DELL Technologies

Best Practices

Dell EMC PowerVault ME4 Series and VMware vSphere

Abstract

This document provides best practices for deploying VMware® vSphere® with Dell EMC™ PowerVault™ ME4 Series storage. It includes configuration recommendations for vSphere hosts to achieve an optimal combination of performance and resiliency.

October 2020

Revisions

Date	Description
September 2018	Initial release
March 2020	Minor revisions
October 2020	Adjusted claim rules

Acknowledgments

Author: Darin Schmitz

Table of contents

Executive summary

This document provides best practices for VMware® vSphere® when using a Dell EMC™ PowerVault™ ME4 Series storage array. It does not include sizing, performance, or design guidance, but it provides information about the features and benefits of using ME4 Series arrays for VMware vSphere environments.

VMware vSphere is an extremely robust, scalable, enterprise-class hypervisor. Correctly configured using the best practices presented in this paper, the vSphere ESXi® hypervisor provides an optimized experience with ME4 Series storage. These recommendations include guidelines for SAN fabric design, HBA settings, and multipath configuration. There are often various methods for accomplishing the described tasks, and this paper provides a starting point for end users and system administrators but is not intended to be a comprehensive configuration guide.

Audience

This document is intended for PowerVault ME4 Series administrators, system administrators, and anyone responsible for configuring ME4 Series systems. It is assumed the readers have prior experience with or training in SAN storage systems and a VMware vSphere environment.

1 Introduction

The PowerVault ME4 Series is next-generation, entry-level storage that is purpose-built and optimized for SAN and DAS virtualized workloads. Available in 2U or dense 5U base systems, the low-cost ME4 Series simplifies the challenges of server capacity expansion and small-scale SAN consolidation with up to 336 drives or 4 PB capacity. It also comes with all-inclusive software, incredible performance, and built-in simplicity with a new web-based HTML5 management UI, ME Storage Manager. Connecting ME4 Series storage to a PowerEdge server or to a SAN ensures that business applications will get high-speed and reliable access to their data—without compromise.

Product features include the following:

Simplicity: ME4 Series storage includes a web-based management UI (HTML5), installs in 15 minutes, configures in 15 minutes, and easily deploys in 2U or 5U systems.

Performance: Compared to the predecessor MD3 Series, the ME4 Series packs a lot of power and scale with the Intel® Xeon® Processor D-1500 product family. The ME4 Series processing power delivers incredible performance gains over the MD3 Series, as well as increased capacity, bandwidth, and drive count.

Connectivity: ME4 Series storage goes to the next level with robust and flexible connectivity starting with a 12Gb SAS back-end interface, and a front-end interface options including four 16Gb FC ports per controller, four 10Gb iSCSI ports per controller (SFP+ or BaseT), or four 12Gb SAS ports per controller.

Scalability: Both 2U and 5U base systems are available, with the 2U system supporting either 12 or 24 drives and the 5U system supporting 84 drives. Each of the 2U (ME4012 and ME4024) and 5U (ME4084) base systems supports optional expansion enclosures of 12, 24, and 84 drives, allowing you to use up to 336 drives. Drive mixing is also allowed.

All-inclusive software: ME4 Series software provides volume copy, snapshots, IP/FC replication, VMware® vCenter Server® and VMware Site Recovery Manager™ integration, SSD read cache, thin provisioning, three-level tiering, ADAPT (distributed RAID), and controller-based encryption (SEDs) with internal key management.

Management: An integrated HTML5 web-based management interface (ME Storage Manager) is included.

For more information, see the ME4 Series product page.

2 ME4 Series features

Although the ME4 Series is targeted at the entry level of the SAN market, it contains many advanced and enterprise-class features detailed in the following sections. It is recommended that both the storage administrator and the VMware administrator have a solid understanding of how these storage features can benefit the vSphere environment prior to deployment.

Note: The ME4 Series array uses the term Virtual Volume, which is not associated with the VMware vSphere Virtual Volumes™ feature.

2.1 Virtual and linear storage

ME4 Series arrays use two storage technologies that share a common user interface: the virtual method and the linear method.

The linear method maps logical host requests directly to physical storage. In some cases, the mapping is one-to-one, while in most cases, the mapping is across groups of physical storage devices, or slices of them. While the linear method of mapping is highly efficient, it lacks flexibility. This makes it difficult to alter the physical layout after it is established.

The virtual method maps logical storage requests to physical storage (disks) through a layer of virtualization, such that logical host I/O requests are first mapped onto pages of storage and then each page is mapped onto physical storage. Within each page, the mapping is linear, but there is no direct relationship between adjacent logical pages and their physical storage. A page is a range of contiguous logical block addresses (LBAs) in a disk group, which is one of up to 16 RAID sets that are grouped into a pool. Thus, a virtual volume as seen by a host represents a portion of storage in a pool. Multiple virtual volumes can be created in a pool, sharing its resources. This allows for a high level of flexibility, and the most efficient use of available physical resources.

Some advantages of using virtual storage include the following:

- It allows performance to scale as the number of disks in the pool increases.
- It virtualizes physical storage, allowing volumes to share available resources in a highly efficient way.
- It allows a volume to be comprised of more than 16 disks.

Virtual storage provides the foundation for data-management features such as thin provisioning, automated tiered storage, read cache, and the quick disk rebuild feature. Because these storage features are valuable in most environments, virtual storage is recommended when deploying VMware vSphere environments. Linear storage pools are most suited to sequential workloads such as video archiving.

2.2 Enhancing performance with minimal SSDs

While the cost of SSDs continues to drop, there is still a significant price gap between SSDs and traditional spinning HDDs. Not all environments require the performance of an all-flash ME4 Series array, however, ME4 Series arrays can use a small number of SSD drives to gain a significant performance increase. Both the automated tiered storage and read flash cache features of the ME4 Series array use a small number of SSDs to provide a significant performance boost to a traditional low-cost, all-HDD SAN solution.

2.2.1 Automated tiered storage

Automated tiered storage (ATS) automatically moves data residing in one class of disks to a more appropriate class of disks based on data access patterns, with no manual configuration necessary. Frequently accessed, hot data can move to disks with higher performance, while infrequently accessed, cool data can move to disks with lower performance and lower costs.

Each virtual disk group, depending on the type of disks it uses, is automatically assigned to one of the following tiers:

Performance: This highest tier uses SSDs, providing the best performance but also the highest cost.

Standard: This middle tier uses enterprise-class SAS hard drives, which provide good performance with mid-level cost and capacity.

Archive: This lowest tier uses nearline SAS hard drives, which provide the lowest performance with the lowest cost and highest capacity.

A volume's tier affinity setting enables tuning the tier-migration algorithm when creating or modifying the volume so that the volume data automatically moves to a specific tier, if possible. If space is not available in a volume's preferred tier, another tier will be used. There are three volume tier affinity settings:

No affinity: This is the default setting. It uses the highest available performing tiers first and only uses the archive tier when space is exhausted in the other tiers. Volume data will swap into higher performing tiers based on frequency of access and tier space availability.

Archive: This setting prioritizes the volume data to the lowest performing tier available. Volume data can move to higher performing tiers based on frequency of access and available space in the tiers.

Performance: This setting prioritizes volume data to the higher performing tiers. If no space is available, lower performing tier space is used. Performance affinity volume data will swap into higher tiers based on frequency of access or when space is made available.

2.2.2 Read flash cache

Unlike tiering, where a single copy of specific blocks of data resides in either spinning disks or SSDs, the read flash cache feature uses one or two SSD disks per pool as a read cache for hot or frequently read pages only. Read cache does not add to the overall capacity of the pool to which it has been added, nor does it improve write performance. Read flash cache can be added from the pool without any adverse effect on the volumes and their data in the pool, other than to impact the read-access performance. A separate copy of the data is always maintained on the HDDs. Taken together, these attributes have several advantages:

- Controller read cache is effectively extended by two orders of magnitude or more.
- The performance cost of moving data to read-cache is lower than a full migration of data from a lower tier to a higher tier.
- Read-cache is not fault tolerant, lowering system cost.

2.3 Asymmetric Logical Unit Access

ME4 Series storage uses Unified LUN Presentation (ULP), which can expose all LUNs through all host ports on both controllers. The storage system appears as an active-active system to the host. The host can choose any available path to access a LUN regardless of disk-group ownership. When ULP is in use, the controllers'

operating/redundancy mode is shown as active/active ULP. ULP uses the Asymmetric Logical Unit Access (ALUA) extensions to negotiate paths with the ALUA-aware operating systems. If the hosts are not ALUA-aware, all paths are treated as equal even though some paths might have better latency than others.

vSphere ESXi is an ALUA aware operating system, and no additional configuration is required. Each datastore will have two, four, or eight active paths depending upon controller configuration (SAS, combined FC/iSCSI controller, or dedicated FC/iSCSI) with half of the paths identified as active optimized and the other half identified as active non-optimized.

2.4 RAID data protection levels

ME4 Series arrays support RAID data protection levels NRAID, 0, 1, 10, 3, 5, 50, 6 and ADAPT. ADAPT is a special RAID implementation that offers some unique benefits. It can withstand two drive failures with very fast rebuilds. Spare capacity is distributed across all drives instead of dedicated spare drives. ADAPT disk groups can have up to 128 drives and allow mixing different drive sizes. Data is stored across all disks evenly. The storage system automatically rebalances the data when new drives are added or when the distribution of data has become imbalanced.

It is recommended to choose the right RAID level that best suits the type of workloads in the environment. Review the information in ME4 Series Administrator's Guide on Dell.com/support which details the benefits of each RAID level, the minimum and maximum disks requirements, and the recommendation of RAID levels for popular workloads.

3 Connectivity considerations

ME4 Series storage supports and is certified with VMware vSphere for server connectivity with iSCSI (1 Gb and 10 Gb), Fibre Channel (8 Gb and 16 Gb, direct-attached and SAN-attached), and direct-attached SAS. While the PowerVault ME4012 or ME4024 array can be configured with a single controller, for maximum storage availability and performance, it is a best practice to use dual controller configurations. A dual-controller configuration improves application availability because in the event of a controller failure, the affected controller fails over to the partner controller with little interruption to data flow. A failed controller can be replaced without the need to shut down the storage system.

3.1 Direct-attached storage

ME4 Series arrays support direct-attached Fibre Channel (8 Gb and 16 Gb) and direct-attached SAS connectivity. Using direct-attached hosts removes the financial costs associated with a SAN fabric from the environment but limits the scale to which the environment can grow. While ME4 Series storage can support up to eight direct-attached servers, this is achieved by providing only a non-redundant, single connection to each server. As a best practice, each host should have a dual-path configuration with a single path to each controller, enabling storage access to continue in the event of controller failure. This limits the number of direct-attached servers to four but enables controller redundancy and increased performance.

Figure 1 shows a configuration with four servers, each with two Fibre Channel or SAS connections to the ME4 Series array.

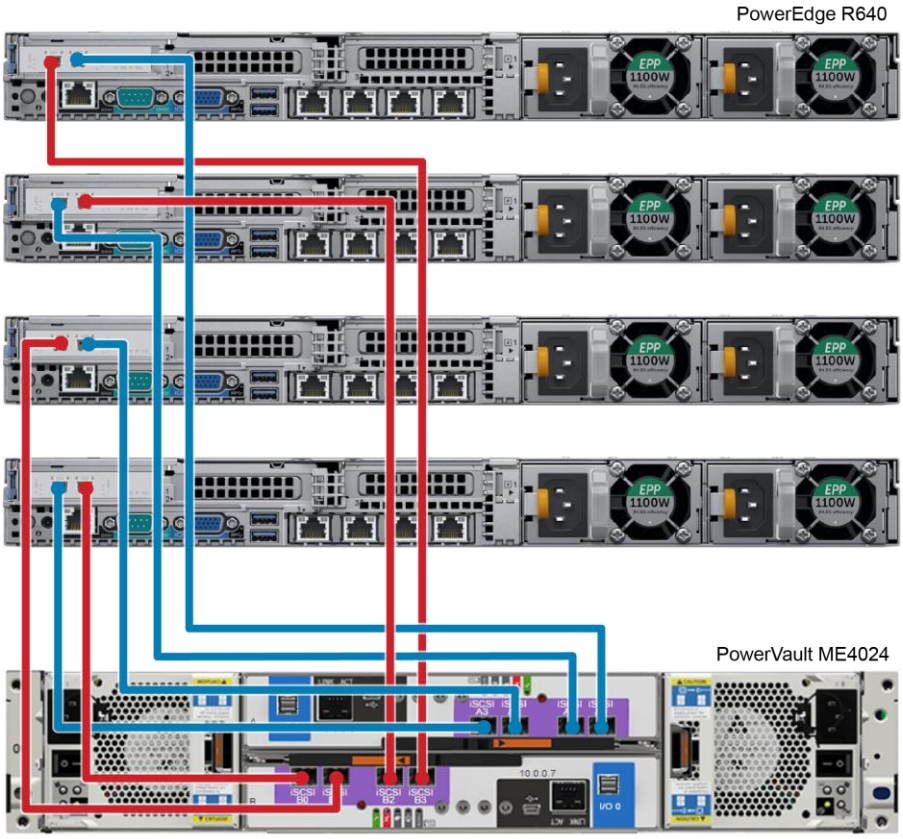

Figure 1 Connecting four hosts directly to a PowerVault ME4024 array with dual paths

3.2 SAN-attached storage

ME4 Series arrays support SAN-attached Fibre Channel (8 Gb and 16 Gb) and iSCSI (10 Gb and 1 Gb) connectivity. A switch-attached solution (or SAN) places a Fibre Channel or Ethernet switch between the servers and the controller enclosures within the storage system. Using switches, a SAN shares a storage system among multiple servers reducing the number of storage systems required for a particular environment. Using switches increases the number of servers that can be connected to the storage system to scale to greater than four servers, which is the limit for a direct-attached environment.

When designing a SAN, using two switches is recommended. This enables the creation of a redundant transport fabric between the server and the ME4 Series storage, and allows an individual switch to be taken out of service for maintenance or due to failure without impacting the availability of access to the storage.

When cabling the ME4 Series controllers in a switched environment, pay close attention to the layout of the cables in both Fibre Channel and Ethernet fabrics. In Figure 2, controller A (the left-most ME4084 controller) has ports 0 and 2 connected to the top switch, and ports 1 and 3 are connected to the bottom switch, which is repeated in a similar fashion with controller B. The servers are configured with each server having connections to each switch. This cabling ensures that access to storage remains available between an individual server and the ME4 Series array during switch maintenance.

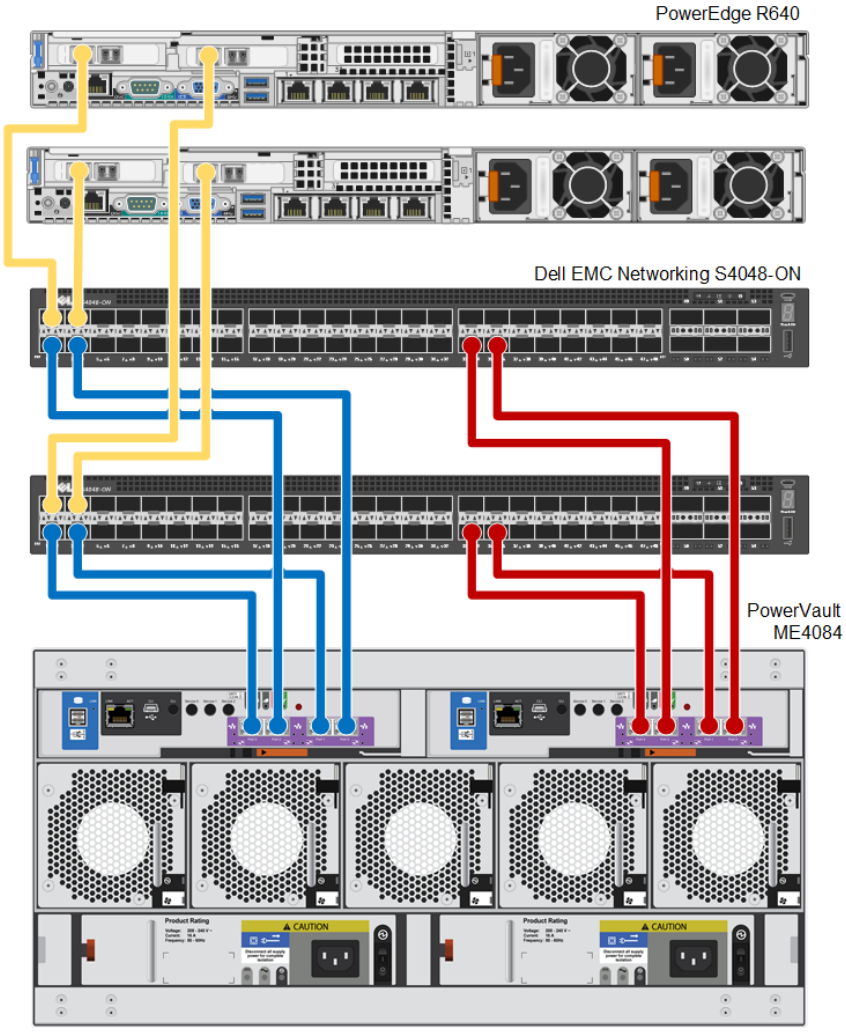

Figure 2 Connecting two hosts to an ME4084 array using two switches

3.2.1 iSCSI fabric settings

This section details recommended and required settings when creating an iSCSI-based SAN.

Note: 1 Gb iSCSI is supported only with the 10GBaseT controller and not the converged network controller.

3.2.1.1 Flow control settings

Ethernet flow control is a mechanism for temporarily pausing data transmission when data is being transmitted faster than its target port can accept the data. Flow control allows a switch port to stop network traffic sending a PAUSE frame. The PAUSE frame temporarily pauses transmission until the port is again able to service requests.

The following settings are recommended when enabling flow control:

- A minimum of receive (RX) flow control should be enabled for all switch interfaces used by servers or storage systems for iSCSI traffic.
- Symmetric flow control should be enabled for all server interfaces used for iSCSI traffic. ME4 Series automatically enables this feature.

3.2.1.2 Jumbo frames

Jumbo frames increase the efficiency of Ethernet networking and reduce CPU load by including a larger amount of data in each Ethernet packet. The default Ethernet packet size, or MTU (maximum transmission unit), is 1,500 bytes. With Jumbo frames, this is increased to 9,000 bytes.

Note: PowerVault ME4 Series storage supports a maximum 8900-byte payload, allowing 100 bytes of overhead for the MTU of 9000.

When enabling Jumbo frames, all devices in the path must be enabled for Jumbo frames for this frame size to be successfully negotiated. This included server NICs or iSCSI HBAs, switches, and the ME4 Series storage. In a vSphere environment, this also included the virtual switches and VMkernel adapters configured for iSCSI traffic.

To enable Jumbo frames on the ME4 Series system, click **System Settings > Ports > Advanced Settings** and select the **Enable Jumbo Frames** check box.

3.2.1.3 Jumbo frames and flow control

Some switches have limited buffer sizes and can support either Jumbo frames or flow control, but cannot support both at the same time. If you must choose between the two features, it is recommended to choose flow control.

Note: All switches listed in the Dell EMC Storage Compatibility Matrix support Jumbo frames and flow control at the same time.

3.2.2 Fibre Channel zoning

Fibre Channel zones are used to segment the fabric to restrict access. A zone contains paths between initiators (server HBAs) and targets (storage array front-end ports). Either physical ports (port zoning) on the Fibre Channel switches or the WWNs (name zoning) of the end devices can be used in zoning. It is recommended to use name zoning because it offers better flexibility. With name zoning, server HBAs and storage array ports are not tied to specific physical ports on the switch.

Zoning Fibre Channel switches for vSphere ESXi hosts is essentially no different than zoning any other hosts to the ME4 Series array.

Zoning rules and recommendations:

- The ME4 Series array and ESXi hosts should be connected to two different Fibre Channel switches (fabrics) for high availability and redundancy.
- Name zoning using WWNs is recommended.
- When defining the zones, it is a best practice to use single-initiator (host port), multiple-target (ME4 ports) zones. For example, for each Fibre Channel HBA port on the server, create a server zone that includes the HBA port WWN and all the physical WWNs on the ME4 Series array controllers on the same fabric. See Table 1 for an example.

Table 1 Fibre Channel zoning examples

Fabrics **(dual-switch configuration)**	FC HBA port **(dual-port HBA configuration)**	ME4 FC ports **(FC port configuration)**
Fabric one zone	Port 0	A0, B0, A2, B2
Fabric two zone	Port 1	A1, B1, A3, B3

Note: It is recommended to use name zoning and create single-initiator, multiple-target zones.

3.3 Physical port selection

In a system configured to use all FC or all iSCSI, but where only two ports are needed, use ports 0 and 2 or ports 1 and 3 to ensure better I/O balance on the front end. This is because ports 0 and 1 share a converged network controller chip, and ports 2 and 3 share a separate converged network controller chip.

4 Host bus adapters

This section provides host bus adapter (HBA) information for SAS, Fibre Channel, and iSCSI cards that provide the most effective communication between the server and the ME4 Series array.

4.1 Fibre Channel and SAS HBAs

To obtain drivers for the Fibre Channel or 12 Gb SAS HBAs shipped in 13th-generation and 14th-generation Dell EMC PowerEdge™ servers, download the Dell-customized ESXi embedded ISO image from Dell Support. The drivers are fully compatible with the ME4 Series array and do not require further configuration.

4.2 iSCSI HBAs

The ME4 Series array is only certified with the vSphere ESXi software iSCSI initiator. No dependent, independent, or iSCSI offload cards are supported.

5 ME4 Series array settings

This section includes ME4 Series array settings that ensure a smooth and consistent data-center environment.

5.1 Missing LUN Response

The setting for **Missing LUN Resp**onse can be found in ME Storage Manager under **Action > Advanced Settings > Cache**.

The default setting of **Illegal Request** is compatible with a VMware vSphere environment and should not be changed. Some operating systems do not look beyond LUN 0 if they do not find a LUN 0, or cannot work with noncontiguous LUNs. This parameter addresses these situations by enabling the host drivers to continue probing for LUNs until they reach the LUN to which they have access. This parameter controls the SCSI sense data returned for volumes that are not accessible because they do not exist or have been hidden through volume mapping.

In a vSphere environment, ESXi interprets the **Not Ready** reply as a temporary condition. If a LUN is removed from an ESXi host without properly un-mounting the datastore first, and if the missing LUN response is set to Not Ready, ESXi may continue to query for this LUN indefinitely.

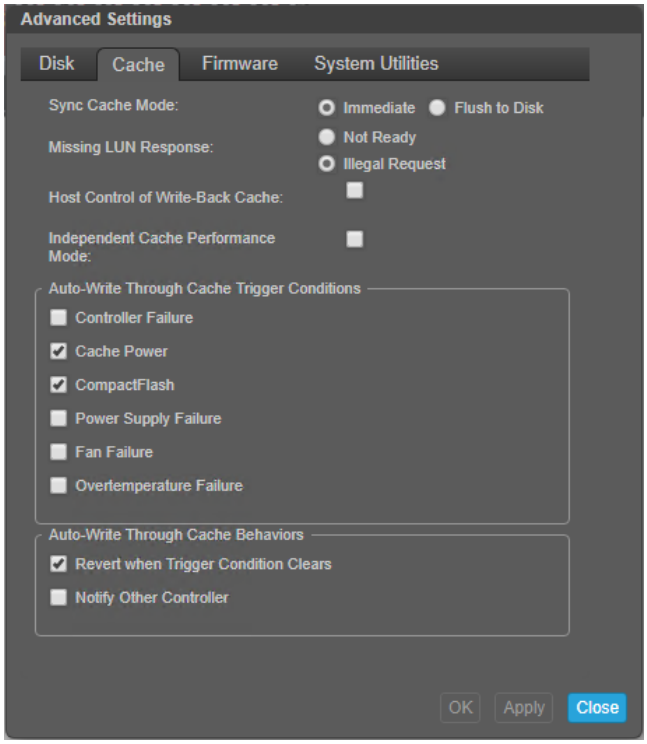

Figure 3 Missing LUN Response setting

5.2 Host groups

For ease of management with ME4 Series arrays, initiators that represent a server can be grouped into an object called a host, and multiple host objects can be organized into an object called a host group. Doing so enables mapping operations for all initiators and hosts in a group to be performed in one step, rather than mapping each initiator or host individually. A maximum of 32 host groups can exist.

5.3 Log file timestamps

Debugging and troubleshooting any data-center issue involves reviewing multiple log files from different sources. Tracking an issue across multiple log files, whether manually of through a third-party log file aggregator, depends upon accurate timestamp information. Ensure that the various components that make up the vSphere environment use the same NTP time source and time zone off set. These settings can be found in ME Storage Manager under Action > System Settings > Date and Time.

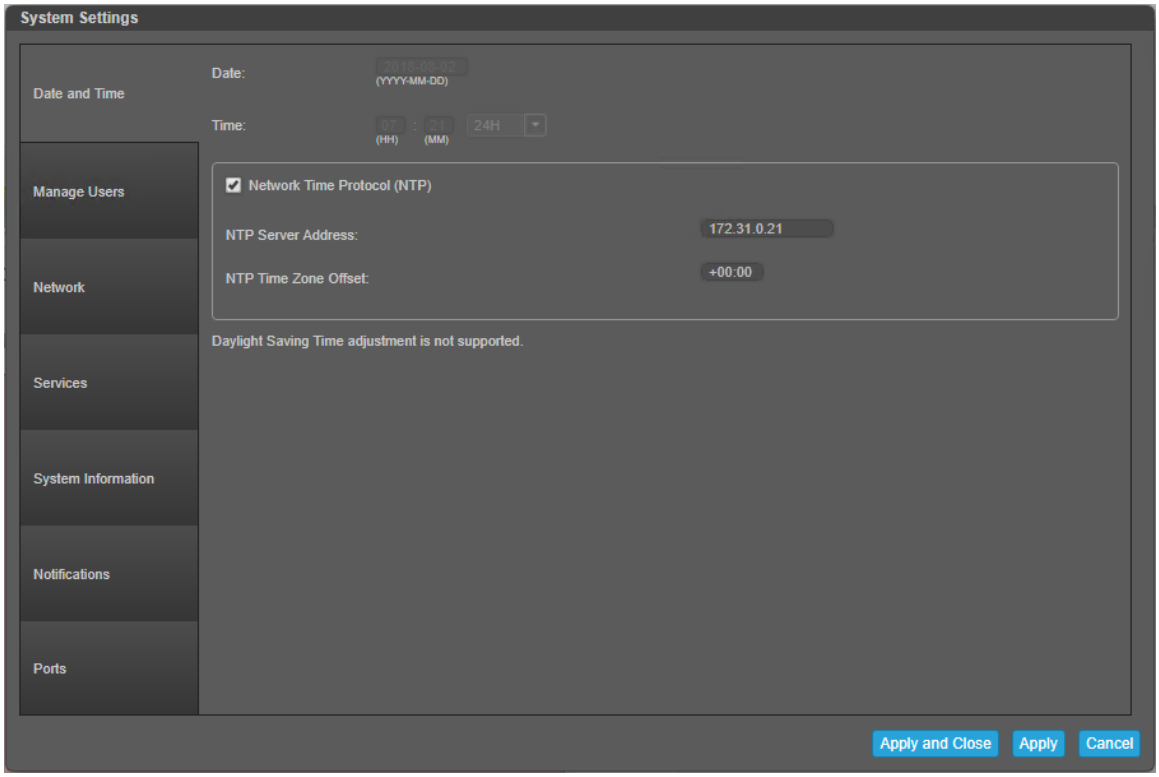

Figure 4 Network Time Protocol system settings

6 VMware vSphere settings

The following configuration settings are recommended for the VMware ESXi™ hosts.

6.1 Recommended iSCSI vSwitch configuration

To configure the VMware iSCSI software initiator for multipathing, see the VMware articles, Configuring Software iSCSI Adapter and Setting Up iSCSI Network located at VMware Docs.

For users with previous experience configuring iSCSI for multipathing, here are a few key points for a successful configuration:

- Create two VMkernel adapters, one for each SAN fabric.
- Make sure each VMkernel adapter is on its own vSwitch with a single vmnic (physical NIC adapter).
- Remember if Jumbo frames is used, it must be enabled on both the VMkernel adapters and the vSwitches.
- Enable the software iSCSI initiator.
- Add an IP address from an iSCSI port on controller A to the software iSCSI imitator under dynamic discovery.
- Add a second IP address from an iSCSI port on controller B to the software iSCSI initiator under dynamic discovery.
- Ensure that each IP address used in the prior steps represents each of the subnets used in the SAN fabric.
- Rescan the iSCSI software adapter.

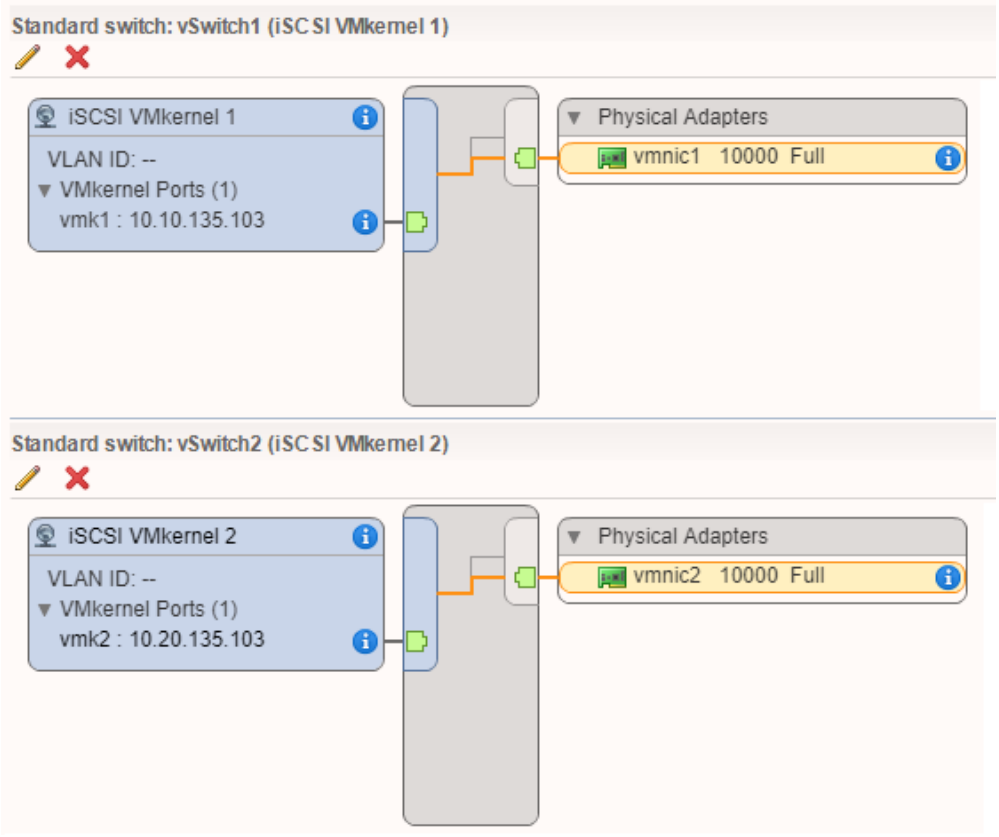

Figure 5 Combined image of vSwitch layout (vSphere Flex interface)

6.2 Recommended multipathing (MPIO) settings

Block storage (iSCSI, FC, or SAS to vSphere hosts from the ME4 Series array) has the native path selection policy (PSP) of most recently used (MRU) applied by default. If the vSphere hosts are connected to the ME4 Series array through SAN fabrics that follow the best practices as described in section 3, Connectivity considerations, multiple paths will be presented to each volume. Half of the paths go to the active controller that owns storage pool from which the volume is created, and the remaining paths go to the passive failover or alternative controller. Since the ME4 Series array is ALUA compliant, and recognized as such by vSphere ESXi, I/O will be correctly routed to the owning or active controller.

With MRU, only one of the two or four paths to the active controller is transporting I/O, with the remaining path or paths only transporting I/O if current path fails. Changing the PSP to round robin (RR) enables the I/O workload to be distributed across all the available paths to the active controller, resulting in better bandwidth optimization. It is recommended to use round robin for SAN-attached volumes.

In addition to changing the PSP to round robin, it is also recommended to change the default number of I/Os between switching paths. The default setting waits until 1,000 I/Os have been sent before switching to the next available path. This may not fully utilize the entire available bandwidth to the SAN when multiple paths are available. It is recommended to change the default number of I/Os between switching paths from 1,000 to 3, as described in the following subsections.

Note: In direct-attached configurations, such as SAS and direct-attached Fibre Chanel, there are typically only two connections, one to each controller, and therefore two paths. In such a configuration, with only one path to the active or owning controller, round robin has no benefit over MRU.

Applying these setting to all the datastores mounted to all the ESXi hosts in a vSphere environment can be achieved in different way, of which two examples are shown in the following subsections.

6.2.1 Modify SATP claim rule

Modifying the SATP claim rule is advantageous because it will apply to all current and future datastores that are added to the ESXi host, but it requires a reboot to be applied. Once the rule is created and a reboot occurs, all current, and future datastores will have the recommended setting applied to them.

To automatically set multipathing to round robin, and set the IOPS path change condition for all current and future volumes mapped to an ESXi host, create a claim rule with the following command:

```
esxcli storage nmp satp rule add --vendor "DellEMC" --model "ME4" --satp
"VMW_SATP_ALUA" --psp "VMW_PSP_RR" --psp-option "iops=3" --claim-
option="tpgs_on"
```

SATP claim rules cannot be edited; they can only be added or removed. To make changes to an SATP claim rule, it must be removed and then re-added. To remove the claim rule, issue the following command:

```
esxcli storage nmp satp rule remove --vendor "DellEMC" --model "ME4" --satp
"VMW_SATP_ALUA" --psp "VMW_PSP_RR" --psp-option "iops=3" --claim-
option="tpgs_on"
```

A reboot is required for claim rule changes to take effect.

Note: For OEM models, the --model must be changed to **VA4** instead of **ME4.**

6.3 ESXi iSCSI setting: delayed ACK

Delayed ACK is a TCP/IP method of allowing segment acknowledgments to transport upon each other or on other data that is passed over a connection with the goal of reducing I/O overhead. One side effect of delayed ACK is that if the pipeline is not filled, acknowledgment of the data is delayed. This can be seen as higher latency during lower I/O periods. Latency is measured from the time data is sent to when the acknowledgment is received. With disk I/O, any increase in latency can result in decreased performance. If higher latency during lower I/O periods is observed in the environment, disabling delayed ACK may resolve the issue. Otherwise, consult Dell Support.

VMware KB article 1002598 provides additional information, including instructions for disabling delayed ACK on ESXi.

6.4 Virtual SCSI controllers

When adding additional virtual hard drives (VMDKs) to a virtual machine, there are two virtual SCSI controller changes that should be considered.

Virtual machines can be configured with additional SCSI controllers. Each SCSI controller not only enables a greater number of VMDKs to be added to the virtual machine, they also add a separate disk queue. These separate disk queues can prevent contention for I/O between VMDKs. Adding virtual SCSI controllers involves the same process as adding any virtual hardware to a virtual machine, through the Edit settings menu.

There are several virtual SCSI controllers. By default, any additional virtual SCSI controllers will be of the default type for that guest operating system. For example, the default virtual SCSI controller with Windows Server 2012 is LSI Logic SAS. However, the VMware Paravirtual (PVSCSI) virtual SCSI controller has a lower host CPU cost per I/O, freeing up those CPU cycles for more valuable uses. VMware has a detailed white paper that takes a close look at the IOPS, latency, and cost of the PVSCSI and LSI Logic SAS controllers. Check the VMware Guest OS Compatibility Guide to ensure that the Paravirtual controller is compatible with the virtual machine's operating system.

6.5 Datastore size and virtual machines per datastore

While administrators continually try to maintain optimized data layout and performance, the size of the datastore becomes a question. Because every environment is different, there is no single answer to the size and number of LUNs. However, the typical recommendation is 10 to 30 VMs per datastore. Several factors in this decision include the speed of the disks, RAID type, and workload intensity of the virtual machines.

VMware currently supports a maximum datastore size of 64 TB. However, in most circumstances, a much smaller, more manageable size would be recommended to accommodate a reasonable number of virtual machines per datastore. Having one virtual machine per datastore would pose an abundance of administrative overhead, and putting all virtual machines on a single datastore would likely cause a performance bottleneck. VMware currently supports a maximum of 2,048 powered-on virtual machines per VMFS datastore. However, in most circumstances and environments, a target of 15 to 25 virtual machines per datastore is the conservative recommendation. By maintaining a smaller number of virtual machines per datastore, potential for I/O contention is greatly reduced, resulting in more consistent performance across the environment.

Determine the most beneficial compromise by monitoring the performance environment to find volumes that may be underperforming. In addition, monitor the queue depth with esxtop to see if there are outstanding I/Os

to a volume indicating that too many VMs may reside on that datastore. It is also important to consider the recovery point objective and recovery time objective of backups and replication for business continuity and recovery.

7 VMware integrations

The ME4 Series array has several integrations and touchpoints with the VMware vSphere ecosystem. Like the vSphere Web Client plug-in, some are more visible then others, such as the VAAI primitives in the firmware. However, a clear understanding of the functionality and benefits they provide is vital to efficient virtual-environment design.

These integrations and touchpoints include VMware vStorage APIs for Array Integration (VAAI) and VMware Storage I/O Control (SIOC).

7.1 VMware vStorage APIs for Array Integration

VMware vSphere recognizes that the underlying storage it is using may be capable of more than just storing data. Through its storage partners (including Dell Technologies), VMware developed a set of APIs which leverage the SCSI T10 specification to use the advanced capabilities that exist in intelligent storage products such as ME4 Series arrays.

This set of APIs is referred to as vStorage APIs for Array Integration (VAAI) and includes the following SCSI primitives:

- Full copy
- Block zeroing
- Hardware-assisted locking
- Thin provisioning space reclamation

7.1.1 Full copy

A common day-to-day IT task involves deploying servers to support new business applications. Virtualization changed this from a labor-intensive task of racking a server and installing the operating system to a simple task that required only a couple of mouse clicks to deploy a virtual machine from a preconfigured template. While this change has resulted in substantial time savings, there was still a significant amount of time spent watching the progress bar as the virtual machine deployed. Traditionally, the process of deploying a virtual machine involved all of its data being read from the array, across the network to the ESXi host, and then written back across the network to the array. This placed a non-production workload on both the network and the ESXi host, in addition to the production workload of the running environment. Now, with the full copy primitive, ESXi can offload this task to the array where it can be completed much more efficiently, with a significant workload reduction for the ESXi host and the network. The benefits of full copy do not end with deploying virtual machines from templates. They also extend to virtual-machine tasks such as Storage vMotion, and virtual machine cloning.

7.1.2 Block zeroing

Fault-tolerant virtual machines require VMDKs that are eager-zeroed thick. These differ from standard thick or thin VMDKs in that the blocks are zeroed out at the time that the VMDK is created. For large disks, this can take a significant amount of time as each zero is written from the server to the array, and an acknowledgment of each write is sent back from the array to the server. With the block zeroing primitive, the ESXi host offloads to the ME4 Series array the task of zeroing out the blocks, and permits the host to continue creating the fault-tolerant virtual machine while the storage completes the zeroing task in the background. By offloading the block zeroing to the ME4 Series array, fault-tolerant virtual machines can be created much faster.

7.1.3 Hardware-assisted locking

To protect Virtual Machine File System (VMFS) metadata, the hardware-assisted locking primitive provides a more granular method than SCSI reservations. Previously, whenever a virtual machine powered on, powered off, grew a thin provisioned virtual disk, or was moved with vMotion to another host, a SCSI reservation lock would be issued by the ESXi host to the underlying volume of the datastore. This prevented other hosts from also issuing a SCSI reservation to service a similar request. While SCSI reservations are short-lived, the impact can be noticed when powering on a large number of virtual machines simultaneously, as typically observed in a virtual desktop infrastructure (VDI) environment. The hardware-assisted locking primitive resolves this by working with the ME4 Series array to lock only the necessary blocks rather than the entire volume. This enables other hosts to perform similar operations against that same volume at the same time.

7.1.4 Thin provisioning space reclamation

The thin provisioning space reclamation primitive, also known as unmap, enables thin-provisioned datastores to be rethinned to only consume the actual space they are consuming on the array. This frees up space on the array that has been deleted by ESXi, allowing thin-provisioned volumes to remain thin and reducing overall storage costs. Traditionally, the size of a thin-provisioned volume, as shown at the storage layer, reflects the maximum space consumption that occurred at some point since it was created. This is because ESXi did not inform the array that particular blocks of data had been deleted and no longer needed to be stored by the array. The T10 SCSI primitive unmap enables this information to be communicate to the array, through the SCSI storage stack. This unmap primitive is referred to as thin provisioning space reclamation by VMware.

With the release of vSphere 6.7, VMware updated the unmap API to run automatically in the background without user intervention as part of VMFS-6. This is dependent upon arrays using 1 MB or smaller pages. The ME4 Series array uses 4 MB pages, and therefore is incompatible with automatic unmap.

The command for running unmap is a part of the esxcli command set of the ESXi operating system. This enables the command to be accessed from many scripting tools and to be called remotely with vSphere vCLI or PowerCLI. The syntax of the unmap command is as follows:

```
esxcli storage vmfs unmap --volume-label=volume_label
```

Note: VMware recommends limiting unmap operations to an off-peak operating timeframe.

7.2 VMware Storage I/O Control

Storage I/O Control (SIOC) ensures that the excessive storage I/O demands of a particular VMDK do not negatively impact the storage I/O needs of other VMDKs residing on the same datastore. Previously, this has been resolved though administrative tasks such as careful VM placement, reactive monitoring of VMDK I/O, and oversizing of the environment to handle occasional I/O spikes.

With SIOC, the reactive monitoring task is conducted by vSphere across all ESXi hosts and the reactive action is performed automatically and instantaneously by vSphere, enabling administrators to more efficiently utilize their storage environments.

The advantages of using Storage I/O Control include the following:

Performance protection: SIOC ensures that all VMDKs receive a fair share or an assigned share of I/O needs, regardless of the I/O they demand during period of congestion.

Better utilization of storage assets: The storage environment no longer needs to be oversized to cover occasional I/O peaks. Rather, these peaks are leveled out by SIOC.

SIOC works by monitoring the I/O latency of a datastore. When that latency exceeds the threshold that has been set (30 ms by default), SIOC will engage and enforce the assigned disk shares. By default, all VMDKs receive the same number of shares, and therefore during times of contention, excessive consumers will be restricted. SIOC achieves this by restricting the number of queue slots available to the VMDKs that are consuming more than their assigned share and provides the previously deprived VMDKs with improved storage performance. Alternatively, a VMDK may be assigned a greater or lesser number of shares than other VMDKs, resulting in SIOC favoring or disfavoring that VMDK to a greater degree, but only during I/O contention.

While SIOC does not eliminate the need for SAN monitoring, it means that the SAN does not need to be actively monitored, freeing up the storage administrator to deal with more important tasks. If SIOC is engaging for significant periods of time, the administrator may need to add additional I/O capacity or relocate I/O-intensive VMDKs.

As a best practice, when using datastores backed by an ME4 Series array configured with a single tier of disks, SIOC is recommended to balance out the I/O needs of VMDKs that share the same datastore.

7.2.1 Storage I/O Control and automated tiered storage

Storage I/O Control is a great equalizer ensuring that each VMDK gets its fair share of I/O when there is contention, but other options are available. Traditionally, SAN storage provides a datastore with only one tier of performance, and often the choice is either fast RAID 10 SSD or slow RAID 6 NL-SAS. However, today's modern SANs designed with ME4 Series storage can spread a volume across multiple tiers of storage with varying performance characteristics. Automated tiered storage (ATS) will move data between the different tiers of storage depending upon the performance needs of the data. This enables the ME4 Series array to adjust to the changing demands of the virtual machine's application.

ATS resolves storage performance issues by relocating highly active data to higher performing tiers of storage. A virtual machine's VMDK with highly active data can trigger SIOC's throttling mechanism, and a slowing of the highly active data, thus preventing ATS from repositioning the data to a higher performing tier of storage.

As a best practice, when using datastores backed by an ME4 Series array configured with ATS, SIOC should be left at the default setting of **Disabled**.

A Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Storage and data protection technical white papers and videos provide expertise that helps to ensure customer success with Dell EMC storage and data protection products.

A.1 Related resources

See the following referenced or recommended resources related to this document:

- ME4 Series product page
- How to download the Dell-customized ESXi Embedded ISO image
- Achieving a Million I/O Operations per Second from a Single VMware vSphere 5.0 Host
- Best Practices for Running VMware vSphere on iSCSI
- Configuring Software iSCSI Adapters
- Setting Up iSCSI Network

The following ME4 Series publications and additional resources are available at Dell.com/support.

- Administrator's Guide
- Deployment Guide
- CLI Guide
- Owner's Manual
- Support Matrix