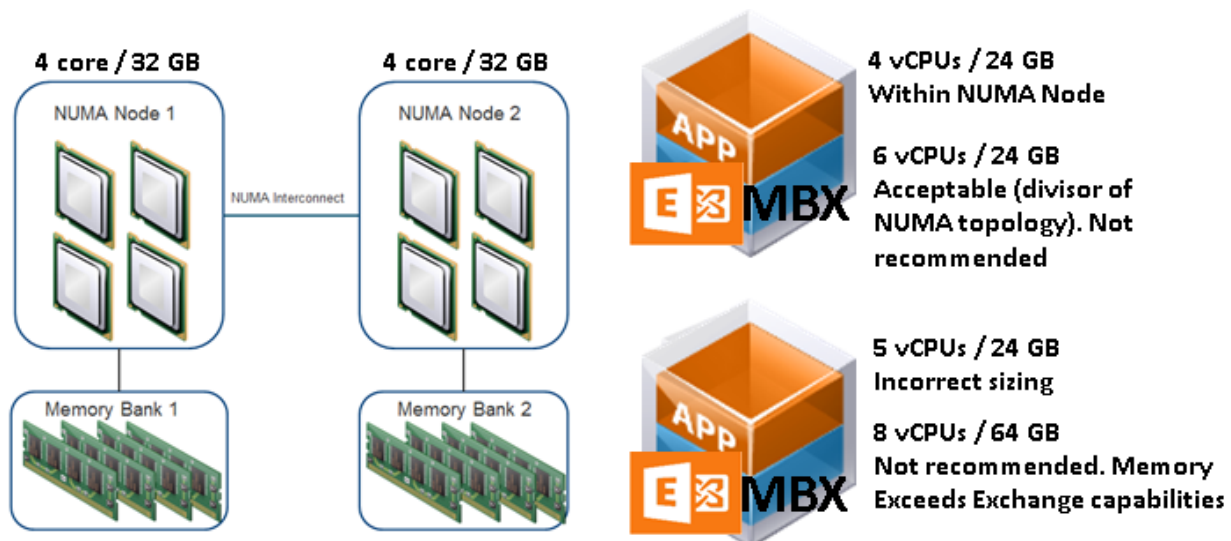
A virtual machine allocated with 6 vCPUs and 24 GB of memory must span NUMA nodes and might incur some memory access latency. The associated latency can be minimized or avoided through the use of the appropriate combination of vNUMA control options in the VM's Advanced Configuration options. See "Specifying NUMA Control" (http://pubs.vmware.com/vsphere-60/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-601-resource-management-guide.pdf#_OPENTOPIC_TOC_PROCESSING_d172e22169) in the VMware vSphere Resource Management Guide.

While a virtual machine allocated with 8 vCPUs and 64GB can evenly span NUMA nodes and not incur the memory access latency issues described earlier, such configuration is neither recommended nor supported because the allocated memory size exceeds Microsoft's recommendations for Exchange Server 2016 sizing.

For large environments, you can test each configuration to determine whether additional latency associated with remote memory addressing warrants creating additional, smaller virtual machines.

Verify that all ESXi hosts have NUMA enabled in the system BIOS. In some systems NUMA is enabled by disabling node interleaving.

2.1.8 vNUMA and CPU Hot Plug

Enabling CPU hot add for a VM on vSphere disables vNUMA for the VM. Because Exchange Server does not benefit for either vNUMA or CPU hot add, VMware recommends that you do not enable CPU hot add for an Exchange Server 2016 VM.

2.2 Memory Configuration Guidelines

This section provides guidelines for memory allocation to Exchange Server 2016 virtual machines. These guidelines consider vSphere memory overhead and the virtual machine memory settings.

2.2.1 ESXi Memory Management Concepts

vSphere virtualizes guest physical memory by adding an extra level of address translation. Shadow page tables make it possible to provide this additional translation with little or no overhead. Managing memory in the hypervisor enables the following:

- Memory sharing across virtual machines that have similar data – Same guest operating systems.

- Memory over-commitment – Allocating more memory to virtual machines than is physically available on the ESXi host.

- A memory balloon technique – Virtual machines that do not need all the memory they have been allocated give memory to virtual machines that require additional allocated memory.

For a more detailed discussion of vSphere memory management concepts, see the "Memory Virtualization Basics" and "Administering Memory Resources" sections in the *vSphere Resource Management* guide (http://pubs.vmware.com/vsphere-60/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-601-resource-management-guide.pdf) *for more detailed information.*

2.2.2 Virtual Machine Memory Concepts

The following figure illustrates the use of memory settings parameters in the virtual machine.

Figure 3. Virtual Machine Memory Settings
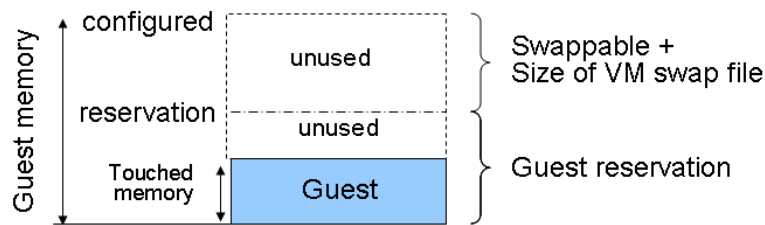
The vSphere memory settings for a virtual machine include the following parameters:

- Configured memory – Memory size of virtual machine assigned at creation.

- Touched memory – Memory used by the virtual machine. vSphere allocates guest operating system memory on demand.

- Swappable – Virtual machine memory that can be reclaimed by the balloon driver or by vSphere swapping. Ballooning occurs before vSphere swapping. If this memory is in use by the virtual machine (touched and in use), the balloon driver causes the guest operating system to swap. Also, this value is the size of the per virtual machine swap file that is created on the VMware vSphere Virtual Machine File System (VMFS) file system.

- If the balloon driver is unable to reclaim memory quickly enough or is disabled or not installed, vSphere forcibly reclaims memory from the virtual machine using the VMkernel swap file.

2.2.3 Allocating Memory to Exchange Virtual Machines

Microsoft has developed a thorough sizing methodology for Exchange Server that has matured with recent versions of Exchange Server. VMware recommends using the memory sizing guidelines set by Microsoft. The amount of memory required for an Exchange Server 2016 is driven by the expected amount of loads that will be generated AND the total number of Exchange Servers that will be supporting the loads in the environment. Load characteristics include (but are not limited to) the following:

- Number of mailbox users

- Profile of mailbox usage pattern (size and number of emails sent and received)

- Type and number of devices used for accessing emails

- Anti-virus and other messaging security and hygiene solutions deployed on the server

- Type and frequency of backup solution in use

- High availability and resilience requirements

As Exchange Servers are memory-intensive, and performance is a key factor, such as in production environments, VMware recommends the following practices:

- Do not overcommit memory on ESXi hosts running Exchange workloads. If memory over-commitment cannot be avoided, use vSphere memory allocation options to guarantee required memory size to the Exchange Server, or to limit memory access for other, non-essential VMs in the vSphere cluster.

- For production systems, it is possible to achieve this objective by setting a memory reservation to the configured size of the Exchange Server virtual machine.

    Note the following:

    o Setting memory reservations might limit vSphere vMotion. A virtual machine can be migrated only if the target ESXi host has free physical memory equal to or greater than the size of the reservation.

- o Setting the memory reservation to the configured size of the virtual machine results in a per-virtual machine VMkernel swap file of near-zero bytes that consumes less storage and eliminates ESXi host-level swapping. The guest operating system within the virtual machine still requires its own page file.

- o Reservations are recommended only when it is possible that memory might become overcommitted on hosts running Exchange virtual machines, when SLAs dictate that memory be guaranteed, or when there is a desire to reclaim space used by a virtual machine swap file.

- o There is a slight, appreciable performance benefit to enabling memory reservation even if you do not expect memory to be overcommitted in the vSphere cluster.

- It is important to right-size the configured memory of a virtual machine. This might be difficult to determine in an Exchange environment because the Exchange JET cache is allocated based on memory present during service start-up. Understand the expected mailbox profile and recommended mailbox cache allocation to determine the best starting point for memory allocation.

- Do not disable the balloon driver (which is installed with VMware Tools™) or any other ESXi memory management mechanism.

  Note the following:

  - o Transparent Page Sharing (TPS): Page sharing enables ESXi hosts to more efficiently utilize it available physical memory to support more workloads. TPS is useful in scenarios where multiple VM siblings share the same characteristics (such as same OS and applications). In this configuration, vSphere is able to avoid redundancy by sharing similar pages among the different Exchange Server VMs. This sharing is transparent to the applications and processes inside the VM.

  - o For security reasons, inter-VM page sharing is disabled by default on current versions of vSphere. While a VM continues to benefit from TPS in this configuration (the VM is able to share pages internally among its own processes and components), a greater benefit can be realized by enabling inter-VM page sharing. See "Sharing Memory Across Virtual Machine" in the *vSphere Resource Management* guide (http://pubs.vmware.com/vsphere-60/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-601-resource-management-guide.pdf).

Enable DRS to balance workloads in the ESXi host cluster. DRS and reservations can give critical workloads the resources they require to operate optimally. More recommendations for using DRS with Exchange Server 2016 are available in the *"Using vSphere Technologies with Exchange Server 2016"* section below.

2.2.4 Memory Hot Add, Over-subscription, and Dynamic Memory

vSphere exposes the ability to add memory to a VM while the VM is powered on. Modern operating systems (including Windows) support this feature and are able to instantaneously detect and use the hot-added memory.

While vSphere also enables the ability to hot remove memory from a VM, this capability is not supported by current versions of Windows, so it is unavailable to a Windows VM on vSphere.

vSphere enables an administrator to allocate virtual memory to VMs beyond the physically available memory size of the ESXi host. This condition is called over-allocation, or over-subscription. Over-subscription is possible and non-intrusive and it is an essential core benefit of virtualization, because repeated testing has shown that VMs do not all fully utilize their allocated resources at the same time. If all VMs request their resources at the same time at any point in time, this creates resource over-commitment on the ESXi or cluster.

Transient resource over-commitment is possible within a virtual environment. Frequent or sustained occurrence of such incidents is problematic for critical applications such as Exchange Server.

Dynamic memory is a Microsoft Hyper-V construct that does not have a direct equivalence on vSphere. Even in an over-commitment scenario, the VM on vSphere is never induced to believe that its allocated amount of memory has been physically reduced. vSphere uses other memory management techniques for arbitrating contentions during a resource over-commitment condition

In Exchange Server 2016, the JET cache is allocated based on the amount of memory available to the operating system at the time of service start-up. After being allocated, the JET cache is distributed among active and passive databases. With this model of memory pre-allocation for use by Exchange databases, adding memory to a running Exchange virtual machine provides no additional benefit unless the virtual machine was rebooted or Exchange services were restarted. Consequently, memory hot add is neither useable by nor beneficial to an Exchange Server VM and is, therefore neither recommended nor supported. In contrast, removing memory that JET had allocated for database consumption impacts performance of the store worker and indexing processes by increasing processing and storage I/O.

Microsoft support for the virtualization of Exchange Server 2016 states that the over-subscription and dynamic allocation of memory for Exchange virtual machines is not supported. For the avoidance of confusion, refer to the preceding paragraphs to understand why these requirements are not relevant to Exchange Servers virtualized on the vSphere platform.

Over-subscription is different from over-commitment. Over-subscription is benign and does not impact the VMs. Over-commitment is the adverse extension of over-subscription and should be avoided in all cases. However, if you expect that your ESXi cluster might occasionally experience resource contention as a result of memory over-commitment, VMware recommends that you judiciously reserve all of the memory that you have allocated to your Exchange Server VMs

Because ESXi does not support hot unplug (hot removal) of memory from a Windows VM, the only way (outside of unsupported programmatic interference) to reduce the amount of memory presented to a VM running Exchange Server 2016 is by powering off the VM and changing the memory allocation. When the VM powers on again, the OS will see the new memory size, and Exchange Server will re-allocate the available memory to its worker processes. This is not dynamic memory.

2.3   Storage Virtualization

VMFS is a cluster file system that provides storage virtualization optimized for virtual machines. Each virtual machine is encapsulated in a small set of files, and VMFS is the default storage system for these files on physical SCSI disks and partitions. VMware supports Fiber Channel, iSCSI, and network-attached storage (NAS) shared-storage protocols.

It is preferable to deploy virtual machine files on shared storage to take advantage of vSphere vMotion, vSphere HA, and DRS. This is considered a best practice for mission-critical Exchange Server deployments that are often installed on third-party, shared-storage management solutions.

VMware storage virtualization can be categorized into three pillars of storage technology, as illustrated in the following figure. The storage array is the physical storage pillar, comprising physical disks presented as logical storage volumes in the next pillar. Storage volumes, presented from physical storage, are formatted as VMFS datastores or with native file systems when mounted as raw device mappings. Virtual machines consist of virtual disks or raw device mappings that are presented to the guest operating system as SCSI disks that can be partitioned and formatted using any supported file system.

Figure 4. VMware Storage Virtualization
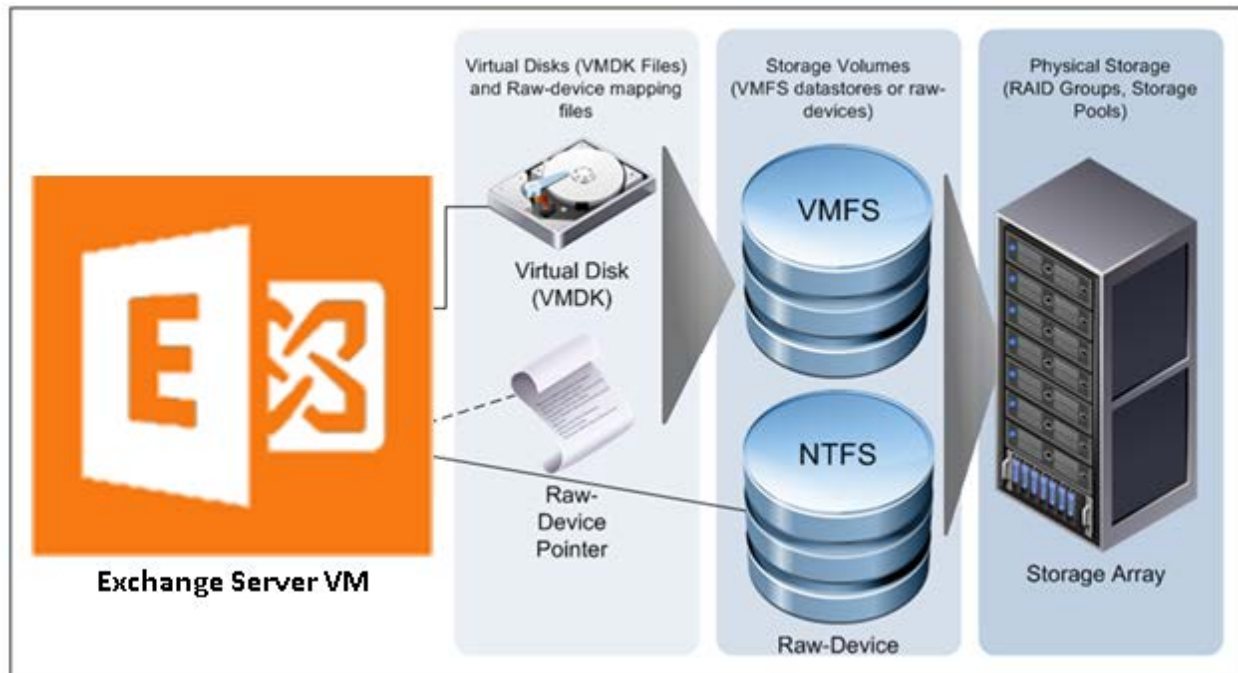
Exchange Server has improved significantly in the last couple of releases, and Exchange Server 2016 continues these improvements. The improvements have made Exchange Server less storage I/O intensive than before. This reality informs Microsoft's preference for commodity-class direct-attached storage (DAS) for Exchange Server. While the case for DAS and JBOD storage for Exchange appears reasonable from the /O perspective, the associated operational and administrative overhead for an enterprise-level production Exchange Server infrastructure do not justify the guidance.

To overcome the increased failure rate and shorter lifespan of commodity storage, Microsoft routinely recommends multiple copies of the Exchange data across larger sets of storage and Exchange Servers than operationally necessary.

While VMware supports the use of converged storage solutions for virtualizing Exchange Server on vSphere, VMware recommends that customers using such solutions (VMware Virtual SAN™) thoroughly benchmark and validate the suitability of such solutions and engage directly with the applicable vendors for configuration, sizing, performance and availability guidance.

Even with the reduced I/O requirements of an Exchange Server instance, without careful planning, storage access, availability and latencies can still manifest in an Exchange server infrastructure.

VMware recommends that you set up a minimum of four paths from an ESXi host to a storage array. To accomplish this, the host requires at least two host bus adapter (HBA) ports.

Figure 5. Storage Multi-pathing Requirements for vSphere
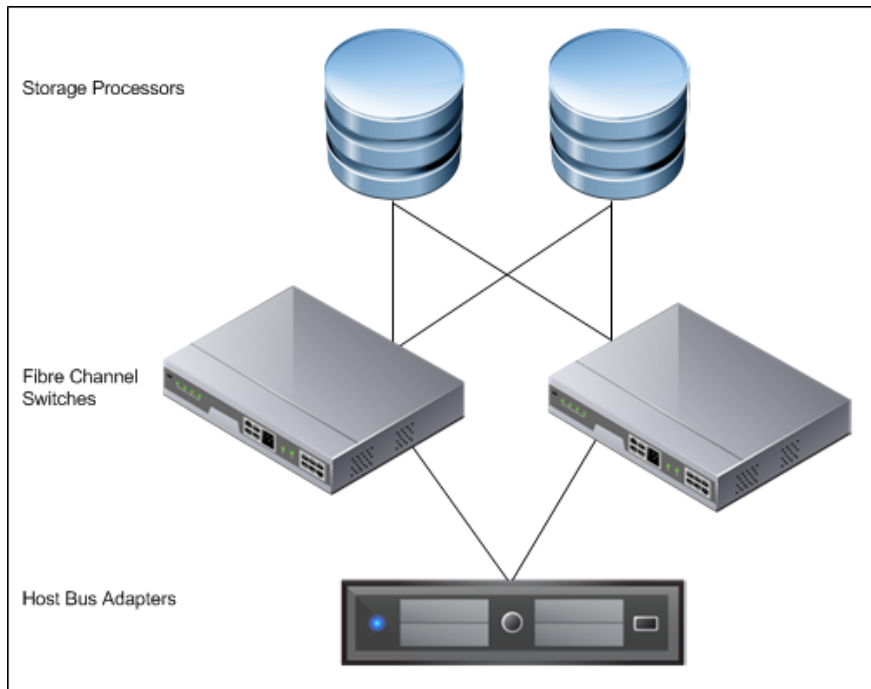
The terms used in the preceding figure are:

- HBA – A device that connects one or more peripheral units to a computer and manages data storage and I/O processing.

- Fibre Channel (FC) – A gigabit-speed networking technology used to build storage area networks (SANs) and to transmit data.

- Storage Processor (SP) – A SAN component that processes HBA requests routed through an FC switch and handles the RAID/volume functionality of the disk array.

2.3.1 Raw Device Mapping

VMFS also supports Raw Device Mapping (RDM). RDM allows a virtual machine to directly access a volume on the physical storage subsystem and can be used only with Fiber Channel or iSCSI. RDM provides a symbolic link or mount point from a VMFS volume to a raw volume. The mapping makes volumes appear as files in a VMFS volume. The mapping file, rather than the raw volume, is referenced in the virtual machine configuration. Connectivity from the virtual machine to the raw volume is direct, and all data is stored using the native file system, NTFS. In the case of a failure of the VMFS datastore holding the RDM mapping file, a new mapping file can be created. Access to the raw volume and its data is restored, and no data loss occurs.

There is no performance difference between a VMFS datastore and an RDM. The following are the only conditions that impose a requirement for RDM disks for VMs in current versions of vSphere:

- If the backup solution performs hardware-based VSS snapshots of VMs (or otherwise requires direct storage access).

- When the VM will be participating in a Windows Server Failover Clustering configuration that requires the clustered VMs to share the same disks.

Because Exchange Server clustering option does not require sharing disks among the nodes, the only scenario for RDM disks for a virtualized Exchange Server on vSphere is where the backup solution vendor requires such configuration.

The decision to use VMFS or RDM for Exchange data should be based on technical requirements. The following table summarizes the considerations when making a decision between the two.

Table 2. VMFS and Raw Disk Mapping Considerations for Exchange Server 2016

VMFS	RDM
• Volume can contain many virtual machine disk files, reducing management overhead. • Increases storage utilization and provides better flexibility and easier administration and management. • Supports existing and future vSphere storage virtualization features. • Fully supports VMware vCenter™ Site Recovery Manager™. • Supports the use of vSphere vMotion, vSphere HA, and DRS. • Supports VMFS volumes up to 64TB and virtual disks/VMDK files up to 2TB. • Has open virtual disk capacity of up to 25TB per host (8TB by default) – More information is available at http://kb.vmware.com/kb/1004424.	• Ideal if disks must be dedicated to a single virtual machine. • Requires more LUNs, making it is easier to reach the limit of 255 LUNs that might be presented to an ESXi host. • Might be required to leverage array-level backup and replication tools (VSS) integrated with Exchange databases. • Facilitates data migration between physical and virtual machines using the LUN swing method. • Fully supports vCenter Site Recovery Manager. • Supports vSphere vMotion, vSphere HA, and DRS. • Supports presenting volumes of up to 64TB to the guest operating system – Physical-mode RDM only.

2.3.2 In-Guest iSCSI and Network-Attached Storage

Similar to RDM, in-guest iSCSI initiator-attached LUNs provide dedicated storage to a virtual machine. Storage presented using in-guest iSCSI is formatted natively using NTFS within the Windows guest operating system and bypasses the storage management of the ESXi host. Presenting storage in this way requires that additional attention be provided to the networking infrastructure and configuration at the vSphere level and the physical level. Some of the benefits from using in-guest iSCSI attached storage include the ability to allocate more than 256 LUNs to virtual machines on a single ESXi host and retaining the ability to use array-level backup and replication tools.

Although VMware testing has shown that NAS-attached virtual disks perform well for Exchange workloads, Microsoft does not currently support accessing Exchange data (mailbox databases, transport queue, and logs) stored on network-attached storage. This includes accessing Exchange data using a UNC path from within the guest operating system, as well as virtual machines with VMDK files located on NFS-attached storage.

When using in-guest iSCSI to present storage to an Exchange Server VM, confirm that the iSCSI NIC is exempted from Windows Server failover clustering or any other non-storage-related processes or components. Similarly, VMware recommends that customers use jumbo frames or 10 GB networks to support such storage configuration option.

2.3.3   Virtual SCSI Adapters

VMware provides two commonly used virtual SCSI adapters for Windows Server 2012: LSI Logic SAS and VMware Paravirtual SCSI (PVSCSI). The default adapter when creating new virtual machines running Windows on vSphere is LSI Logic SAS, and this adapter can satisfy the requirements of most workloads with no additional drivers.

The Paravirtual SCSI adapter is a high-performance vSCSI adapter developed by VMware to provide optimal performance for virtualized business critical applications. The advantage of the PVSCSI adapter is that the added performance is delivered while minimizing the use of hypervisor CPU resources. This leads to lower hypervisor overhead required to run storage I/O-intensive applications.

Note   For environments running ESXi versions prior to 5.0 Update 1 consider using PVSCSI. Refer to *Windows 2008 R2 virtual machine using a paravirtual SCSI adapter reports the error: Operating system error 1117 encountered* (http://kb.vmware.com/kb/2004578).

Exchange Server 2016 has greatly reduced the amount of I/O generated to access mailbox data. However, storage latency is still a factor. In environments supporting thousands of users per Mailbox server, PVSCSI might prove beneficial. The decision on whether to use LSI Logic SAS or PVSCSI should be made based on Jetstress testing of the predicted workload using both adapters. Additionally, organizations must consider any management overhead an implementation of PVSCSI might introduce. Because many organizations have standardized on LSI Logic SAS, if the latency and throughput difference is negligible with the proposed configuration, the best option might be the one with the least impact to the current environment.

Virtual machines can be deployed with up to four virtual SCSI adapters. Each vSCSI adapter can accommodate up to 15 storage devices for a total of 60 storage devices per virtual machine. During allocation of storage devices, each device is assigned to a vSCSI adapter in sequential order. Not until a vSCSI adapter reaches its fifteenth device will a new vSCSI adapter be created. To provide better parallelism of storage I/O, equally distribute storage devices among the four available vSCSI adapters, as shown in the following figure.

Figure 6. Storage Distribution with Multiple vSCSI Adapters
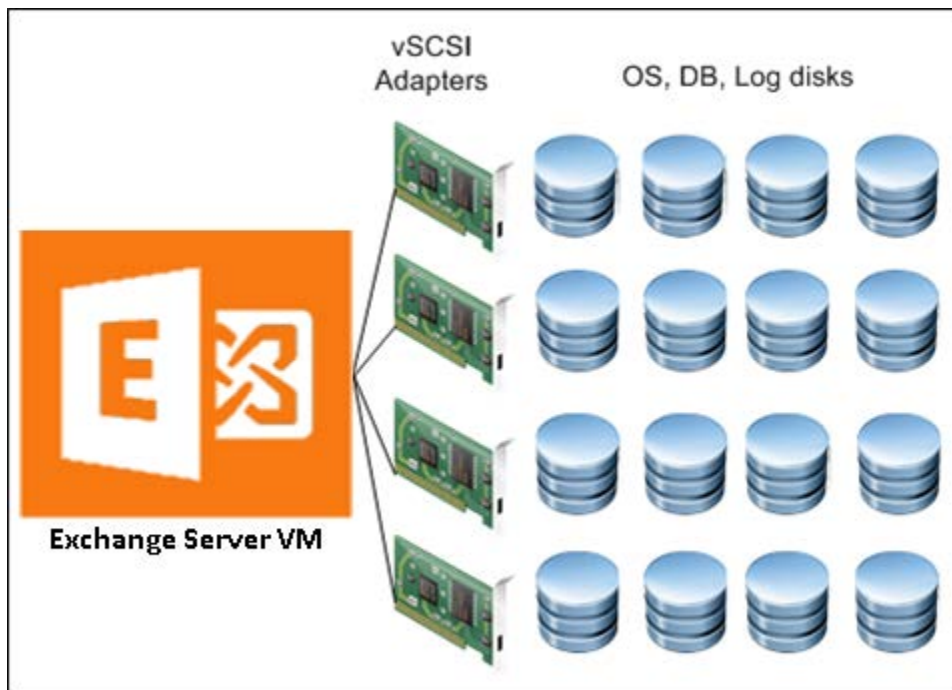
2.3.4 Virtual SCSI Queue Depth

Queue depth denotes the number of I/Os that can pass through a storage path at one time. All other I/Os beyond this number are queued until the path has more room to accommodate them. Because there are multiple points through which an I/O will pass from within a Guest Operating System before it reaches the physical storage device, customers should pay special attention to their storage configuration to avoid unnecessary and undesirable I/O queuing for their Exchange Server VMs.

When presenting storage to a VM, the virtual disk is connected to one virtual SCSI controller. Each SCSI controller has a finite (and configurable) queue depth, which varies among the virtual SCSI controllers available in vSphere.

The PVSCSI is the optimal SCSI controller for an I/O-intensive application on vSphere. This controller has queue depth of 64 by default (double the size of an LSI Logic SAS controller). The PVSCSI controller's queue depth can also be increased to 254, providing even more increased I/O bandwidth for the virtualized workload.

Because of these increased performance and throughput benefits, VMware recommends that customers choose PVSCSI as the virtual SCSI controllers for their Exchange Server data, databases and logs volumes.

Note the following:

- Using PVSCSI for Exchange Server and Exchange DAG is officially supported on vSphere 6.0, although such configurations have been in common use since vSphere 5.1

- The PVSCSI drivers is not native to the Windows, therefore customers using PVSCSI controllers must keep the VMware Tools instances on their VMs updated on a regular basis. VMware Tools is the vehicle for delivering the PVSCSI drivers to the OS.

Figure 7 Common Points of Storage IO Queues
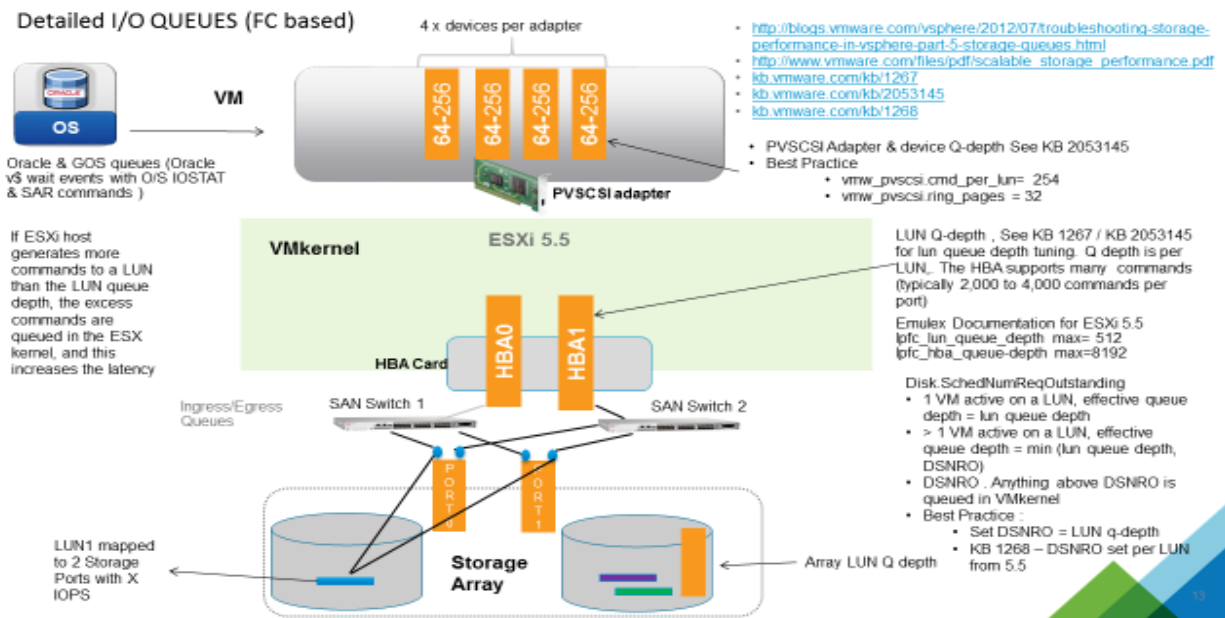
While increasing the default queue depth of a virtual SCSI controller can be beneficial to an Exchange Server VM, the configuration can also introduce unintended adverse effects in overall performance if not done properly. VMware highly recommends that customers consult and work with the appropriate storage vendor's support personnel to evaluate the impact of such changes and obtain recommendations for

other adjustments that may be required to support the increase in queue depth of a virtual SCSI controller. See *Large-scale workloads with intensive I/O patterns might require queue depths significantly greater than Paravirtual SCSI default values* (http://kb.vmware.com/kb/2053145) and *Changing the queue depth for QLogic, Emulex, and Brocade* HBAs (http://kb.vmware.com/kb/1267) for further information.

2.3.5 Exchange Server 2016 on All-Flash Storage Array

As all-flash storage is gaining increasing popularity in corporate data centers, many people are examining the value of deploying Exchange Server on all-flash storage arrays. Exchange Server is generally perceived as a workload which requires lots of capacity, but relatively less demanding on performance when compared to other database workloads. All-flash storage arrays are viewed as more expensive than the traditional storage arrays when it comes to physical capacity cost per GB. Why would customers be interested in an all-flash storage option for Exchange Server 2016?

While people are well aware of the performance benefits of all-flash storage, the latest generation of all-flash storage also offers:

- Built-in data services, such as, always-on thin provisioning, inline data deduplication, and inline data compression that provide compelling data reduction ratio.

- Flash optimized data protection that replaces traditional RAID methodologies can simplify the Exchange Server sizing and capacity planning efforts while minimizing protection overhead and performance penalty.

- Instant space efficient copies via VSS integration that significantly increases efficiency and operational agility for Database Availability Groups (DAGs) and can be used for local data protection and potentially replace lagged copies.

Exchange Server administrators are constantly facing a number of challenges when upgrading:

- The architecture for Exchange Server is rigid by nature; the design is intended to be built once and maintained with minor changes until decommission. Exchange Server admins are tasked with predicting projected mailbox growth and usage patterns for the upcoming 4- to 5-year period. Storage is over-provisioned to avoid the potentially costlier mistake of under-sizing. The nature of SSD allows modern all-flash vendors to support 100% storage on demand with always-on thin provisioning without any performance impacts. The initial acquisition cost of storage is significantly driven down.

- Exchange Server has long ago moved away from single instance storage. Exchange Server Continuous Replication (Database Availability Groups / DAG) consumes between 2x and 6x the capacity required to store the production copy of databases. Most companies report that capacity requirements have multiplied by a factor of 6x after migrating from Exchange Server 2003/2007 to version 2010/2013. The Exchange Server data that used to consume 12TB of space in an Exchange Server 2003 single copy cluster, now consumes 72TB in a three-copy DAG. With the right all-flash storage, initial database copies can be reduced in capacity. Passive DAG copies can be created in seconds via VSS integrated copy technology, and consume no additional space.

- Distribution Groups (DLs) is a de facto method for transmitting messages within organizations. Coupled with the fact that nearly 70% of all email messages contain attachments and the fact that every attachment is stored repeatedly in every DL member's inbox, mailboxes and mailbox databases are larger than ever before. There are massive opportunities to increase storage efficiency with all-flash storage solutions that offer inline data deduplication and compression.

- Many organizations are forced to run their Exchange Server in Online Mode for several reasons, including VDI, Security/Governance (cannot use OSTs), Workflow applications (cannot tolerate cached versions of mailbox items, and HIPPA regulations. Online Mode increases the IO requirements compared to Cached Mode by 270%. Performance can still be a critical consideration for Exchange Server deployment in many cases.

During a study of Exchange Server on EMC XtremIO, EMC found incredible efficiencies resulting in reducing the total disk required to manage an Exchange Server environment. The efficiencies generated significant opportunities to reduce costs over a five-year period. So much so that when compared to

alternative solutions such as a VNX solution or Microsoft's preferred architecture which included a direct attached storage solution, the XtremIO solution total cost of ownership (TCO) was fairly similar while offering tangibly improved performance and simpler storage management.

Figure 8 - Cost of Ownership Comparison
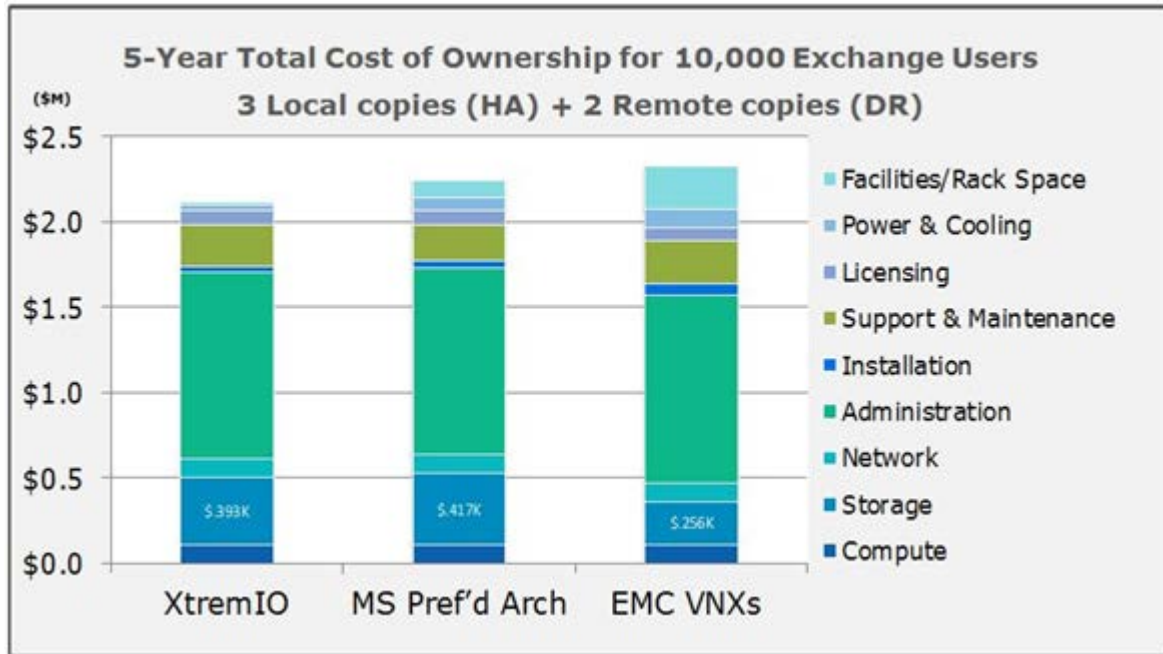
The study was conducted as a "straw man TCO" based on 10,000 seats, 2GB average mailbox size, with 150 messages sent/received per user, per day. The TCO includes all of the various aspects of installing, managing, cooling, supporting, and paying for facilities costs that you would typically find in any other TCO model. The "Comparative Costs of XtremIO" graph shows three Exchange 2010 implementations based on three different storage devices. All other aspects of the Exchange implementation are held as constant as possible. For example, same number of servers, Ethernet ports, admins, mailboxes, databases and database copies -- everything is the same except the storage and costs associated with the storage such as maintenance, installation, facilities costs, power and cooling. The result, all three storage configurations resulted in average mailbox costs that were within 25-cents of each other. And, based on the "deal" you might get from your hardware vendor, could actually be the same price. The prices used in this TCO are the typical "street" prices seen in the open marketplace. There are no special discounts, no list prices, no one-time offers. These are real prices of real configurations obtained in April of 2015.

The data reduction ratio for EMC XtremIO from the combinations of thin provisioning, deduplication, compression, and space efficient copies with XtremIO Virtual Copies was 7:1. The data reduction ratio increases when more mailboxes and DAG copies are placed onto the array, making XtremIO even more attractive for larger deployments.

Figure 9 - Data Reduction Ration on XtremIO

All-flash array is typically more expensive than the traditional storage array or DAS, once you get past the initial hardware acquisition cost and start to look at the efficiencies and operational agility aspects, it may offer attractive TCO, and can be a compelling storage platform for Exchange deployments.

http://www.emc.com/collateral/technical-documentation/h14246-xtremio-exchange-ra.pdf

https://community.emc.com/community/connect/everything_microsoft/blog/2015/05/20/preliminary-testing-of-exchange-on-xtremio

2.4 Networking Configuration Guidelines

This section covers design guidelines for the virtual networking environment and provides configuration examples at the ESXi host level for Exchange Server 2016 installations.

Note   The examples do not reflect design requirements and do not cover all possible Exchange network design scenarios.

2.4.1 Virtual Networking Concepts

The virtual networking layer comprises the virtual network devices through which virtual machines and the ESXi host interface with the rest of the network and users. In addition, ESXi hosts use the virtual networking layer to communicate with iSCSI SANs and NAS storage.

The virtual networking layer includes virtual network adapters and the virtual switches. Virtual switches are the key networking components in vSphere. The following figure provides an overview of virtual networking in vSphere.

Figure 10. vSphere Virtual Networking Overview
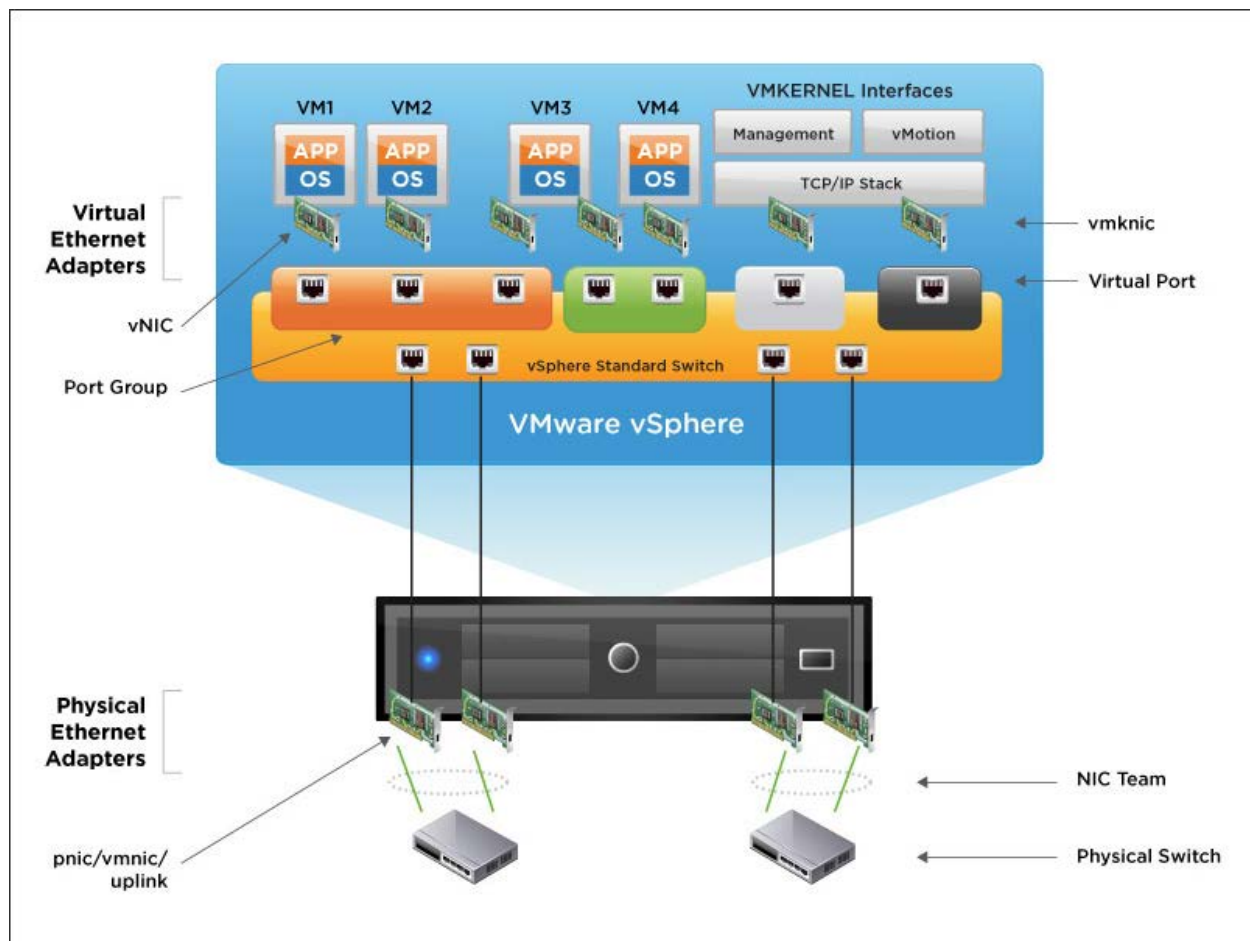
As shown in the preceding figure, the following components make up the virtual network:

- Physical switch – vSphere host-facing edge of the physical local area network.

- NIC team – Group of physical NICs connected to the same physical/logical networks to provide redundancy.

- Physical network interface (pnic/vmnic/uplink) – Provides connectivity between the ESXi host and the local area network.

- vSphere switch (standard and distributed) – The virtual switch is created in software and provides connectivity between virtual machines. Virtual switches must uplink to a physical NIC (also known as vmnic) to provide virtual machines with connectivity to the LAN, otherwise virtual machine traffic is contained within the virtual switch.

- Port group – Used to create a logical boundary within a virtual switch. This boundary can provide VLAN segmentation when 802.1q trunking is passed from the physical switch, or it can create a boundary for policy settings.

- Virtual NIC (vNIC) – Provides connectivity between the virtual machine and the virtual switch.

- VMkernel (vmknic) – Interface for hypervisor functions, such as connectivity for NFS, iSCSI, vSphere vMotion, and VMware vSphere Fault Tolerance logging.

- Virtual port – Provides connectivity between a vmknic and a virtual switch.

2.4.2 Virtual Networking Best Practices

The standard VMware networking best practices apply to running Exchange on vSphere:

- The choice between standard and distributed switches should be made outside of the Exchange design. Standard switches provide a straightforward configuration on a per-host level. For reduced management overhead and increased functionality, the distributed virtual switch should be considered. Both virtual switch types provide the functionality needed by Exchange Server 2016.

- Traffic types should be separated to keep like traffic contained to designated networks. vSphere can use separate interfaces for management, vSphere vMotion, and network-based storage traffic. Additional interfaces can be used for virtual machine traffic. Within virtual machines, different interfaces can be used to keep certain traffic separated. Use 802.1q VLAN tagging and virtual switch port groups to logically separate traffic. Use separate physical interfaces and dedicated port groups or virtual switches to physically separate traffic. This is shown in Figure 11.

- Leverage network interface teaming capabilities to provide redundant uplinks for virtual switches. To use this capability, assign at least two physical network interfaces per virtual switch.

- Use the VMXNET3 network adapter – This is a paravirtualized network device that provides better throughput with reduced hypervisor CPU utilization. It is imperative that customers regularly update VMware Tools on VMs using VMXNet3 virtual adapters.

- Previous Microsoft clustering configuration guidance recommends that, for Exchange Server 2016 virtual machines participating in a Database Availability Group (DAG), customers should configure at least two virtual network interfaces, connected to different VLANs or networks. These interfaces provide access for client access and replication traffic. This guidance is no longer operative in recent versions of Windows clustering. Consequently, VMware recommends that customer adheres to the latest guidance from Microsoft regarding whether or not to separate client access traffic and replication traffic onto different network adapters.

- Where possible, consider enabling jumbo frames on the virtual switches where you have enabled vSphere vMotion traffic. This is to improve vSphere vMotion operations, especially for Exchange Server VMs participating in a DAG configuration. You must ensure that jumbo frames are also enabled on your physical network infrastructure before making this configuration on the virtual switches

- Follow the guidelines on guest operating system networking considerations and hardware networking consideration in the *Performance Best Practices for VMware vSphere 6.0* guide (http://www.vmware.com/files/pdf/techpaper/VMware-PerfBest-Practices-vSphere6-0.pdf?vmw_so_vex_mande_12).

2.4.3 Sample Exchange Virtual Network Configuration

Because of the flexibility of virtual networking, the topology can take many different forms. There is no single recommended practice because each provides its own sets of benefits. The following figure shows two examples of host-level configuration based on the most common configurations, which are single and multiple virtual switches.

Figure 11. Sample Virtual Network Configuration
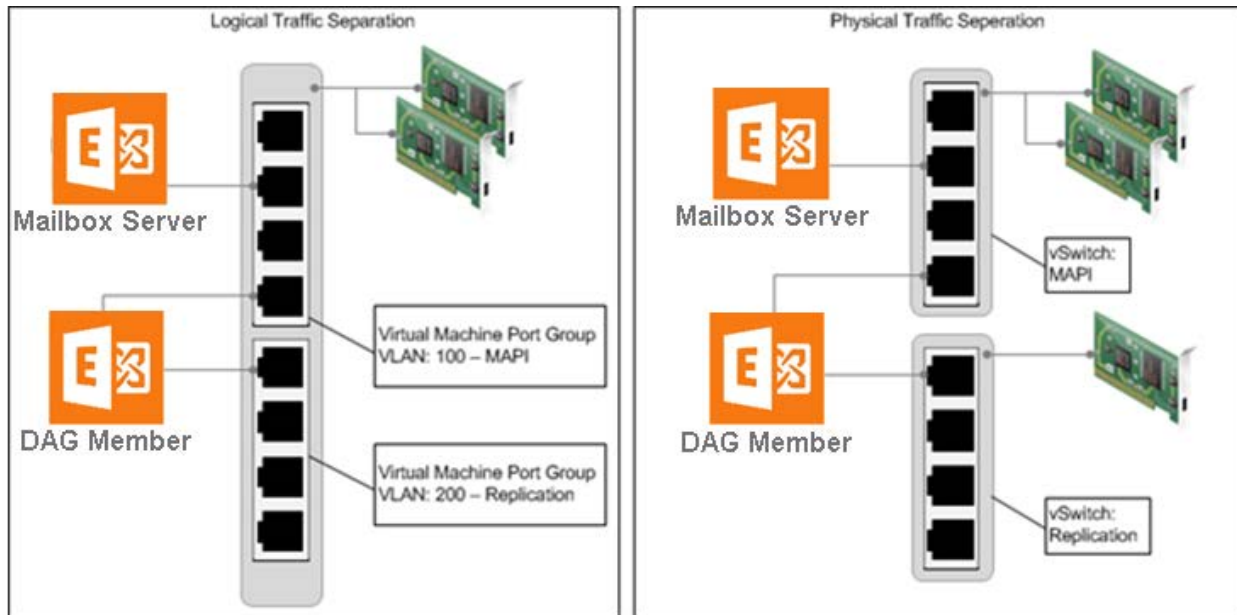
There is no requirement to separate client access and DAG replication traffic onto different network adapters. However, this configuration is still a general practice among many customers. Although VMware encourages customers to validate this configuration with their Microsoft support representatives, the diagram depicts a configuration for completeness.

In the vSphere environment, traffic separation can be established using virtual or physical networks. The preceding figure provides examples of these two scenarios.

- The scenario on the left depicts an ESXi host with two network interfaces, teamed for redundancy and using virtual networks and port groups to provide the traffic separation for client access and DAG replication traffic. This scenario can also utilize VMware vSphere Network I/O Control for dynamic traffic prioritization.

- The scenario on the right depicts an ESXi host with multiple network interfaces. Physical traffic separation is accomplished by allocating two vmnics on one network to a virtual switch. These vmnics are teamed and dedicated to client access network traffic. DAG replication traffic uses a third vmnic on a separate virtual switch.

In both scenarios, the DAG member virtual machine is connected to both networks, according to best practice.

2.5   Power Management

The vSphere ESXi hypervisor provides a high performance and competitive platform that effectively runs many Tier 1 application workloads in virtual machines. By default, ESXi has been heavily tuned for driving high I/O throughput efficiently by utilizing fewer CPU cycles and conserving power, as required by a wide range of workloads. However, many applications require I/O latency to be minimized, even at the expense of higher CPU utilization and greater power consumption.

VMware defines latency-sensitive applications as workloads that require optimizing for a few microseconds to a few tens of microseconds end-to-end latencies. This does not apply to applications or workloads in the hundreds of microseconds to tens of milliseconds end-to-end-latencies. Microsoft Exchange Server is not considered a "latency sensitive" application. However, given the adverse impact of incorrect power settings in a Windows Operating System, customers should pay special attention to power management on an Exchange Server VM on the vSphere platform. See *Best Practices for Performance Tuning of Latency-Sensitive Workloads in vSphere VMs*

(http://www.vmware.com/files/pdf/techpaper/VMW-Tuning-Latency-Sensitive-Workloads.pdf). Keep in mind that many of the prescriptions provided in this document do not apply to Exchange Server workloads and might induce sub-optimal performance if applied to Exchange Server VMs.

Server hardware and operating systems are engineered to minimize power consumption. Windows and vSphere ESXi hypervisor both favor minimized power consumption over performance. While previous versions of ESXi default to a "high performance" power schemes, vSphere 6.0 defaults a "balanced" power scheme. For critical applications such as Exchange Server, the default power scheme in vSphere 6.0 is not recommended.

Figure 12- Default ESXi 6.0 Power Management Setting
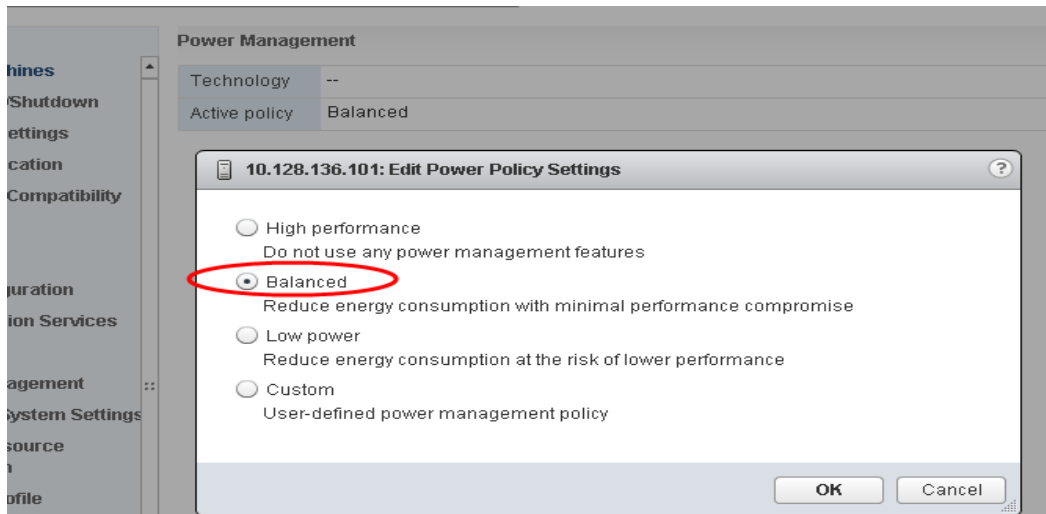
There are three distinct areas of power management in a vSphere hypervisor virtual environment: Server hardware, Hypervisor and Guest OS. The following section provides power management and power setting recommendations covering all of these areas.

2.5.1 Server Hardware BIOS Settings

Most servers with new Intel and AMD processors provide power savings features that use several techniques to dynamically detect the load on a system and put various components of the server, including the CPU, chipsets and peripheral devices into low power states when the system is mostly idle. Hardware-level power management adds latency to the path where an idle system (in one of several power savings modes) responds to an external event. Consequently, VMware recommends the following BIOS-level power management settings for Exchange Server 2016 on vSphere:

- Set Power Management (or its vendor-specific equivalent label) to "OS controlled". This will enable the ESXi hypervisor to make the most judicious power management decision based on actual utilization and the condition of the running virtual machines.

- Disable all processor C-states (including the C1E halt State). These enhanced power management schemes can introduce memory latency and sub-optimal CPU state changes (Halt-to-Full), resulting in reduced performance for the virtual machine.

- Enable Turbo Boost, if available.

2.5.2 ESXi Host Power Settings

ESXi can take advantage of several power management features that the host hardware provides to adjust the trade-off between performance and power use. You can control how ESXi uses these features by selecting a power management policy.

In general, selecting a high performance policy provides more absolute performance, but at lower efficiency (performance per watt). Lower-power policies provide lower absolute performance, but at higher efficiency. ESXi provides five power management policies. If the host does not support power management, or if the BIOS settings specify that the host operating system is not allowed to manage power, only the "Not Supported" policy is available.

Table 3. CPU Power Management Policies

Power Management Policy	Description
High Performance	The VMkernel detects certain power management features, but will not use them unless the BIOS requests them for power capping or thermal events. This is the recommended power policy for an Exchange Server running on ESXi.
Balanced (Default)	The VMkernel uses the available power management features conservatively to reduce host energy consumption with minimal compromise to performance.
Low Power	The VMkernel aggressively uses available power management features to reduce host energy consumption at the risk of lower performance.
Custom	The VMkernel bases its power management policy on the values of several advanced configuration parameters. You can set these parameters in the vSphere Web Client Advanced Settings dialog box.
Not supported	The host does not support any power management features or power management is not enabled in the BIOS.

VMware recommends setting the high performance ESXi host power policy for Exchange Server virtual machines. You select a policy for a host using the vSphere Web Client. If you do not select a policy, ESXi uses **Balanced** by default.

Figure 13 - Recommended ESXi Host Power Management Setting
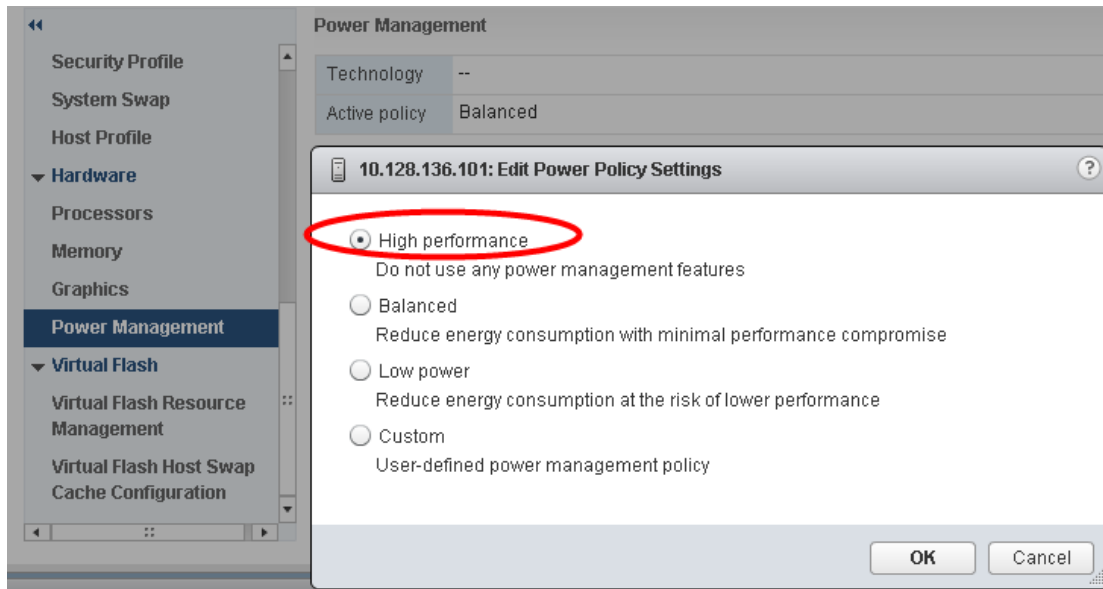
When a CPU runs at lower frequency, it can also run at lower voltage, which saves power. This type of power management is called dynamic voltage and frequency scaling (DVFS). ESXi attempts to adjust CPU frequencies so that virtual machine performance is not affected.

When a CPU is idle, ESXi can take advantage of deep halt states (known as C-states). The deeper the C-state, the less power the CPU uses, but the longer it takes for the CPU to resume running. When a CPU becomes idle, ESXi applies an algorithm to predict how long it will be in an idle state and chooses an appropriate C state to enter. In power management policies that do not use deep C-states, ESXi uses only the shallowest halt state (C1) for idle CPUs.

2.5.3 Windows Guest Power Settings

The default power policy option in Windows Server 2012 is "Balanced". This configuration allows Windows OS to save power consumption by periodically throttling power to the CPU and turning off devices such as the network cards in the guest when Windows determines that they are idle or unused. This capability is inefficient for an Exchange Server workload due to the latency and disruption introduced by the act of powering-off and powering-on CPUs and devices. Allowing Windows to throttling CPUs can result in what Microsoft describes as core parking. See the "Power Management" section of Troubleshooting High CPU utilization issues in Exchange 2013 (http://blogs.technet.com/b/exchange/archive/2015/04/30/troubleshooting-high-cpu-utilization-issues-in-exchange-2013.aspx) for more information.

Figure 14. Windows CPU Core Parking
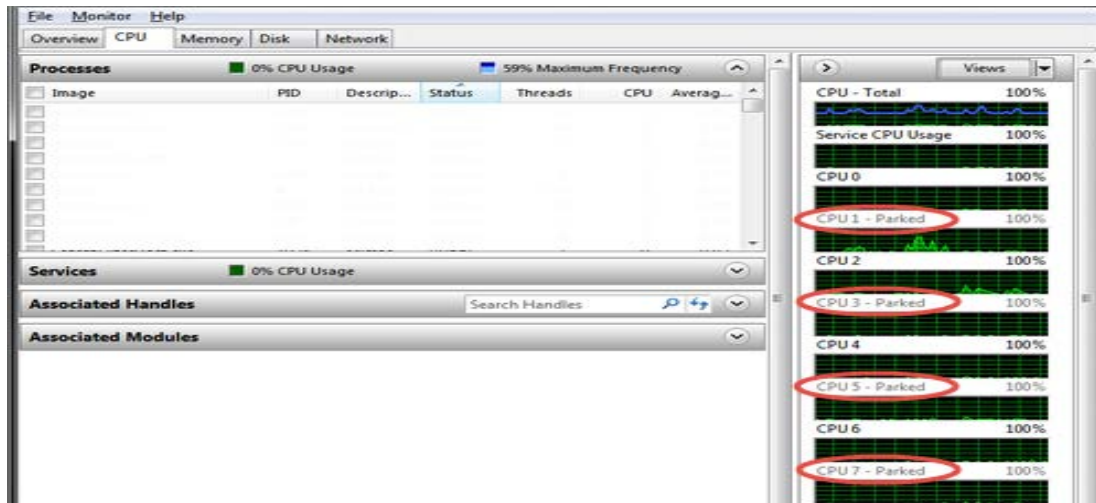
Microsoft recommends the high performance power management policy for applications requiring stability and performance. VMware supports this recommendation and encourages customers to incorporate it into their server tuning and administration practices for virtualized Exchange Server VMs.

Figure 15. Recommended Windows Guest Power Scheme
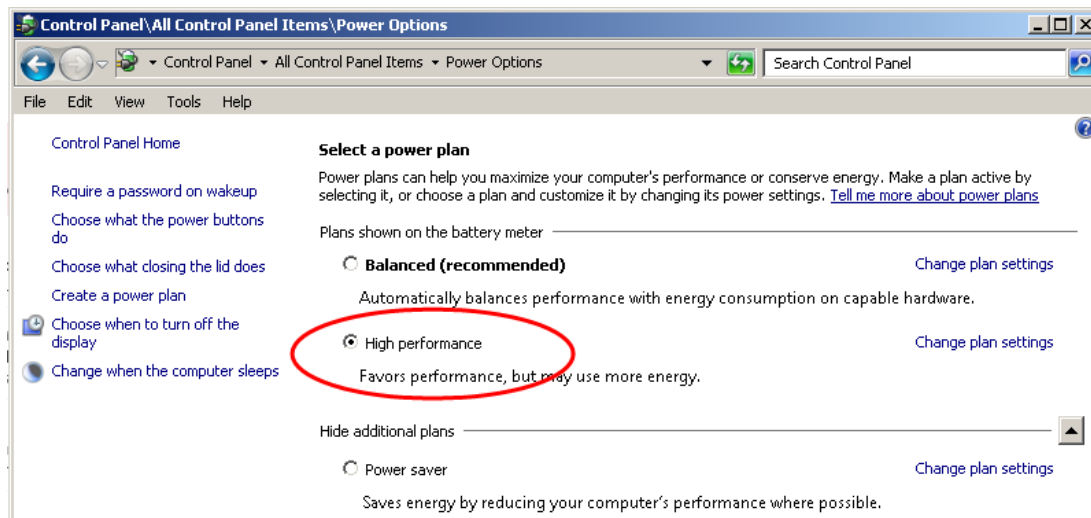
3. Using vSphere Technologies with Exchange Server 2016

The *Microsoft Exchange 2013 on VMware Availability and Recovery Options* guide
(http://www.vmware.com/files/pdf/Exchange_2013_on_VMware_Availability_and_Recovery_Options.pdf)
takes an in-depth look at the options available for building a highly available and site-resilient Exchange
Server 2013 environment on vSphere.

Note The 2016 version of this guide is not yet available.

This section explores the technologies that make those options possible, mainly vSphere HA, vSphere
Distributed Resource Scheduler, and vSphere vMotion. This also includes proven best practices for using
these technologies with critical applications such as Exchange Server 2016.

Although all of the Exchange Server roles have been capable of taking advantage of these advanced
vSphere features, official support by Microsoft for their use with DAG members was not available until
Exchange 2010 SP1. Exchange Server 2016 was released with the same support for these features as
Exchange 2010 SP1, validating the continued effort by both VMware and Microsoft to provide support for
the features customers believe are valuable for virtualized Exchange Server environments.

3.1 Overview of vSphere Technologies

vSphere vMotion technology enables the migration of virtual machines from one ESXi host to another
without service interruption. This migration allows you to move Exchange Server virtual machines from a
heavily loaded ESXi host to one that is lightly loaded or to offload them to allow for hardware
maintenance without any downtime.

3.1.1 vSphere HA

With vSphere HA, Exchange Server virtual machines on a failed ESXi host can be restarted on another
ESXi host. This feature provides a cost-effective failover alternative to third-party clustering and
replication solutions.

If you use vSphere HA, be aware of the following:

- vSphere HA handles ESXi host hardware failure and does not monitor the status of the Exchange
  services. These must be monitored separately.

- A vSphere HA heartbeat is sent using the vSphere VMkernel network, so optimal uplink bandwidth
  and redundancy in this network are strongly recommended.

- Allowing two nodes from the same DAG to run on the same ESXi host for an extended period is not
  recommended when using symmetrical mailbox database distribution. This condition will create a
  single-point-of-failure scenario if the two nodes have the only copies of one or more mailbox
  databases. DRS anti-affinity or guest-to-host affinity rules should be used to mitigate the risk of
  running active and passive mailbox databases on the same ESXi host.

3.1.2 DRS

DRS takes vSphere vMotion a step further by adding an intelligent scheduler. DRS allows you to set
resource assignment policies that reflect business needs. DRS does the calculations and automatically
handles the details of physical resource assignments. It dynamically monitors the workload of the running
virtual machines and the resource utilization of the physical servers within a vSphere cluster.

vSphere vMotion and DRS perform best under the following conditions:

- The source and target ESXi hosts must be connected to the same network and the same shared
  storage.

- A dedicated gigabit (or higher) network for vSphere vMotion is recommended.

- The vSphere vMotion or DRS target host must have enough resources to accommodate the VM that will be migrated.

- The virtual machine must not have physical devices such as a CD-ROM or floppy disk attached during the vSphere vMotion operation.

- The source and destination hosts must have compatible CPU models, otherwise migration with vSphere vMotion fails. If the vSphere cluster contains hosts with different CPU generations, then Enhance vSphere vMotion Compatibility (EVC) must be enabled in the cluster to allow vSphere vMotion operations to succeed. See *EVC and CPU Compatibility FAQ* (http://kb.vmware.com/kb/1005764) for more information on EVC.

- Virtual machines with smaller memory sizes are better candidates for migration than larger ones.

- Persistent resource over-commitment in a vSphere cluster can impede the efficiency of vSphere vMotion operations

3.1.3 vSphere vMotion

Support for using vSphere vMotion with Exchange virtual machines with Exchange DAG members has existed since early 2011 when Exchange 2010 SP1 was released. A well-designed and purpose-built vSphere infrastructure can provide seamless migration of running workloads with no special configuration. In the case of Exchange Server 2016 DAG members, this is essential for utilizing vSphere vMotion with no interference to the cluster services. Even with ideal conditions, the heavy load of Exchange workloads and memory usage can cause a vSphere vMotion operation to trigger a database failover. Database failovers are not necessarily a problem if the environment is designed to properly distribute the load and can help to validate the cluster health by activating databases that might normally go for weeks or months without accepting a user load. However, many administrators prefer that database activations be a planned activity or only done in the case of a failure. For this reason, VMware has studied the effect of vSphere vMotion on Exchange DAG members and provided the following best practice recommendations.

3.1.3.1. Cluster Heartbeat Settings

Exchange Server's Database Availability Group (DAG) relies on the Windows Server Failover Clustering (WSFC) Service. By default, each clustered Exchange Server node exchanges heartbeat packets with its partners every second. If the node does not receive a response from any partner after five consecutive probes, the probing node considers the partner to be non-responsive (unavailable). The WSFC service on the probing node then initiates corrective actions, including taking ownership of clustered resources previously owned by the non-responsive partner.

A clustered node considers its partner unavailable if it is unable to exchange heartbeat packets with that partner after five attempts (five seconds), which triggers actions necessary to provide continued availability of the resources located on that partner. This is what happens when there is an unintended and unplanned cluster resource failover.

An aggressive cluster failover threshold can be problematic and disruptive. For example, a failover might not have been required because the non-responsiveness was a transient issue that went away after six seconds.

During a vSphere vMotion operation, a virtual machine's memory pages are gradually copied from its current parent host to its target parent host and the virtual machine is subsequently switched over to the target host. During this copy and transition procedure, the virtual machine is quiesced (VMware refers to this as *stunning*). Under normal operating condition, the stun period is very brief and not noticeable. However, the quiescing period might last longer than the five-second threshold under some operating conditions, including the following:

- Size of the VM (CPU and RAM).

- Current outstanding operations, threads, and processes on the VM.

- Total CPU subscription ratio in the vSphere cluster.

- Network bandwidth or congestion state.

The net effect of a stun operation is that the stunned VM is unable to exchange heartbeat for the duration of the stun operation. If that lasts longer than five seconds and the VM is in a cluster relationship, then its cluster partners consider it unavailable.

This behavior is not peculiar to vSphere vMotion, or even virtualization in general. If your backup solution takes VSS-based snapshots of your server, it is also likely quiescing the server.

This is a known behavior. See *Tuning Failover Cluster Network Thresholds* (http://blogs.msdn.com/b/clustering/archive/2012/11/21/10370765.aspx) for a detailed discussion that includes avoiding an unintended cluster failover incident (and its associated disruptive effects) when performing a vSphere vMotion operation on a DAG node. There are several configuration options described in the following sections that you can use to overcome these disruptive effects.

Two Windows Server Failover Cluster settings control the behavior of the cluster service to missed heartbeat probes. These are:

- `SameSubnetDelay` – Controls how often a node sends heartbeat probe packets.

- `SameSubnetThreshold` – Controls how many probe misses the node must tolerate before taking actions.

If the clustered nodes are in different subnets, the corresponding settings are:

- `CrossSubnetDelay`

- `CrossSubnetThreshold`

The following are the default values for these settings:

```
PS C:\Windows\system32> get-cluster | fl *subnet*

CrossSubnetDelay          : 1000
CrossSubnetThreshold      : 5
PlumbAllCrossSubnetRoutes : 0
SameSubnetDelay           : 1000
SameSubnetThreshold       : 5
```

A `SubnetDelay` value of 1000ms and a `SubnetThreshold` of 5 denotes that heartbeat probes are sent every second with a tolerance for a maximum of 5 missed heartbeats. These are the settings that enabled the unexpected failover described in the preceding section.

Adjusting either of these parameters to 10 seconds (for cluster nodes in the same subnet) or 20 seconds (for cluster nodes in different subnets) is the most effective way to prevent the unintended cluster failover issue described in the preceding section. The following are the settings to accomplish each of the recommendations:

```
(get-cluster).SameSubnetThreshold = 10
```

```
(get-cluster).CrossSubnetThreshold = 20
```

```
PS C:\Windows\system32> (get-cluster).SameSubnetThreshold = 10
PS C:\Windows\system32> (get-cluster).CrossSubnetThreshold = 20
PS C:\Windows\system32> get-cluster | fl *subnet*


CrossSubnetDelay          : 1000
CrossSubnetThreshold      : 20
PlumbAllCrossSubnetRoutes : 0
SameSubnetDelay           : 1000
SameSubnetThreshold       : 10
```

Windows Server 2012 includes an additional cluster service logging component that tracks dropped heartbeat packets. The `RouteHistoryLength` setting and its default value (10) must also be modified so that it remains useful when investigating heartbeat-related issues. According to Microsoft, the recommended value for this setting must be double that of the `SubnetThreshold` setting. If the value `SameSubnetThrshold` and `CrossSubnetThreshold` follow the VMware recommendations, the value of `RouteHistoryLength` must be double the value of the `CrossSubnetThreshold` value. Therefore, the VMware recommended value for `RouteHistoryLength` setting is 40.

```
(get-cluster).RouteHistoryLength = 40
```

```
PS C:\Windows\system32> get-cluster | fl RouteHistoryLength


RouteHistoryLength : 10



PS C:\Windows\system32> (get-cluster).RouteHistoryLength = 40
PS C:\Windows\system32> get-cluster | fl RouteHistoryLength


RouteHistoryLength : 40
```

There is no need to restart the cluster service for these changes to take effect.

The unintended database failover during vSphere vMotion is a direct result of the default heartbeat timeout threshold in Windows Server Failover Cluster.

3.1.3.2. Multiple vSphere vMotion Interfaces

Database failover due to vSphere vMotion operations can be mitigated by using multiple dedicated vSphere vMotion network interfaces. In most cases, the interfaces that are used for vSphere vMotion are also used for management traffic. Because management traffic is relatively light, this does not add significant overhead.

vSphere provides the ability to use multiple vmnic interfaces for vSphere vMotion traffic to effectively load balance the vSphere vMotion traffic. Testing has shown up to a 25% increase in throughput achieved when multiple vSphere vMotion interfaces are used. In the test case with two 2 Gbps interfaces configured for vSphere vMotion and no cluster heartbeat modifications, vSphere vMotion operations succeeded with no database failover.

Enabling multiple interfaces for vSphere vMotion requires configuring multiple VMkernel ports on different port groups. Each port group is assigned multiple vmnic interfaces as either active or standby. See *Multiple-NIC vMotion in vSphere* (http://kb.vmware.com/kb/2007467) for detailed configuration procedures.

Before deciding on using this configuration option, see *vMotion migrations fail when using multiple VMkernel ports for vMotion in different IP subnets* (http://kb.vmware.com/kb/2052092).

3.1.3.3. Enable Jumbo Frames for vSphere vMotion Interfaces

Standard Ethernet frames are limited to a length of approximately 1500 bytes. Jumbo frames can contain a payload of up to 9000 bytes. Support for jumbo frames on VMkernel ports was added to vSphere 4.0 for both ESX and ESXi. This added feature means that large frames can be used for all VMkernel traffic, including vSphere vMotion.

Using jumbo frames reduces the processing overhead to provide the best possible performance by reducing the number of frames that must be generated and transmitted by the system. During testing, VMware had an opportunity to test vSphere vMotion migration of DAG nodes with and without jumbo frames enabled. Results showed that, with jumbo frames enabled for all VMkernel ports and on the VMware vNetwork Distributed Switch, vSphere vMotion migrations of DAG member virtual machines completed successfully. During these migrations, no database failovers occurred, and there was no need to modify the cluster heartbeat setting.

The use of jumbo frames requires that all network hops between the vSphere hosts support the larger frame size. This includes the systems and all network equipment in between. Switches that do not support (or are not configured to accept) large frames will drop them. Routers and Layer 3 switches might fragment the large frames into smaller frames that must then be reassembled and this can cause both performance degradation and a pronounced incidence of unintended database failovers during a vSphere vMotion operation. Do not enable jumbo frames within a vSphere infrastructure unless the underlying physical network devices are configured to support this setting.

3.2　vSphere Distributed Resource Scheduler

Distributed resource scheduling provides active load balancing of virtual machine workloads within a vSphere cluster. Aside from the active monitoring and load balancing functions, DRS provides the following features:

- Virtual machine placement during power on, based on resource requirements and availability.

- Virtual machine evacuation during ESXi host maintenance mode.

- Virtual machine and host groups for grouping like objects.

- Rules to keep virtual machines together or apart and on or off of a set of hosts.

DRS helps make a virtualized Exchange Server 2016 environment more agile. The following sections provide recommendations for using DRS with Exchange Server 2016.

3.2.1　Enable DRS in Fully Automated Mode

DRS provides the following levels of automation:

- Manual – Migration recommendations are provided by DRS. No migrations are performed by DRS.

- Partially automated – Virtual machines are automatically placed on hosts during power-on, migration recommendations are provided by DRS, and no migrations are performed by DRS.

- Fully automated – Virtual machines are automatically placed on hosts during power on and are automatically migrated between hosts to optimize resource usage.

When designed according to VMware recommendations, vSphere clusters that have been purpose built for Exchange Server 2016 possess sufficient resources and do not experience many DRS migrations. However, when an ESXi host is placed in maintenance mode, DRS makes recommendations on placement of virtual machines running on that host. To leverage automatic host evacuation, the DRS automation level must be set to Fully Automated.

If the vSphere cluster hosting Exchange Server 2016 is a shared environment, DRS in fully automated mode helps to maintain resource optimization among the multiple workloads.

3.2.2   Use Anti-Affinity Rules for Exchange Virtual Machines

DRS provides rules for keeping virtual machines apart or together on the same ESXi host or group of hosts. In an Exchange environment, the common use case for anti-affinity rules is to keep Exchange virtual machines with the same roles installed separated from each other.

Microsoft recommends symmetrically distributing mailbox databases among DAG members. Unlike traditional active/passive configurations, this design allows all DAG members to support active users as well as reserve a portion of compute power for failover capacity. In the case of failure of a single DAG member, all remaining members might take part in supporting the failed databases. Because of this, it is recommended that no two members of the same DAG run on the same ESXi host for an extended period.

Figure 16. vSphere Distributed Resource Scheduler Anti-Affinity Rule
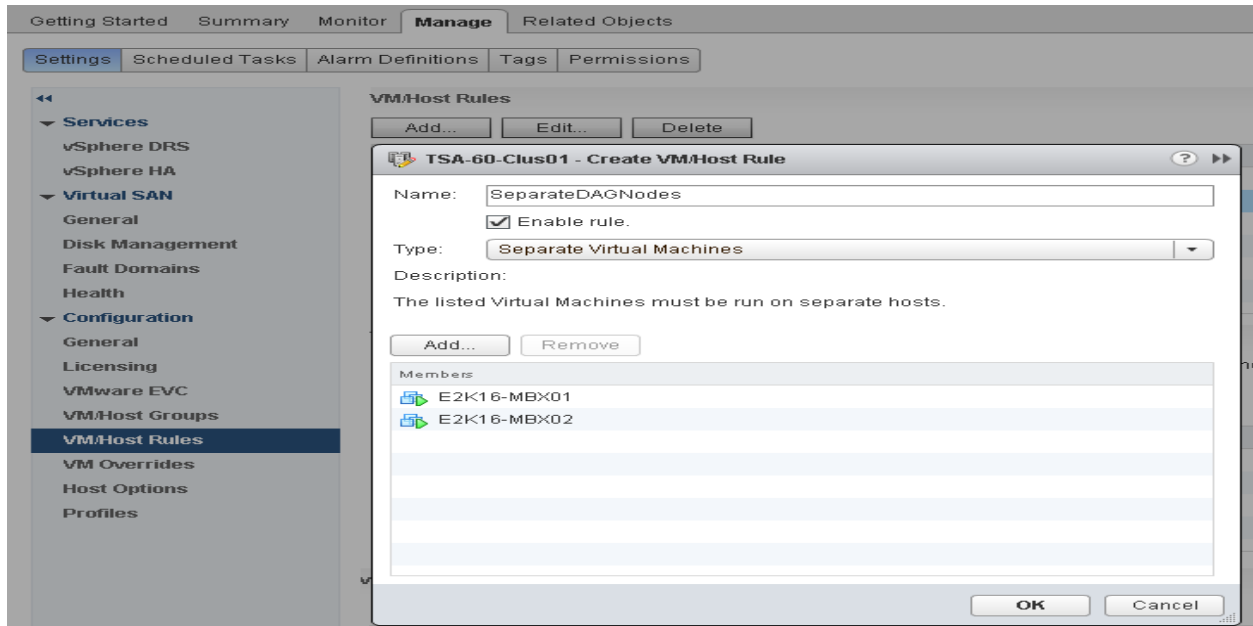
Anti-affinity rules enforce virtual machine separation during power on operations and vSphere vMotion migrations due to a DRS recommendation, including a host entering maintenance mode. Prior to vSphere 5.5, if a virtual machine is enabled for vSphere HA and a host experiences a failure, vSphere HA might power on a virtual machine and violate a DRS anti-affinity rule. This is because vSphere HA does not inspect DRS rules during a recovery task. However, during the next DRS evaluation (every 5 minutes), the virtual machine is migrated to fix the violation.

To avoid this condition when utilizing DRS with vSphere 5.5, VMware encourages customers to apply the following vSphere HA Advanced Configuration option to their vSphere Clusters:

das.respectVmVmAntiAffinityRules = TRUE

This setting instructs vSphere HA to inspect and respect vm-vm anti-affinity rules when restarting virtual machines after a host failure. The Exchange Server VMs separated by an anti-affinity will not be co-located on the same ESXi host.

Figure 17. HA Advanced Configuration Option for DRS Anti-Affinity Rules
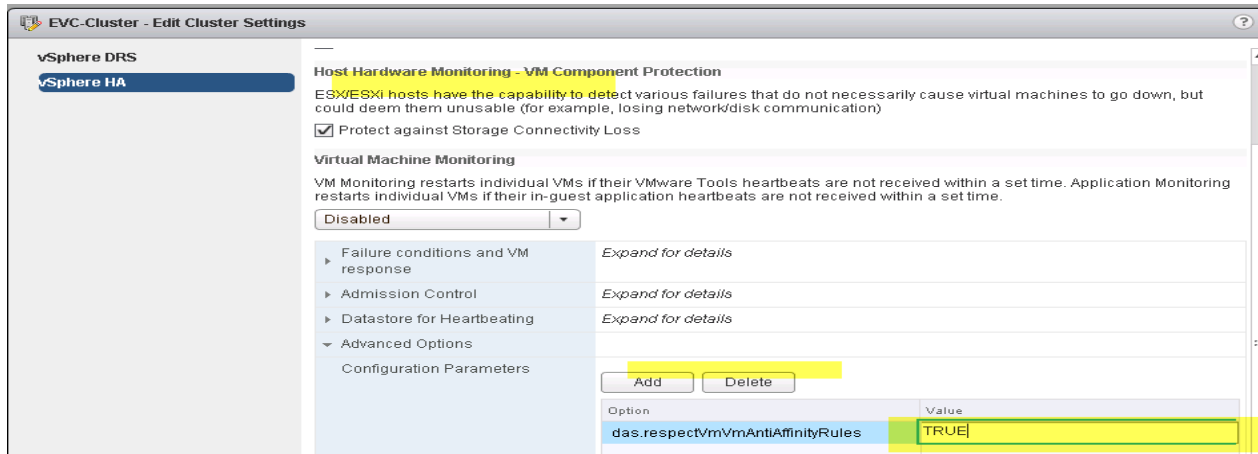
As shown in the following figure, vSphere 6.0 includes an improved, GUI-based configuration option to control how vSphere HA responds to all DRS rules in a cluster. It is no longer required to configure the Advanced Configuration parameter manually.

Figure 18. Improved vSphere HA and DRS Interoperability in vSphere 6.0

3.2.3 DRS Groups and Group-Based Rules

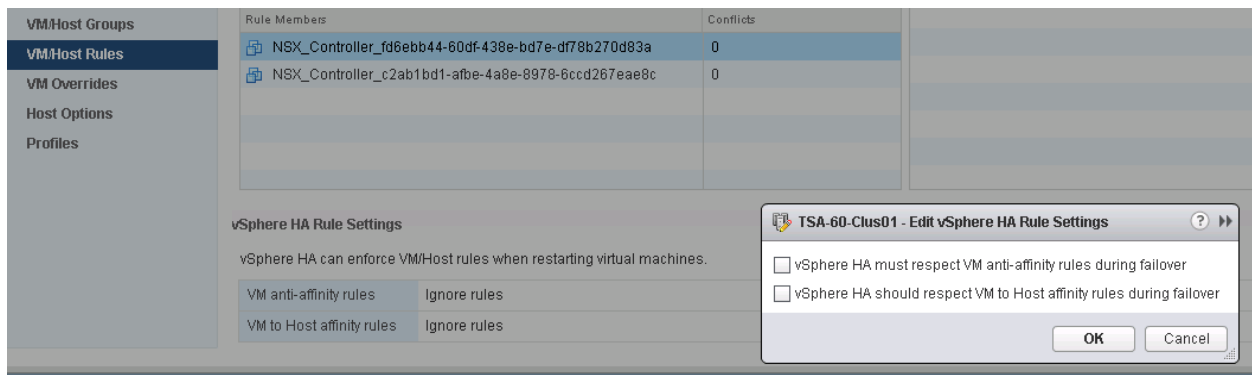Defining DRS groups helps enforce virtual machine placement in a way that is not possible with affinity and anti-affinity rules only. Host groups can contain hosts licensed to run an application or hosts in a blade chassis or rack. Virtual machine groups can contain virtual machines with a similar purpose. With groups defined, the *virtual machines to hosts*-type rule is available for use. Virtual machines to hosts rules can be created with four variations:

- VM group must run on hosts in group.

- VM group should run on hosts in group.

- VM group must not run on hosts in group.

- VM group should not run on hosts in group.

Must run on rules provide hard enforcement of virtual machine placement. If a rule stipulates that a group of virtual machines must run on a group of ESXi hosts, both DRS and vSphere HA obey these rules. If all hosts in the group are down, the virtual machines are unable to run on any other host in the vSphere cluster.

In the following figure, two virtual machine groups and two host groups are defined. Two *must run on* rules, shown in the solid black ovals, keep the virtual machines in each group running on their respective host group. The virtual machine in the middle is not tied to a group or a rule and might roam. In the case of a failure of all hosts in the group, all virtual machines bound to those hosts by a *must run on* rule stay offline until a host from that group is brought back online.

Figure 19. Must Run on Rule Example
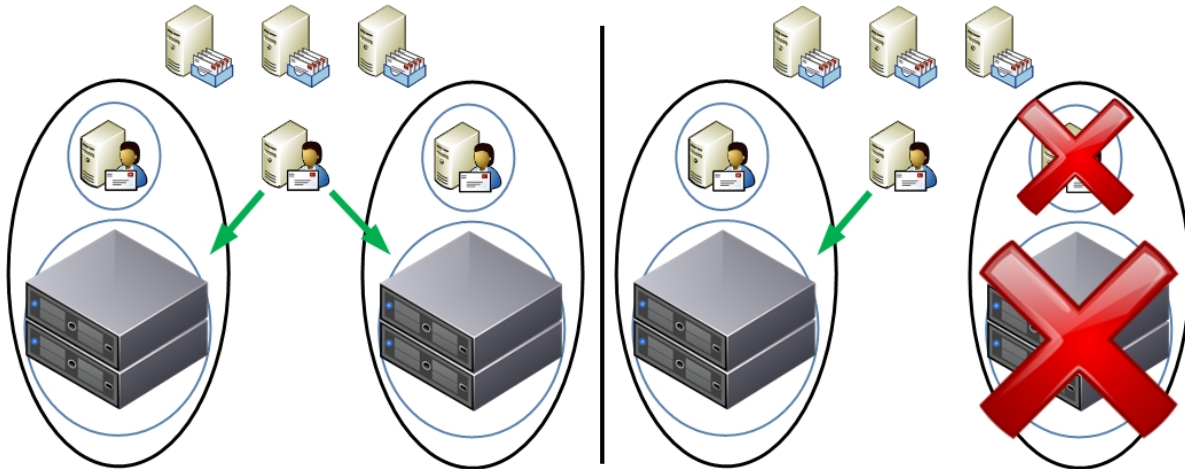
Should run on rules provide soft enforcement of virtual machine placement. If a rule stipulates that a group of virtual machines *should run on* a group of ESXi hosts, those virtual machines will always be preferentially placed on hosts in the host group. They can still run on other hosts in the vSphere cluster outside of the host group, if needed (for example, if all the hosts in the host group are unavailable or otherwise unsuitable for the VM.

In the following figure two virtual machine groups and two host groups are defined. Two *should run on* rules, shown in the broken green ovals, keep the virtual machines in each group running on their respective host group. The virtual machine in the middle is not tied to a group or a rule and might roam. In the case of a failure of all hosts in the group, virtual machines bound to those hosts by a *should run on* rule can be brought back online by vSphere HA.

Figure 20. Should Run on Rule Example
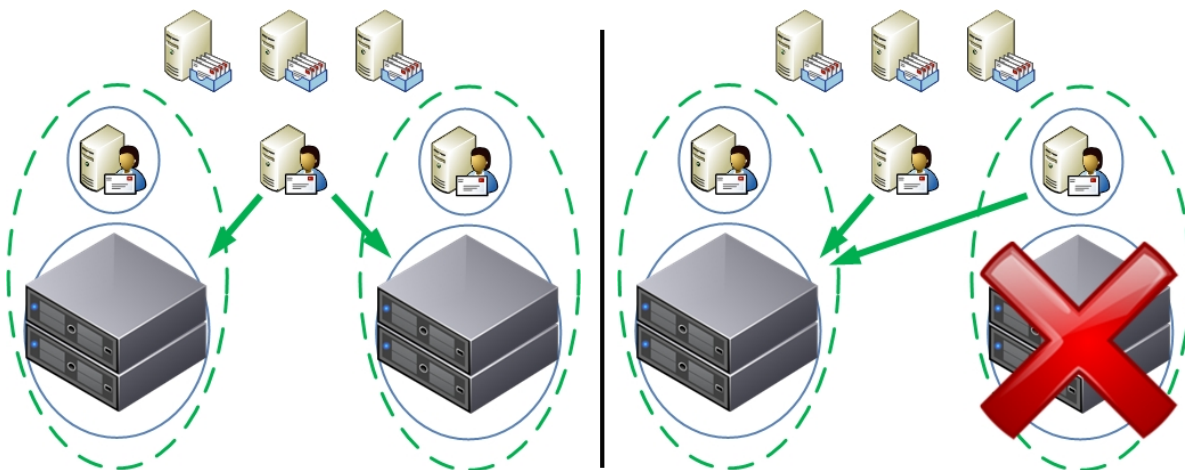
In an Exchange Server 2016 environment, virtual machine to host rules can be used to provide soft or hard enforcement of virtual machine placement. As an example, consider creating groups of ESXi hosts

based on a failure domain, such as a blade chassis or server rack. Create two virtual machine groups with each containing half of the Exchange Server virtual machines, and create rules to link each virtual machine group to a host group. In the case of a complete chassis or rack failure, any virtual machines that have failed can be powered back on by vSphere HA.

3.3   vSphere High Availability

By providing a higher level of availability than is possible out-of-the-box for most applications, vSphere HA has become the default HA solution for vSphere virtual machines. Regardless of operating system or application, vSphere HA can provide protection from ESXi host failures, guest operating system failures, and, with the help of third-party add-ons, application failures.

Exchange Server 2016 environments are built for high availability. Mailbox servers are deployed in DAGs for mailbox database high availability. Theoretically, this provides all of the availability required by an Exchange Server. However, in the case of a hardware failure, utilization of the remaining Exchange servers rises as new connections are established, and DAG protection is compromised as passive databases are activated. In a physical deployment, an administrator needs to address the problem quickly to restore availability levels and mitigate any further outages. This condition informs Microsoft's recommendation at least four (4) DAG copies in a production Exchange Server infrastructure, increasing cost of ownership (storage, licensing, etc) as well as administrative and management overhead. With a vSphere infrastructure, a hardware failure results in virtual machines being powered back on by vSphere HA, restoring availability levels quickly and keeping utilization balanced. With increased availability, customers are able to reduce the number of DAG copies required to efficiently run their Exchange Server infrastructure. The following sections provide recommendations for using vSphere HA with Exchange Server 2016.

3.3.1 Admission Control

The amount of failover capacity provided by hosts in a vSphere cluster determines how many host failures can be tolerated. For example, in a four-host cluster, each host needs to reserve at least 25% of its resources to accommodate the workloads of one host in the event of a failure. Reserving this capacity is a manual task without admission control.

vSphere Admission Control allows administrators to define policies that are used by vCenter to manage failover capacity. How vCenter failover capacity is calculated and reserved is determined by the following policies:

- The number of host failures a cluster can tolerate.

- A percentage of the cluster resources that are reserved as failover spare capacity.

- Dedicated failover hosts.

- Sizes of the VMs in the cluster.

Surviving a physical host failure without compromising performance or availability is a major driver for virtualizing business critical applications. By configuring admission control, VMware vCenter Server™ monitors utilization and manages failover capacity. To provide sufficient resources in the case of a hardware failure, or even during maintenance, enable and configure admission control.

For more information on admission control and the policies see *vSphere HA Admission Control* in the *vSphere Availability Guide* (http://pubs.vmware.com/vsphere-60/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-601-availability-guide.pdf).

3.3.2 Virtual Machine Monitoring

Along with ESXi host monitoring, vSphere HA can also provide monitoring at the virtual machine level. Virtual machine monitoring can detect guest operating system failures and, with the help of third-party software, application failures. vSphere HA establishes a heartbeat with VMware Tools, installed within the guest operating system, and monitors this heartbeat. If heartbeat communication fails between VMware

Tools and vSphere HA, a secondary check is made against network and storage I/O activity. If both network and storage I/O activity has halted, vSphere HA triggers a virtual machine restart.

Enable virtual machine monitoring if there is a desire to reduce downtime due to guest operating system failures that would otherwise require manual intervention. This monitoring feature cannot detect Exchange processes, database health or failures. It is, therefore, not a replacement for Exchange DAG.

3.3.3 Using vSphere HA with Database Availability Groups

In a physical environment, DAGs are often deployed with three or more database copies to protect from hardware and disk failures. In these environments, when a physical server or storage component fails, the DAG is still protected due to the multiple database copies. This comes at the expense of managing multiple database copies, and other administrative overhead. Exchange environments built on vSphere are typically designed with two database copies and utilize vSphere HA and RAID to protect from hardware and storage failures. vSphere HA restarts a DAG member if the host where it was running experiences a hardware failure or becomes otherwise unavailable, and RAID protects databases from storage failure at the array level.

When enabling a vSphere cluster for HA with the intention of protecting DAG members, consider the following:

- Members of the same DAG should not reside on the same vSphere host for an extended period of time when databases are symmetrically distributed between members. Allowing two members to run on the same host for a short period of time (for instance, after a vSphere HA event), even if doing so may violate resource availability constraints and DRS rule, allows the Exchange server VM to become operational and for database replication and protection to resume quicker. DAG members should be separated as soon as operationally feasible (for example, as soon as the ESXi host becomes available or additional capacity has been added to the vSphere cluster.

- To adequately protect from an extended server outage, vSphere clusters should be designed in an N+1 configuration, where N is the number of DAG members. If a hardware failure occurs causing vSphere HA to power on a failed DAG member, Exchange servers and DAG maintain the same levels of performance and protection as during normal runtime.

- Use anti-affinity rules to keep DAG members separated. vSphere HA might violate this rule during a power-on operation (one caused by a host failure), but DRS fixes the violation during the next interval. To eliminate the possibility of DAG members running on the same host (even for a short period), "must not run on virtual machine to host" anti-affinity rules must be used.

4. Exchange Performance on vSphere

Since 2006, VMware and its partners have used testing to successfully demonstrate the viability of running Exchange on the VMware Infrastructure platform. This testing has been confirmed by organizations that have deployed Exchange 2003, 2007, 2010, and 2013 in virtualized production environments and now benefit from the considerable operational advantages and cost savings. The ease of virtualizing an Exchange Server infrastructure and the demonstrable parity in performance, reliability, availability and recoverability of a virtualized Exchange Server workload over the years have combined to dispel the fear that many organizations had about virtualizing Exchange Server.

As previously mentioned, improvements in Exchange Server 2016 and the unambiguous support statement by Microsoft both make the decision about virtualizing Exchange Server a less onerous one. Even customers who have not adopted the new normal "Virtualize First" policy can expect the same optimal performance and reliability that they are familiar on their physical-hardware Exchange Server instances when they virtualize those servers.

Server hardware compute resources continue to expand on a regular basis. It is now quite uncommon to purchase a modern server that has fewer than 40 CPUs and 512GB of RAM. These modern-class servers are unsuitable for directly running Exchange Server 2016, due to the technical limitations in compute resources addressable by Exchange Server. Since an Exchange Server has an upper limit of 96GB RAM and 24 CPUs, directly installing Exchange Server on them not only becomes cost-ineffective, it can also induce performance degradation.

This makes virtualization the most optimal and cost-effective configuration option for Exchange Server 2016. Virtualization helps customers overcome these limitations by efficiently partitioning and maximizing the compute resources and allocating them to the Exchange Server VMs as required, without inducing performance bottleneck.

Many enterprises have adopted virtualization as their default design choice for new workloads. This choice is logical for an Exchange Server infrastructure for the following reasons:

- The Exchange information store (the *Managed Store*) has been rewritten to further optimize resource consumption. This update to the Managed Store has also led to further reduction in storage I/O requirements.

- Advances in server hardware such as multicore processors, higher memory density, and advances in storage technology are far outpacing the performance requirements for applications, including Exchange Server 2016. Virtualization becomes an effective way to leverage the full power of these systems.

- The advances in Exchange Server 2016 and server hardware technology have coincided with advances in vSphere. On vSphere 6.0, Virtual machines support up to 4 TB RAM and 128 vCPUs, both of which are significantly higher than the maximum recommended compute resources for a single Exchange Server. The question of whether or not virtualization can support the performance requirements of an Exchange server is, therefore, moot.

VMware conducts baseline performance testing of every release of Exchange Server in order to be able to understand compute requirements and to provide the applicable guidance to our customers. Because of the continued improvements carried over into Exchange Server 2016 storage IO requirements, our validation tests (using the Microsoft Jetstress tool) revealed that, even under the most strenuous conditions, storage is not a significant cause of performance bottleneck for Exchange Server 2016 if the storage architecture adheres to VMware's recommended configuration practices.

With concerns over relative storage IOs and performance eliminated, many more Exchange administrators are finding the flexibility, enhanced availability, and lower costs associated with virtualization very attractive in supporting an Exchange infrastructure.

4.1   Key Performance Considerations

A variety of factors can affect Exchange Server 2016 performance on vSphere, including processor and memory allocation to the guest virtual machine, storage layout and design, virtual machine placement, and high availability methods. The following are tips for achieving the best possible performance:

- Fully understand your organization's business and technical requirements for implementing Exchange Server 2016.

- Fully understand the Exchange workload requirements.

- Although I/O is reduced in Exchange Server 2016 compared with prior versions of Exchange Server, Exchange 2010, there is still a requirement to provide adequate throughput and low latency. Dedicate physical storage for Exchange to avoid compromising I/O by having other workloads running on the same physical disks.

- Use the correct Microsoft sizing tools to obtain the appropriate configuration guidelines for your Exchange Server 2016 virtual machines. The two primary tools for sizing an Exchange Server 2016 infrastructure are as follows:

  o Exchange Server Role Requirements Calculator (https://gallery.technet.microsoft.com/Exchange-2013-Server-Role-f8a61780) - Obtain the correct version of this tools for your sizing exercise. For example, using the Exchange Server 2010 version of the tool to size an Exchange Server 2016 implementation will result in sizing recommendations that will result in noticeable performance bottleneck.

  o The Role Calculator requires some input parameters that can result in under- or over-sizing if the supplied information is inaccurate. One of the most important of these parameters is what Microsoft describes as the *megacycle*. The megacycle is a computed value derived from the publicly published SPECInt Ratings Values (http://www.spec.org/cgi-bin/osgresults?conf=rint2006) of the processor architecture of a given physical server. Manually calculating the correct "megacycle" value to use is your sizing exercise is not a trivial exercise. To help overcome this challenge, Microsoft has provided the Exchange Processor Query Tool (https://gallery.technet.microsoft.com/Exchange-Processor-Query-b06748a5), which automates the retrieval of the appropriate rating value for a specified hardware and enables a less cumbersome derivation of the correct "megacycle" value.

  VMware strongly encourages customers to utilize these tools when sizing their Exchange Server 2016 workloads.

- Follow the best practices in Section 2, ESXi Host Best Practices for Exchange, to optimize the ESXi host environment for enterprise applications such as Exchange Server.

4.2   Performance Testing

Every Exchange environment is different, with varying business and technical requirements, many server and storage options, and requirements for integrating with third-party software solutions such as antivirus, anti-spam, and smartphones. Due to the many variables, VMware strongly recommends that organizations test performance on their particular mix of server, storage, and software to determine the best design for their Exchange environment. In addition, several VMware server and storage partners have performed testing to validate Exchange performance on vSphere. Both of these options are discussed in this section.

4.2.1   Internal Performance Testing

Microsoft provides tools to measure the performance of Microsoft Exchange Server architectures. The Microsoft Jetstress tool is used for qualifying and validating the underlying storage architecture for Exchange Server workloads. The main purpose of the Jetstress tool is to measure the IO throughput of the disk subsystem supporting the Exchange Servers at the database level. Although the Jetstress tool

has not been specifically updated for Exchange Server 2016, the current version (2013) is suitable for baselining an Exchange Server 2016 storage requirements and configurations.

The Microsoft Exchange Server Load Generator (LoadGen) simulates client access to an Exchange infrastructure for the purpose of measuring and analyzing Exchange Server performance under heavy client activities. At the time of this writing, Microsoft does not plan to update Loadgen specifically for Exchange Server 2016.

Note   The reduction in storage I/O in Exchange Server 2016 might lead to an oversized proposed configuration when using the Exchange Server 2013 sizing tools.

It is important to address a concern with the collection of performance metrics from within virtual machines. Early in the virtualization of high-performance applications, the validity of in-guest performance metrics came into question because of a time skew that can be possible in highly overcommitted environments. With the advancements in hypervisor technology and server hardware, this issue has mostly been addressed, especially when testing is performed on under-committed hardware. This is validated by Microsoft support for running Jetstress within virtual machines. More information on virtual machine support for Jetstress is available in the *Jetstress 2013 Field Guide* (https://gallery.technet.microsoft.com/Jetstress-2013-Field-Guide-2438bc12)

4.2.2 Partner Performance Testing

VMware and its OEM partners have been working together for years to characterize Exchange performance. This testing helps to understand the performance of Exchange in a virtualized environment, qualify best practice recommendations, and better understand any virtualization overhead impact on Exchange Server. At this time, Exchange Server 2016 is still a new product, and it is unclear whether Microsoft will be releasing an updated version of LoadGen for Exchange Server 2016, so there are currently no published Loadgen-based benchmarking studies or references for Exchange Server 2016 from our partners. Even in the absence of the Loadgen tool for Exchange Server 2016, a similarly-sized Exchange Server 2016 is expected to have better performance characteristics than Exchange Server 2013.

4.3   Ongoing Performance Monitoring and Tuning

Traditional Exchange Server performance monitoring leverages the Microsoft Windows performance monitor tool (Perfmon) to collect statistics. Exchange integrates with Perfmon to provide familiar counters that indicate system performance. Exchange administrators should continue to use familiar tools to monitor performance, especially for Exchange specific counters such as RPC averaged latency. In addition to the standard counters familiar to an Exchange administrator, VMware Tools adds two additional Perfmon counters that can be monitored—VM Memory and VM Processor. These counters provide ESXi host-level insight into the resource allocation and usage of the virtual machine.

Figure 21. Virtual Machine Perfmon Counters
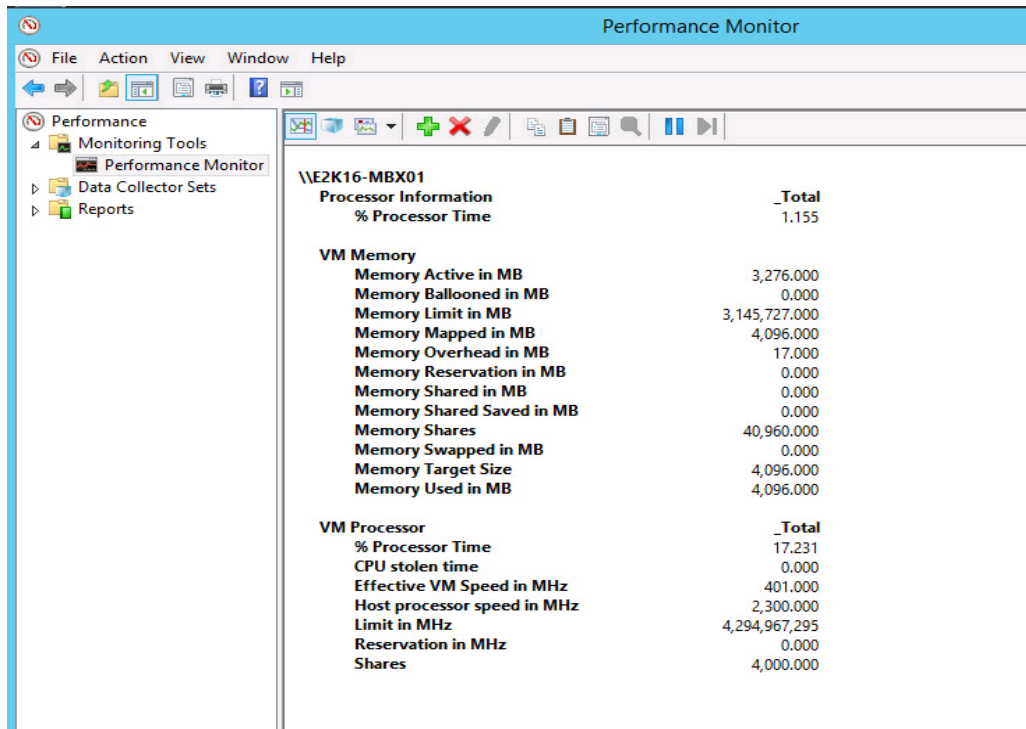
Many of the counters available can be used to help confirm allocations have been set properly when vCenter Server access is not available or for configuration monitoring. The following table lists counters that can be actively monitored.

Table 4. Virtual Machine Perfmon Counters of Interest

Object	Counter	Description
VM Processor	% Processor Time	Processor usage across all vCPUs.
VM Memory	Memory Ballooned	Amount of memory in MB reclaimed by balloon driver.
	Memory Swapped	Amount of memory in MB forcibly swapped to ESXi host swap.
	Memory Used	Physical memory in use by the virtual machine.

vSphere and Exchange administrators can also use the counters listed in the following table to monitor performance at the ESXi host level. Those metrics can then be correlated with metrics from Exchange virtual machines. See the section on performance monitoring utilities in section in *vSphere Monitoring and Performance* (http://pubs.vmware.com/vsphere-60/topic/com.vmware.ICbase/PDF/vsphere-esxi-vcenter-server-601-monitoring-performance-guide.pdf) for a comprehensive list of performance counters and metrics in vSphere and for information on using vSphere native tools to monitor an ESXi host and virtual machine performance.

Table 5. VMware Performance Counters of Interest to Exchange Administrators

Subsystem	esxtop Counters	vCenter Counter
CPU	%RDY	Ready – milliseconds in a 20,000ms window
	%USED	Usage
Memory	%ACTV	Active
	SWW/s	Swapin Rate
	SWR/s	Swapout Rate
Storage	ACTV	Commands
	DAVG/cmd	Device Latency
	KAVG/cmd	Kernel Latency
Network	MbRX/s	packetsRx
	MbTX/s	packetsTx

The preceding table indicates a few key counters that should be added to the list of inspection points for Exchange administrators. Of the CPU counters, the total used time indicates system load. Ready time indicates overloaded CPU resources. A significant swap rate in the memory counters is a clear indication of a shortage of memory, and high device latencies in the storage section point to an overloaded or misconfigured array. Network traffic is not frequently the cause of most Exchange performance problems, except when large amounts of iSCSI storage traffic are using a single network line. Check total throughput on the NICs to see whether the network is saturated.

5. VMware Enhancements for Deployment and Operations

VMware vSphere provides core virtualization functionality. The extensive software portfolio offered by VMware is designed to help customers to achieve the ultimate goal of 100% virtualization and the software-defined data center. This section reviews some of the VMware products that can be used in an Exchange Server 2016 environment virtualized on vSphere.

5.1 VMware NSX for vSphere

Although virtualization has allowed organizations to optimize their compute and storage investments, the network has mostly remained physical. VMware NSX for vSphere solves datacenter challenges found in physical network environments by delivering software-defined network and security. Using existing vSphere compute resources, network services can be delivered quickly to respond to business challenges. VMware NSX is the network virtualization platform for the Software-Defined Data Center (SDDC). By bringing the operational model of a virtual machine to your data center network, you can transform the economics of network and security operations. NSX lets you treat your physical network as a pool of transport capacity, with network and security services attached to VMs with a policy-driven approach.

5.1.1 VMware NSX Edge

Although the Client Access Servers (CAS) role has been coupled with the Mailbox Server role in Exchange Server 2016, client access to the Exchange infrastructure still requires efficient load balancing to provide a highly available and well-performing experience for end users. To provide this functionality, hardware load balancers are deployed in front of the Mailbox servers. If the load balancer solution must be highly available, this can double the hardware investment required. In multisite deployments, this can mean up to four hardware load balancers to provide a highly available load balancing solution.

VMware NSX Edge™ provides load balancing for virtual machines through a virtual appliance. NSX Edge can be deployed in a high availability pair, providing better protection than hardware load balancing solutions without the additional hardware or management overhead. The NSX Edge load balancer enables application or service requests to be distributed across multiple back-end servers in a pool. NSX Edge Load Balancer includes the following functions:

- Virtual IP (VIP) Address: An IP address and service port number used by the user to access the service.

- Server Pool: The pool of back-end servers that need to be load balanced. A VIP address is associated with the server pool.

- Service Monitor: Defines the health-check parameters for a particular type of network traffic. A service monitor is associated with the server pool to monitor the pool members.

- Application Profile: Defines the behaviour of a particular type of network traffic. Examples: the session persistence parameter and SSL parameters.

VMware NSX 6.2 introduces support for more VIPs, from 64 to 1024 per edge. This reduces the number of edges that need to be deployed, thereby reducing the data center footprint and improving scalability in terms of manageability. VMware NSX 6.2 also introduces new improvements in terms of monitoring and failure reporting, support for port range when creating VIPs as well as tighter integration with third party load balancers. VMware NSX 6.2 also adds support for F5 integration.

Prior to VMware NSX 6.2, health monitoring checks were configured with intervals and retries. Probes were sent at configured intervals to test the connectivity with the specific application. The default interval was 5 seconds and the default settings for retries was 3. This type of configuration cannot detect intermittent failures in the load balancer pool, making troubleshooting difficult.

With VMware NSX 6.2, health monitoring is more granular. The edge can immediately report specific probe failures. The system can also track last health check and status changes that a pool has

experienced. It can also report the reason for the failures. All these improvements simplify troubleshooting.

Prior to VMware NSX 6.2, a VIP was a combination of the VIP IP address and a port. Applications required to listen on multiple ports or port ranges need the creation of multiple VIPs, one for every port. This significantly affects the scalability of the solution in terms of operations and troubleshooting and increases the number of edges that need to be deployed.

VMware NSX 6.2 adds support for listening on multiple port ranges. A VIP can be associated with multiple ports or port ranges, thereby improving the scalability and reducing the number of edges that need to be deployed.

NSX Edge supports both Layer 7 (the recommended load-balancing option without session affinity requirements in Exchange Server 2016) and Layer 4 load balancing of HTTP and HTTPS protocols. It supports multiple load balancing methods, such as round robin and least connection. Layer 7 HTTP/HTTPS VIP address are processed after the NSX Edge firewall. NSX Edge uses the faster Layer 4 load-balancer engine. The Layer 4 VIP address is processed before NSX Edge firewall.

The NSX Edge services gateway supports the following deployment models for load-balancer functionality:

- One-arm load balancer

- Inline load balancer

5.1.1.1. One Arm Load Balancer

The advantage of the one arm load balancer model is that it is simpler to deploy and flexible. It allows deployment of load balancer services (VMware NSX Edge appliances) directly on the logical segments where they are needed without requiring any modification on the centralized VMware NSX Edge that provides routing communication to the physical network. On the downside, this option requires provisioning of more VMware NSX Edge instances and mandates the deployment of source NAT that does not allow the servers in the DC to have visibility into the original client IP address.

The load balancer can insert the original IP address of the client into the HTTP header before performing source NAT (a function named Insert X-Forwarded-For HTTP header). This provides the servers visibility into the client IP address, but it is limited to HTTP traffic.

5.1.1.2. Inline Load Balancer

The inline load balancer model is also simple and allows the servers to have full visibility into the original client IP address. It is less flexible from a design perspective, as it usually forces using the load balancer as default gateway for the logical segments where the server farms are deployed and this implies that only centralized (and not distributed) routing must be adopted for those segments. The load balancer is another logical service added to the NSX Edge that is already providing routing services between the logical and the physical networks. Therefore, VMware recommends increasing the form factor of the NSX Edge to X-Large before enabling load balancing services.

Figure 22. Load-Balancing Exchange Server 2016 with NSX Edge
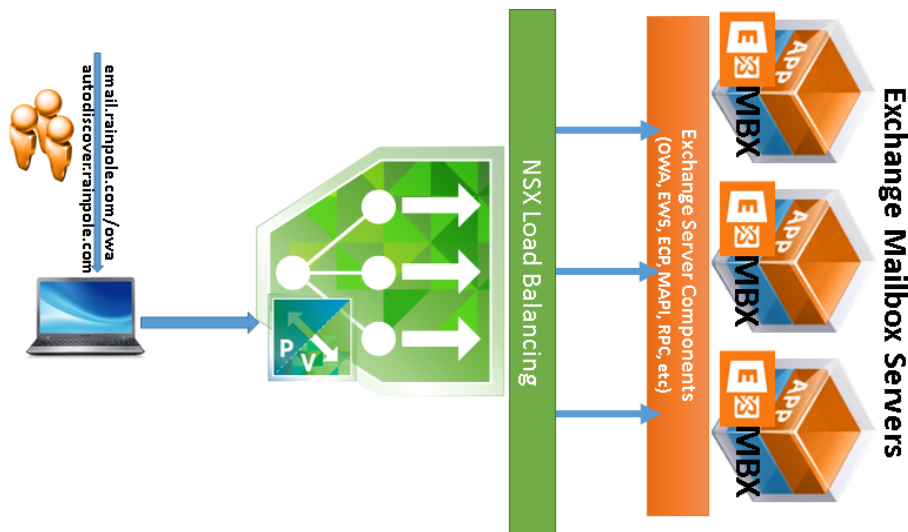
5.1.2   VMware NSX Distributed Firewall

Exchange 2013 can leverage VMware NSX for vSphere to provide application layer isolation against unauthorized access. Isolation at this level typically requires hardware firewalls and multiple VLANs in a physical networking environment. With NSX Distributed Firewall, this capability is delivered in software through in kernel module.

The distributed firewall provides security filtering and service chaining functions on every host prepared for VMware NSX. The throughput scales as hypervisors are added.

The distributed firewall kernel module is embedded in the VMkernel. It is deployed as a VMware infrastructure bundle (VIB) during the host preparation process.

The firewall rules are configured by the administrator using the vSphere Web Client or REST APIs. The NSX Manager stores the rules in its database and pushes the policies to the ESXi hosts through the message bus agent.

You can create different types of firewall rules with VMware NSX.

Application-aware policies can use dynamic tagging and user identity. Infrastructure-aware policies can use vCenter inventory objects as the basis for rules. Network-based policies are the most common (and traditional) types of policies that use Layer 2 and Layer 3 data for the rules.

Note   If you have vShield App version 5.5, you can upgrade to NSX. If you have a prior version of vShield App in your infrastructure, you must upgrade to version 5.5 before upgrading to version 6.0.

After you upgrade firewall to NSX, move the grouping objects used by firewall rules to global scope. To do this, use NSX APIs to create new grouping objects with the same members and then update the relevant firewall rules with the new IDs.

**Figure 23. NSX Distributed Firewall Capability**
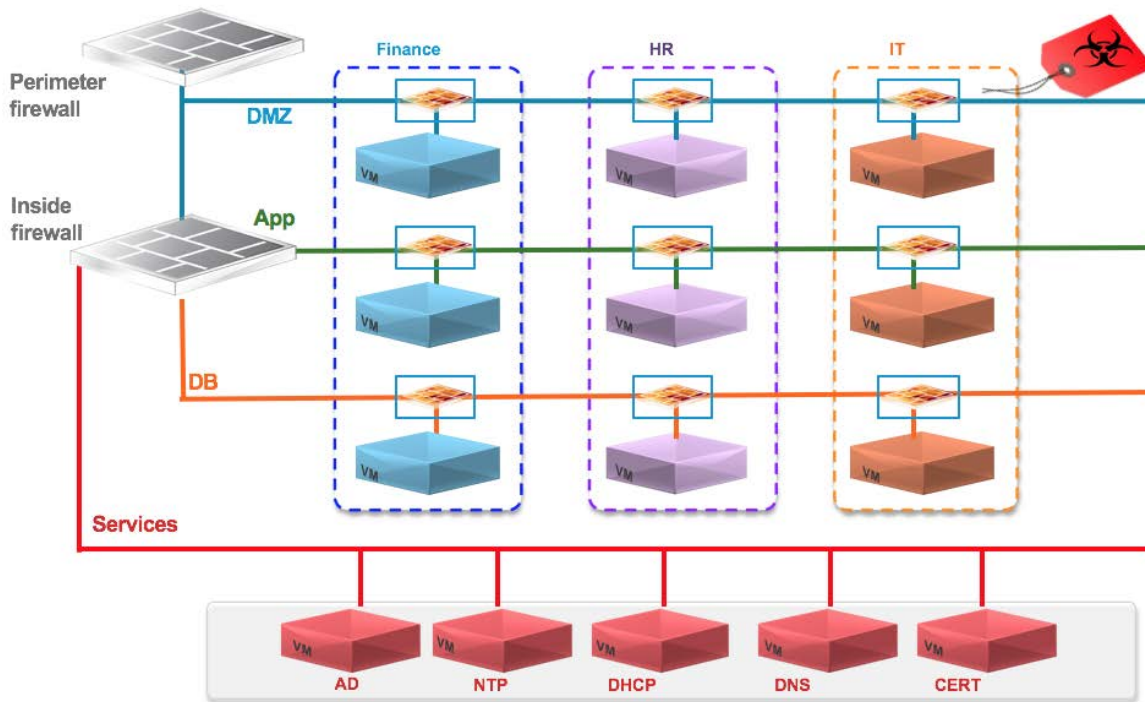
5.2   VMware vRealize Operations Manager

VMware vRealize® Operations Manager™ can provide a holistic approach to performance, capacity, and configuration management. By using patented analytics, service levels can be proactively monitored and maintained. When performance or capacity problems arise in your Exchange environment, vRealize Operations Manager is able to analyze metrics from the application all the way through to the infrastructure to provide insight into problematic components, whether they are compute (physical or virtual), storage, networking, OS, or application related. By establishing trends over time, vRealize Operations Manager can minimize false alerts and proactively alert on the potential root cause of increasing performance problems before end users are impacted.

In an Exchange environment, constant monitoring is required to maintain acceptable service levels, not only for end users, but also for the Exchange components. vRealize Operations Manager includes patented capacity analytics that can eliminate the need for spreadsheets, scripts, or rules of thumb. Quickly run through "what if" capacity scenarios to understand growth trends and identify upcoming compute power shortages or over-provisioned resources. As an application comprising multiple components, Exchange performance and functionality can be affected by changes made at many levels. vRealize Operations Manager monitors configurations across virtual machines and detects unwanted changes to help maintain continuous compliance with operational best practices.

Figure 24. vRealize Operations
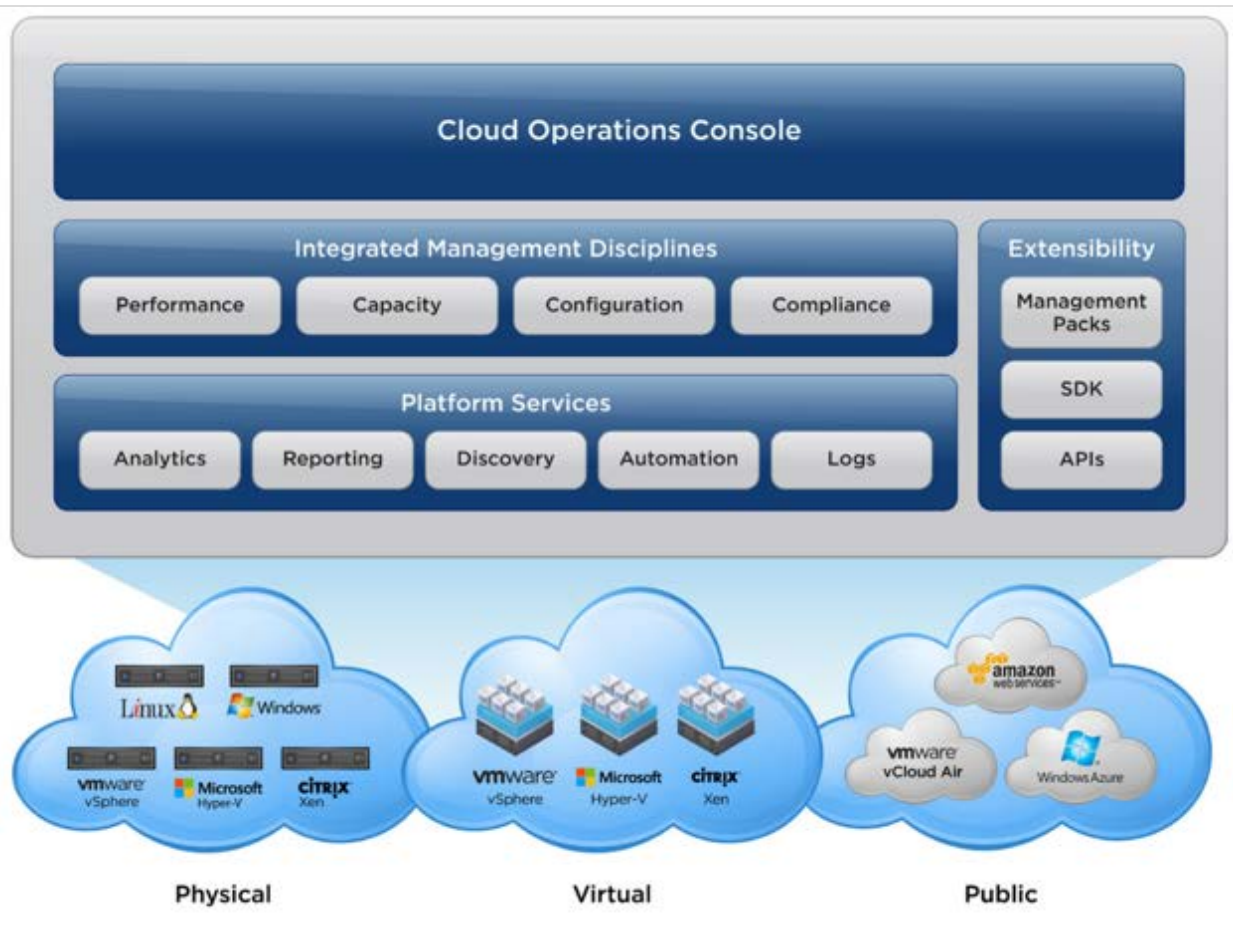
5.3 Site Recovery Manager

Site Recovery Manager takes advantage of virtual machine encapsulation to make testing, validating and initiating disaster recovery (DR) failover a straightforward, integrated vCenter process. Site Recovery Manager runs alongside vCenter Server to provide planning, testing, and automated recovery in the case of a disaster. By using VMware vSphere Replication or storage-based replication technology, Site Recovery Manager eliminates the manual steps required during a failover scenario to provide consistent and predictable results.

The following is a high-level overview of steps that can be performed with Site Recovery Manager during a test or actual failover.

- Shut down production virtual machines, if the production site is still reachable (actual failover).

- Promote recovery storage to primary (actual failover).

- Take and mount a snapshot of recovery storage in read/write mode (test failover).

- Rescan recovery site ESXi hosts to make storage visible.

- Register recovery virtual machines, using placeholder (VMX) files representing protect VMs at DR site.

- Power on registered virtual machines at the recovery site, in the exact order specified in the recovery plan.

- Using information contained in recovery plan, reconfigure VM IP addresses, if required.

- If configured, recovery step pauses for external administrator's tasks.

- Continue with recovery steps upon completion of administrator's actions.

- Verify that VMware Tools starts successfully on recovered virtual machines.

- Execute any in-guest (or SRM server-hosted) scripts and commands specified in the recovery plan.

- Notify administrators about completion.

- Power off recovered virtual machines (test failover).

- Unregister virtual machines (test failover).

- Remove storage snapshot from the recovery side (test failover).

- Provide option to configure protection for recovered VMs, as soon as failed site becomes operational, or to another surviving site.

Exchange Server's native high availability feature (DAGs) can provide high availability by implementing database-level intra- and inter-site database replication for all (or some) Exchange Server databases. Although DAG is an excellent choice for data center high availability, the application-centric nature of a DAG might not be in line with a company's disaster recovery plans. In addition, configuring DAG for the purposes of timely and optimal recovery of Exchange services in the event of a catastrophic data center failure is complex, costly and less reliable than leveraging the features and capabilities of the VMware Site Recovery Manager.

Site Recovery Manager is not a replacement for application-aware clustering solutions (such as DAG) that might be deployed within the guest operating system. Site Recovery Manager provides integration of the replication solution, vSphere, and optionally customer-developed scripts to provide a simple, repeatable, and reportable process for disaster recovery of the entire virtual environment, regardless of the application. Site Recovery Manager complements and enhances Exchange DAG capabilities by streamlining, automating and optimizing recovery operations in the event of a site-level disaster.

Figure 25. VMware Site Recovery Manager – Logical Components
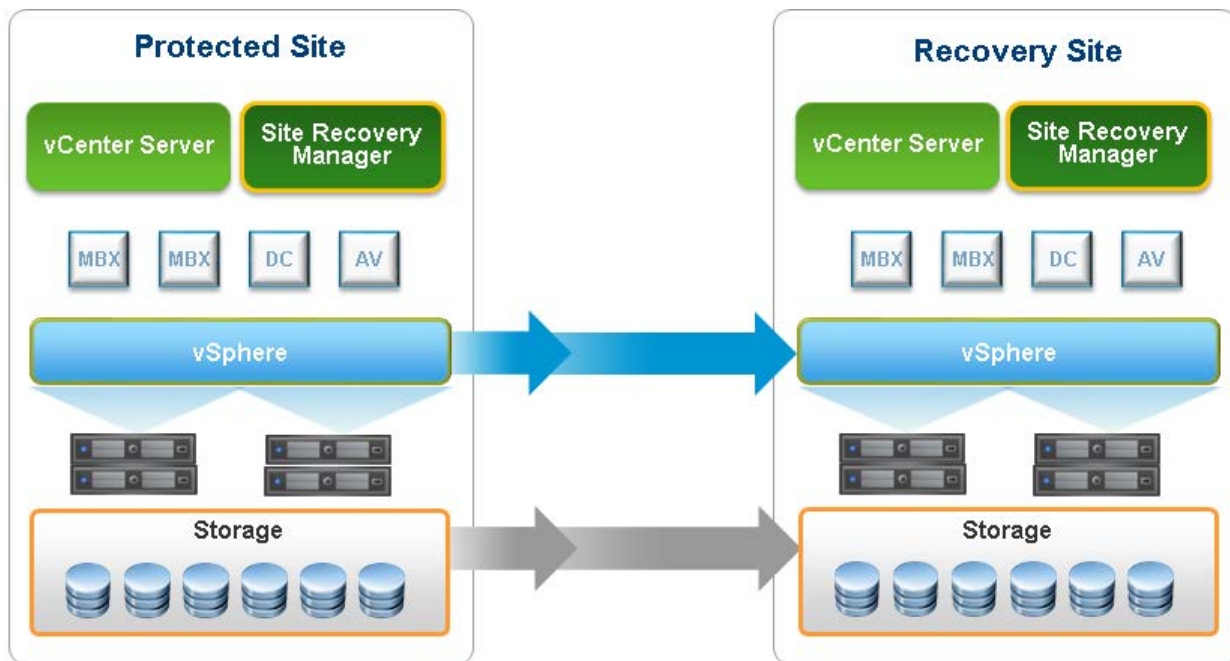
The *Exchange 2016 Preferred Architecture* (http://blogs.technet.com/b/exchange/archive/2015/10/12/the-exchange-2016-preferred-architecture.aspx) prescribes a minimum of four DAG copies AND three geographically-dispersed data centers to achieve a semblance of site resilience and disaster recovery with DAG. Even with these requirements, the following impediments still make Site Recovery Manager a superior DR solution over DAG:

- No testing capability – Numerous changes happen over the life of a given IT infrastructure. Some of these changes invalidate previous configurations, scripts and processes, requiring iterative updates and testing. A disaster recovery plan requires reliability and assurance because an actual disaster event is not the best time to discover that a previously-configured recovery plan has been invalidated by later infrastructure changes. Site Recovery Manager enables continuous, periodic testing and reconfiguration of recovery plans without inducing interruption or service outage in the Exchange Server infrastructure. Simulating recovery from a site disaster event with Exchange DAG requires service interruption for the duration of the simulation exercise. Post-simulation, returning the Exchange infrastructure to its prior state – a one-click operation with Site Recovery Manager – is also a complex undertaking, requiring multiple steps and a lengthy database reseeding operation.

- Cost efficiency – Site Recovery Manager is more cost-efficient, both in terms of administrative efforts and financial costs. Each of the four Exchange Servers required to satisfy the preferred architecture design requirements needs its own server hardware, OS, Exchange, antivirus and other application licenses, in addition to the storage to support the configuration. Even when configured with DAG as prescribed, the associated administrative, management and maintenance efforts required to support the design can become quickly overwhelming and prohibitive. With Site Recovery Manager, you can achieve a better DR solution with just a two-member DAG configuration, providing a better and less costly alternative to the preferred architecture design.

- Unified DR Solution – Exchange DAG is a high availability solution for Exchange Servers and only Exchange Servers. An Exchange Server instance has multiple dependencies (Active Directory, backup and other messaging hygiene and security components, and so on), which are required to be available before Exchange Server services can be successfully recovered in the event of a site disaster. Recovering Exchange Servers alone in this configuration does not add much value unless the dependencies themselves are recovered as well. Site Recovery Manager is application-agnostic

and suitable for protecting and recovering any server, including the dependencies that DAG cannot protect. The unification of a disaster recovery solution improves efficiency and reduces costs. Administrators do not have to manage multiple DR solutions, removing confusion, complexity and stress in an otherwise stressful situation such as a DR event.

As stated by Microsoft:

> *"The specific prescriptive nature of the PA means of course that not every customer will be able to deploy it (for example, customers without multiple data centers). And some of our customers have different business requirements or other needs which necessitate a different architecture. If you fall into those categories, and you want to deploy Exchange on-premises, there are still advantages to adhering as closely as possible to the PA, and deviate only where your requirements widely differ".*

Disaster recovery for a production, enterprise-level Exchange Server infrastructure is a critical design consideration that requires deviation from Microsoft's prescriptive guidance.

The following is a high-level overview of Site Recovery Manager as a DR solution for an Exchange Server infrastructure. See the *Exchange Server High Availability and Disaster Recovery on VMware vSphere Guide* (http://www.vmware.com/files/pdf/Exchange_2013_on_VMware_Availability_and_Recovery_Options.pdf) for a detailed discussion of this topic.

**Figure 26. Challenges with Exchange Server DAG as a DR Solution**

Using Site Recovery Manager to protect Exchange Server components (including DAG) and infrastructure is a fully supported configuration. Because Site Recovery Manager is application-agnostic, it does not interfere with, modify or otherwise affect Exchange Servers. Site Recovery Manager is not involved in the replication of the VM files and data from a protected site to a recovery site. This function is performed by the applicable storage replication components of the customer's choosing. Site Recovery Manager provides the necessary storage API (storage replication adapter) required for to interact with the underlying storage infrastructure. Site Recovery Manager does not need to install any agent or components on the virtual machine.

Site Recovery Manager adds automation and orchestration capabilities to a virtual infrastructure, affording customers the ability to configure a comprehensive recovery plan that includes every facet of

the recovery steps and actions required to restore services to an Exchange infrastructure in a DR scenario. Site Recovery Manager includes the capability to pause a recovery operation to allow for manual administrative intervention where required (for example, reconfigure DNS records or load balancer configuration in a non-stretched network) as well as script callouts.

Figure 27. Faster Exchange Service Recovery with Site Recovery Manager Automated DR Workflows
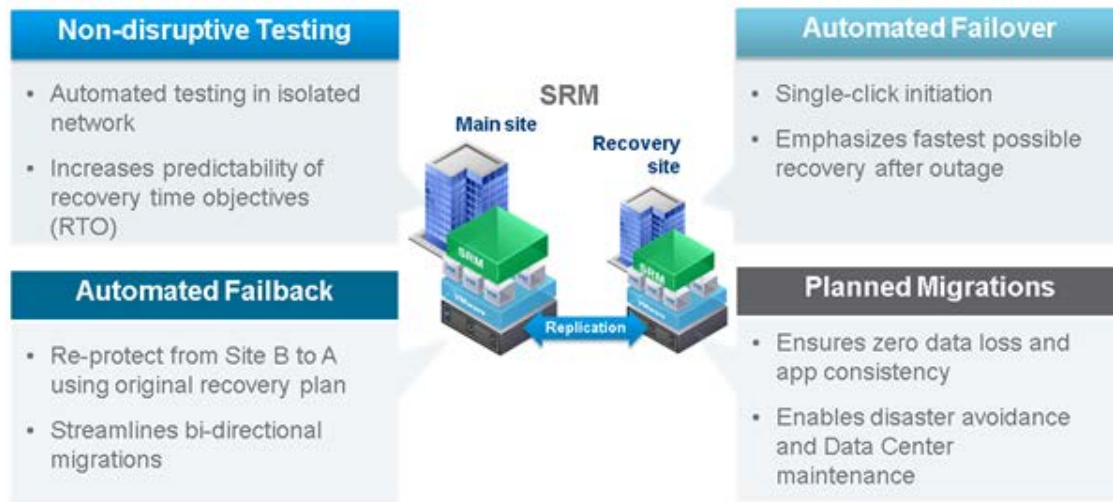
Site Recovery Manager supports all the features of a vSphere infrastructure, including, DRS, vSphere HA, Fault Tolerance, and Virtual SAN. vMotion support includes storage and cross-data center vSphere vMotion operations. While Site Recovery Manager supports configuring an isolated test network for testing disaster recovery plan, it does not require such configuration – a test failover operation auto-generates the fenced network that is required to isolate the recovered Exchange infrastructure from the production environment.

Site Recovery Manager provides – multiple topologies and recovery options for protecting an organization's Exchange Server infrastructure:

- Active-Passive – Site Recovery Manager supports the traditional active-passive DR scenario, where a production site running applications is recovered at a second site that is idle until failover is required. Although the most common configuration, this scenario also means that you are paying a lot of money for a DR site that is idle most of the time.

- Active-Active – To make better use of the recovery site, Site Recovery Manager also enables you to leverage your recovery site for other workloads when you aren't using it for DR. Site Recovery Manager can be configured to automatically shut down or suspend VMs at the recovery site as part of the failover process so that you can easily free up compute capacity for the workloads being recovered.

- Bidirectional – Site Recovery Manager can also provide bidirectional failover protection so that you can run active production workloads at both sites and failover to the other site in either direction. The spare capacity at the other site will be used to run the VMs that are failed over.

- Shared Recovery Sites – Although less common, some of customers need to be able to failover within a given site or campus, for example when a storage array failure occurs or when building maintenance forces movement of workloads to a different campus building.

- Active-Active Datacenters – This is a new topology supported with metro-distance stretched storage solutions. Production apps run at both sites, and the stretched storage provides synchronous reads

and writes on storage when sites are within a metro distance (less than 100 km). Site Recovery Manager is used to orchestrate recover or even live migration of VMs between sites.

Figure 28. Failover Scenarios with Site Recovery Manager