WHITE PAPER

Titan T2000 VMware vSphere

CRS TITAN

POWERED BY
DELL Technologies

Dell PowerVault ME5 Series: VMware vSphere Best Practices

February 2022

H19030

White Paper

Abstract

This document provides best practices for deploying VMware vSphere with Dell PowerVault ME5 storage. It includes configuration recommendations for vSphere hosts to achieve an optimal combination of performance and resiliency.

Copyright

Contents

Executive summary

Overview

This document provides best practices for VMware vSphere when using a Dell PowerVault ME5 storage array. It does not include sizing, performance, or design guidance, but it provides information about the features and benefits of using PowerVault ME5 arrays for VMware vSphere environments.

VMware vSphere is an extremely robust, scalable, enterprise-class hypervisor. When correctly configured using the best practices in this paper, the vSphere ESXi hypervisor provides an optimized experience with PowerVault ME5 storage. These recommendations include guidelines for SAN fabric design, HBA settings, and multipath configuration. While there are various methods to accomplish the described tasks, this paper primarily provides a starting point for end users and system administrators. It is not intended to be a comprehensive configuration guide.

Audience

This document is intended for PowerVault ME5 administrators, system administrators, and anyone responsible for configuring PowerVault ME5 systems. It is assumed the readers have prior experience with or training in SAN storage systems and a VMware vSphere environment.

Revisions

Date	Description
February 2022	Initial release

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Darin Schmitz

Contributor: Jason Boche

Note: For links to other documentation for this topic, see Dell Technologies Info Hub.

Introduction

PowerVault ME5 Series

PowerVault ME5 Series storage is the new gold standard for entry storage that is purpose-built and optimized for SAN and DAS virtualized workloads. Available in 2U or dense 5U base systems, the affordable PowerVault ME5 simplifies the challenges of server capacity expansion and small-scale SAN consolidation with up to 336 drives or 8 PB[1] capacity. It also comes with all-inclusive software, incredible performance, and built-in simplicity with an HTML5 management UI, PowerVault Manager. Connecting PowerVault ME5 storage to a PowerEdge server or to a SAN ensures that business applications get high-speed and reliable access to their data—without compromise.

Product features include the following:

Simplicity: A web-based management UI (HTML5), installs and configures in 25 minutes and easily deploys in 2U or 5U systems.

Performance: Compared to its PowerVault ME4 Series predecessor, PowerVault ME5 packs a lot of power, bandwidth, and scale with faster Intel® Xeon® processors. The PowerVault ME5 processing power delivers incredible performance gains over PowerVault ME4, plus increased capacity, bandwidth, and drive count.

Connectivity: PowerVault ME5 goes to the next level with robust and flexible connectivity starting with a 12 Gb SAS back-end interface. Front-end interface options include the following:

- Four 16/32 Gb FC ports per controller
- Four 10 Gb iSCSI ports per controller (BaseT)
- Four 10/25 Gb iSCSI (optical)
- Four 12 Gb 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 base systems (PowerVault ME5012 and PowerVault ME5024) and the 5U PowerVault ME5084 base system support optional expansion enclosures of 12, 24, and 84 drives. This expansion capability allows you to use up to 336 drives total with the 5U system. PowerVault ME5 also allows drive mixing.

All-inclusive software: PowerVault ME5 software provides volume copy, snapshots, IP/FC replication, VMware vCenter Server, and VMware Site Recovery Manager integration. It also provides an SSD read cache, thin provisioning, three-level tiering, ADAPT (distributed RAID), and controller-based encryption (SEDs) with internal key management**.**

Management: PowerVault ME 5 includes an integrated HTML5 web-based management interface, PowerVault Manager. Both OpenManage Enterprise and Nagios management utilities support PowerVault ME5.

For more information, see the Dell PowerVault ME5 product page at www.dell.com/powervault.

[1] PowerVault ME5 firmware is designed to support up to 8 PB when higher-capacity drives are available for use.

PowerVault ME5 features

Introduction

Although PowerVault ME5 is targeted at the entry level of the SAN market, it contains many advanced and enterprise-class features detailed in the following sections. We recommend that the storage administrator and VMware administrator understand how these storage features can benefit the vSphere environment before deployment.

Note: PowerVault ME5 arrays use the term **virtual volume**, which is not associated with the VMware vSphere Virtual Volumes (vVols) feature.

Virtual and linear storage

PowerVault ME5 arrays use two storage technologies that share a common user interface: the virtual method and the linear method.

Linear method: This method maps logical host requests directly to physical storage. Sometimes, the mapping is one-to-one, while usually, the mapping spans 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.

Virtual method: This method maps logical storage requests to physical storage (disks) through a layer of virtualization. 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. You can create multiple virtual volumes in a pool, sharing its resources. This practice enables 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 consist 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, we recommend using virtual storage when deploying VMware vSphere environments. Linear storage pools are most suited to sequential workloads such as video archiving.

Note: You cannot use data management features in Linear Storage Mode.

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 array. Both the automated tiered storage and read SSD cache features of the PowerVault ME5 arrays can use a small number of SSDs. This use can provide a significant performance boost to a traditional low-cost, all-HDD SAN solution.

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 swaps 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 swaps into higher tiers based on frequency of access or when space is available.

Read flash cache

With tiering, a single copy of specific data blocks resides in either spinning disks or SSDs. However, 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. You can add read flash cache 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 extends 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 the system cost.

Asymmetric Logical Unit Access

PowerVault ME5 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 controller operating or 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 extra configuration is required. Each datastore has two, four, or eight active paths depending upon the controller configuration (SAS, combined FC/iSCSI controller, or dedicated FC/iSCSI). Half of the paths identified as active are optimized, and the other half are identified as active non-optimized.

RAID data protection levels

PowerVault ME5 arrays support RAID data protection levels NRAID, 0, 1, 5, 6, 10 and ADAPT. ADAPT, included on every PowerVault ME5 array, 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.

Note: RAID 0 is only available for read cache and cannot be used in linear mode.

We recommend selecting the RAID level that is most appropriate for the type of workloads in the environment. Review the information in the *PowerVault ME5 Administrator's Guide* on Dell.com/support. This guide details the benefits of each RAID level, the minimum and maximum disks requirements, and the recommendation of RAID levels for popular workloads.

Connectivity considerations

Introduction

PowerVault ME5 storage supports and is certified with VMware vSphere for server connectivity with iSCSI (10/25Gb), Fibre Channel (16/32 Gb, direct-attached, and SAN-attached), and direct-attached SAS. While you can configure the PowerVault ME5012 or PowerVault ME5024 array 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. You can replace a failed controller without having to shut down the storage system.

Direct-attached storage

PowerVault ME5 arrays support direct-attached Fibre Channel (16/32 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 PowerVault ME5 storage can support up to eight direct-attached servers, this configuration is achieved by providing only a non-redundant, single connection to each server. As a best practice, ensure each host has a dual-path configuration with a single path to each controller, enabling storage access to continue in the event of controller failure. This practice limits the number of direct-attached servers to four but enables controller redundancy and increased performance.

The following figure shows a configuration with four servers, each with two Fibre Channel or SAS connections to the PowerVault ME5 array.

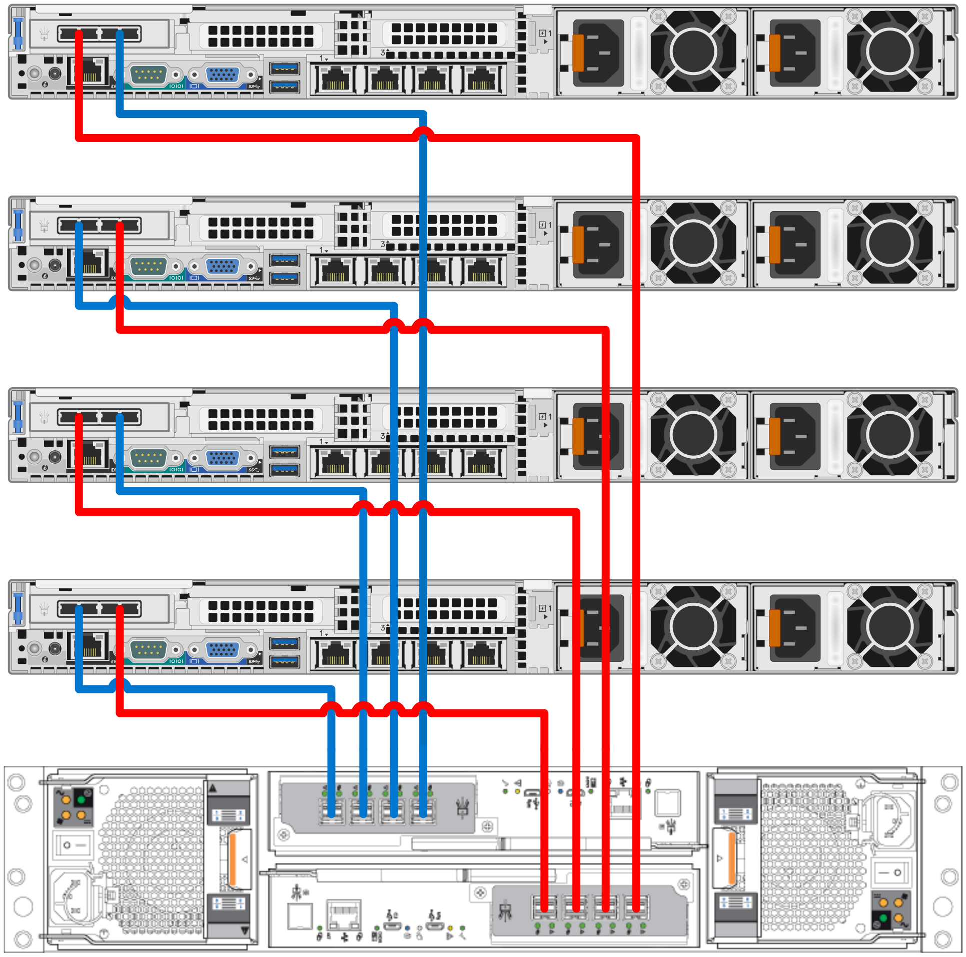


Figure 1. Connecting four hosts directly to a PowerVault ME5024 array with dual paths

SAN-attached storage

PowerVault ME5 arrays support SAN-attached Fibre Channel (16/32 Gb) and iSCSI (10/25Gb) 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, we recommend using two switches. This practice enables you to create a redundant transport fabric between the server and the PowerVault ME5 storage. This configuration also allows you to take an individual switch out of service for maintenance, or due to failure, without impacting storage-access availability.

When cabling the PowerVault ME5 controllers in a switched environment, pay close attention to the layout of the cables in both Fibre Channel and Ethernet fabrics. In the following figure, controller A (the left-most PowerVault ME5084 controller) has ports 0 and 2 connected to the top switch, and ports 1 and 3 are connected to the bottom switch. This configuration repeats 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 PowerVault ME5 array during switch maintenance.

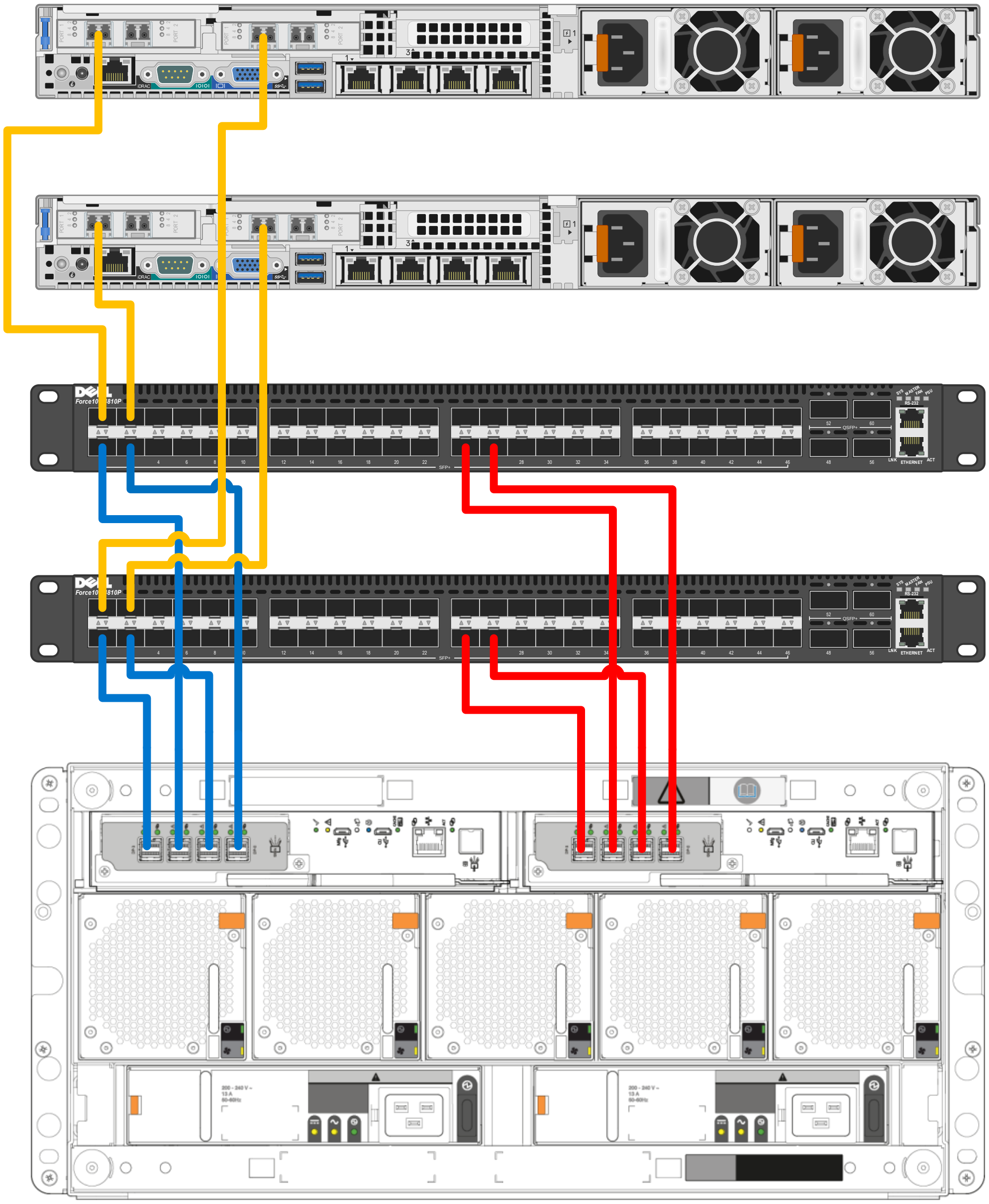


Figure 2. Connecting two hosts to a PowerVault ME5084 array using two switches

iSCSI fabric settings

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

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.

We recommend the following settings when enabling flow control:

- Enable a minimum of receive (RX) flow control for all switch interfaces used by servers or storage systems for iSCSI traffic.
- Enable symmetric flow control for all server interfaces used for iSCSI traffic. PowerVault ME5 automatically enables this feature.

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 size is increased to 9,000 bytes.

When enabling Jumbo frames, all devices in the path must be enabled for Jumbo frames for this frame size to be successfully negotiated. These devices include server NICs or iSCSI HBAs, switches, and the PowerVault ME5 storage. In a vSphere environment, the devices also include the virtual switches and VMkernel adapters configured for iSCSI traffic.

Jumbo frames and flow control

Some switches have limited buffer sizes and can support either Jumbo frames or flow control but cannot support both simultaneously. If you must choose between the two features, we recommend choosing flow control.

Fibre Channel zoning

Fibre Channel zones can segment the fabric to restrict access. A zone contains paths between initiators (server HBAs) and targets (storage array front-end ports). You can use either physical ports (port zoning) on the Fibre Channel switches or the WWNs (name zoning) of the end devices in zoning. We recommend using 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 no different than zoning any other hosts to the array.

Zoning rules and recommendations:

- Connect the PowerVault ME5 array and ESXi hosts to two different Fibre Channel switches (fabrics) for high availability and redundancy.
- Use WWNs for name zoning.https://infohub.delltechnologies.com/l/dell-powervault-me5-series-vmware-vsphere-best-practices/introduction-2421

- When defining the zones, it is a best practice to use single-initiator (host port), multiple-target (PowerVault ME5 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 PowerVault ME5 array controllers on the same fabric. See the following table for an example.

Table 1. Fibre Channel zoning examples

Fabrics (dual-switch configuration)	FC HBA port (dual-port HBA configuration)	PowerVault ME5 FC ports (FC port configuration)
Fabric one zone	Port 0	A0, B0, A2, B2
Fabric two zone	Port 1	A1, B1, A3, B3

Note: We recommend using name zoning and creating single-initiator, multiple-target zones.

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. Ports 0 and 1 share a converged network controller chip, and ports 2 and 3 share a separate converged network controller chip.

Host bus adapters

Introduction

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 PowerVault ME5 array.

Fibre Channel and SAS HBAs

To obtain drivers for the Fibre Channel or SAS HBAs shipped in Dell PowerEdge servers, download the Dell customized ESXi embedded ISO image from Dell Support. The drivers are fully compatible with the PowerVault ME5 array and do not require further configuration.

iSCSI HBAs

The PowerVault ME5 array is only certified with the vSphere ESXi software iSCSI initiator. No dependent, independent, or iSCSI offload cards are supported.

PowerVault ME5 array settings

Introduction

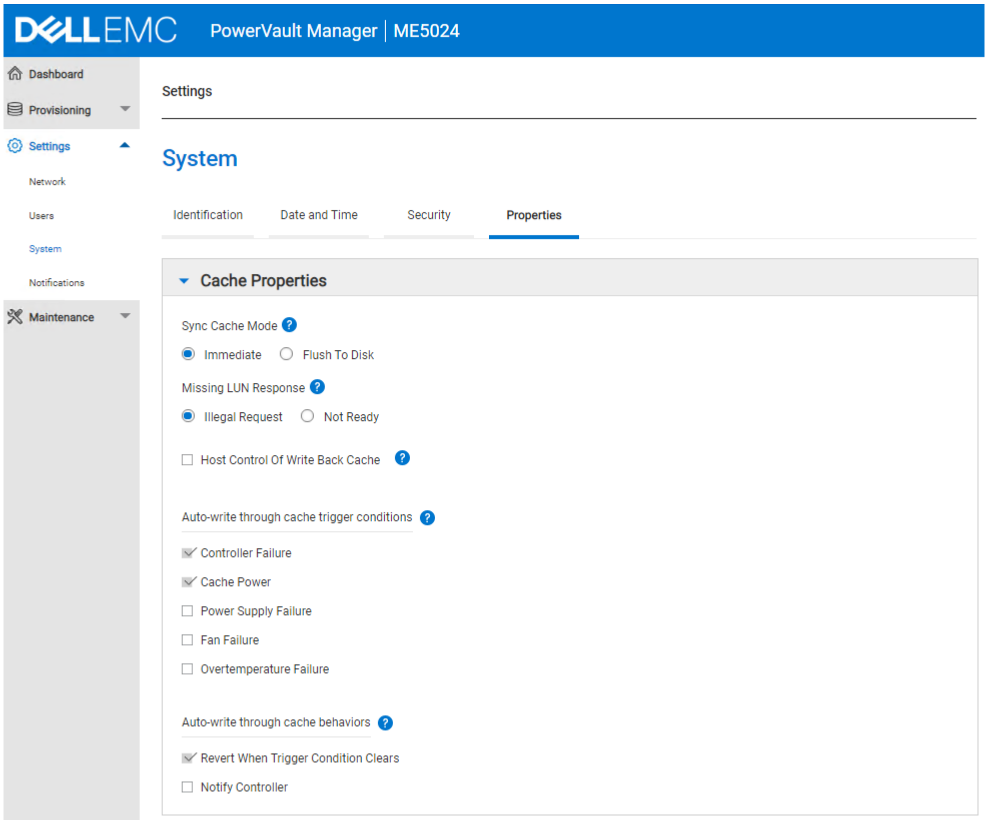
This section includes PowerVault ME5 array settings that ensure a smooth and consistent data-center environment.

Missing LUN Response

The setting for **Missing LUN Response** is in PowerVault Manager under **Settings > System > Properties > Cache Properties**.

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 unmounting the datastore first, and if the missing LUN response is set to Not Ready, ESXi may continue to query for this LUN indefinitely.



Cache Properties

Sync Cache Mode

● Immediate ○ Flush To Disk

Missing LUN Response

● Illegal Request ○ Not Ready

☐ Host Control Of Write Back Cache

Auto-write through cache trigger conditions

☑ Controller Failure

☑ Cache Power

☐ Power Supply Failure

☐ Fan Failure

☐ Overtemperature Failure

Auto-write through cache behaviors

☑ Revert When Trigger Condition Clears

☐ Notify Controller

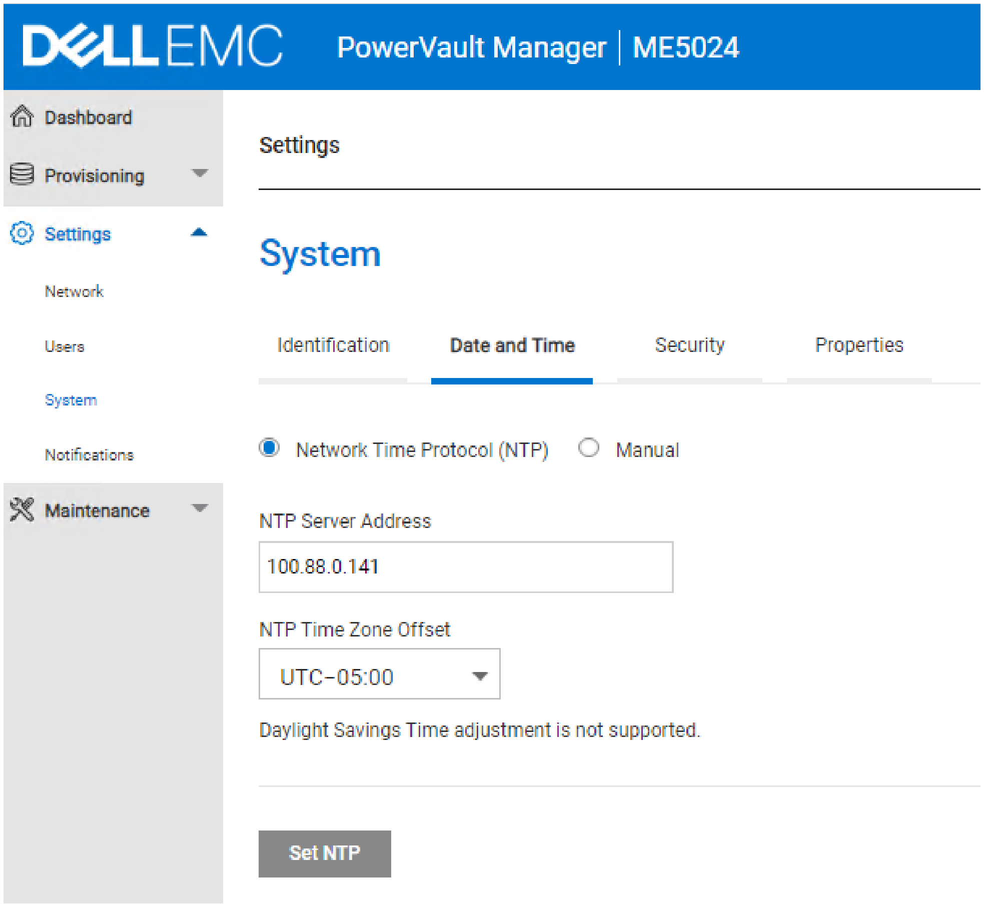
Figure 3. Missing LUN Response setting

Host groups

For ease of management with PowerVault ME5 arrays, group initiators that represent a server into an object called a host. You can also organize multiple host objects 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.

Log file timestamps

Debugging and troubleshooting any data-center issue involves reviewing multiple log file 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. Locate these settings in PowerVault Manager under **Settings > System > Date and Time**.



● Network Time Protocol (NTP) ○ Manual

NTP Server Address
100.88.0.141

NTP Time Zone Offset
UTC-05:00

Daylight Savings Time adjustment is not supported.

Figure 4. Network Time Protocol system settings

VMware vSphere settings

Introduction

We recommend the following configuration settings for the VMware ESXi hosts.

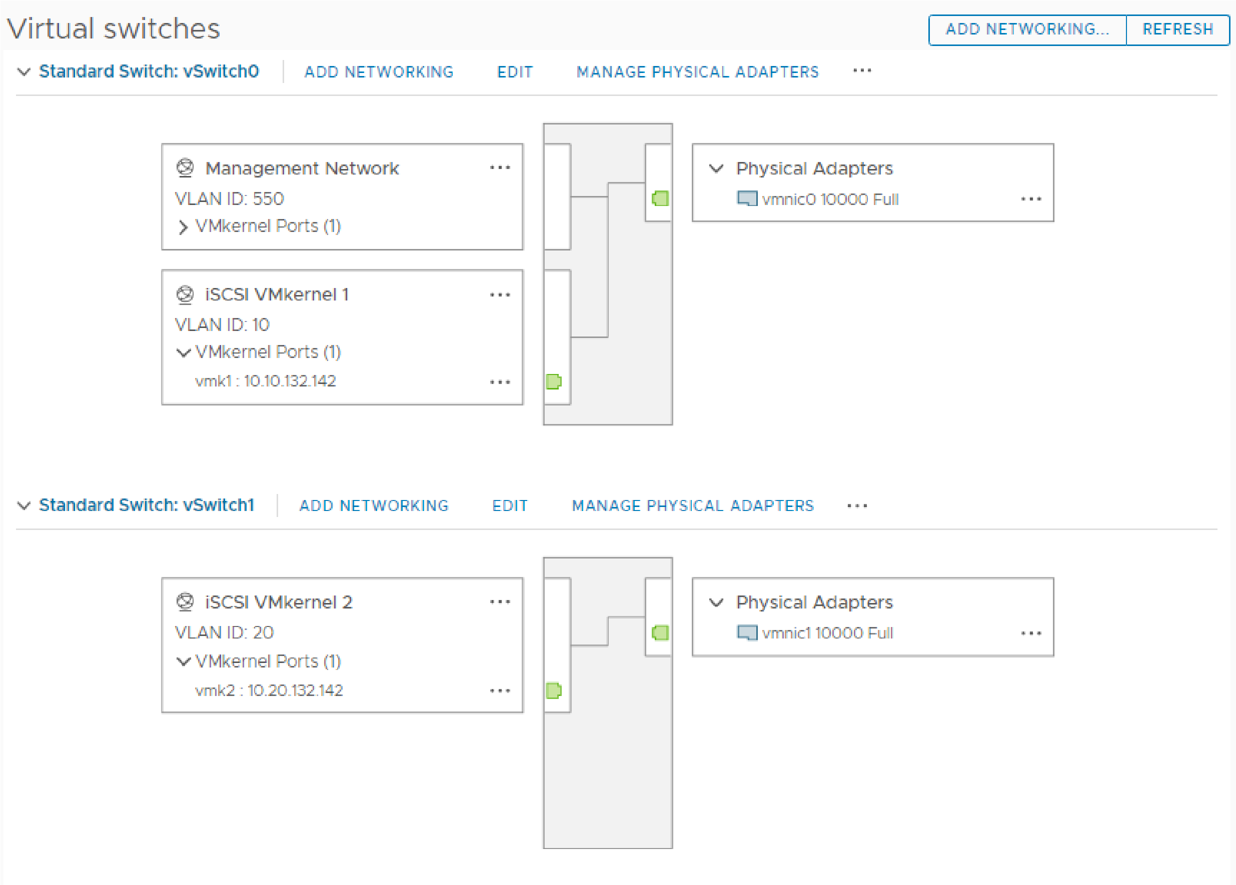
Recommended iSCSI vSwitch configuration

To configure the VMware iSCSI software initiator for multipathing, see the VMware articles, *Configuring Software iSCSI Adapters in the VMware Host Client* and *Setting Up Network for iSCSI and iSER* in VMware documentation.

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.
- Ensure each VMkernel adapter is on its own vSwitch with a single vmnic (physical NIC adapter).
- Remember if Jumbo frames is used, you must enable it 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 previous steps represents each of the subnets used in the SAN fabric.
- Rescan the iSCSI software adapter.



Standard Switch: vSwitch0

Management Network
VLAN ID: 550
VMkernel Ports (1)

iSCSI VMkernel 1
VLAN ID: 10
VMkernel Ports (1)
vmk1 : 10.10.132.142

Physical Adapters
vmnic0 10000 Full

Standard Switch: vSwitch1

iSCSI VMkernel 2
VLAN ID: 20
VMkernel Ports (1)
vmk2 : 10.20.132.142

Physical Adapters
vmnic1 10000 Full

Figure 5. vSwitch layout

Recommended multipathing (MPIO) settings

Block storage (iSCSI, FC, or SAS to vSphere hosts from the PowerVault ME5 array) has the native path selection policy (PSP) of most recently used (MRU) applied by default. If the vSphere hosts connect to the PowerVault ME5 array through SAN fabrics that follow the best practices described in the Connectivity considerations section, multiple paths are 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 PowerVault ME5 array is ALUA compliant and is recognized as such by vSphere ESXi, I/O is correctly routed to the owning or active controller.

With MRU, only one of the two or four paths to the active controller transports I/O, and the remaining path or paths only transport 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. We recommend using round robin for SAN-attached volumes.

Besides changing the PSP to round robin, we also recommend changing the default number of I/Os between switching paths. The default setting waits until 1,000 I/Os are sent before switching to the next available path. This setting may not fully utilize the entire available bandwidth to the SAN when multiple paths are available. We recommend changing the default number of I/Os between switching paths from 1,000 to 1, 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 two paths. In such a configuration, with only one path to the active or owning controller, round robin has no benefit over MRU.

There are other ways that you can apply these setting to all the datastores mounted to all the ESXi hosts in a vSphere environment. Two examples are shown in the following subsections.

Modify SATP claim rule

Modifying the SATP claim rule is advantageous because it applies to all current and future datastores that are added to the ESXi host, but it requires a reboot. After you create the rule and perform a reboot, all current, and future datastores 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
"PowerVault ME5" --satp "VMW_SATP_ALUA" --psp "VMW_PSP_RR" --psp-
option "iops=1"   --claim-option="tpgs_on"
```

SATP claim rules cannot be edited; they can only be added or removed. To change an SATP claim rule, you must remove and readd it. To remove the claim rule, issue the following command:

```
esxcli storage nmp satp rule remove --vendor "DellEMC" --model
"PowerVault ME5" --satp "VMW_SATP_ALUA" --psp "VMW_PSP_RR" --psp-
option "iops=1" --claim-option="tpgs_on"
```

Perform a reboot for the claim rule changes to take effect.

Modify multipathing setting

You can change the PSP from MRU to round robin using the VMware vSphere Web client for each existing datastore. However, this process can become unwieldy if there are many existing new datastores that require changes. It also does not permit changing the number of I/Os between switching paths. It is often more convenient for an administrator to write a short script to change all the devices on a particular host. For more information about the commands required, see VMware KB article 1017760.

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. The goal of this method is to reduce I/O overhead. One side effect of delayed ACK is that if the pipeline is not filled, acknowledgment of the data is delayed. This effect manifests 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 more information, including instructions for disabling delayed ACK on ESXi.

Virtual SCSI controllers

When adding additional virtual hard drives (VMDKs) to a virtual machine, consider the following two virtual SCSI controller changes.

Virtual machines can be configured with extra SCSI controllers. Each SCSI controller not only enables a greater number of VMDKs to be added to the virtual machine, it also adds a separate disk queue. These separate disk queues can prevent contention for I/O between VMDKs. Adding virtual SCSI controllers requires 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 extra virtual SCSI controllers are of the default type for that guest operating system. For example, the default virtual SCSI controller with Windows Server 2022 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. Check the VMware Guest OS Compatibility Guide to ensure that the Paravirtual controller is compatible with the virtual machine operating system.

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, we recommend starting with 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 supports a maximum datastore size of 64 TB. However, in most circumstances, we recommend using a much smaller and more manageable size to accommodate a reasonable number of virtual machines per datastore. Having one virtual machine per datastore can result in high administrative overhead and puts all virtual machines on a single datastore, likely causing 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. After you establish a performance baseline for a datastore, if there is little to no I/O contention, you can add virtual machines.

Determine the most beneficial compromise by monitoring the performance environment to find volumes that may be underperforming. Also, 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.

VMware integrations

Introduction

The PowerVault ME5 array has several integrations and touchpoints with the VMware vSphere ecosystem. Like the vSphere Web Client plug-in, some are more visible than 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).

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 use the SCSI T10 specification to take advantage of the advanced capabilities of intelligent storage products such as PowerVault ME5 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

Full copy

A common day-to-day IT task involves deploying servers to support new business applications. Virtualization changed this task from a labor-intensive process of racking a server and installing the operating system to a simple task. With only a couple of mouse clicks, you could 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, deploying a virtual machine required reading its data from the array across the network to the ESXi host, and writing the data back across the network to the array. This placed a nonproduction workload on both the network and the ESXi host, concurrently with 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. This primitive also results in 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. Benefits also extend to virtual-machine tasks such as Storage vMotion and virtual-machine cloning.

Block zeroing

Fault-tolerant virtual machines require VMDKs that are eager-zeroed thick. These VMDKs differ from standard thick or thin VMDKs in that the blocks are zeroed out when the VMDK is created. For large disks, this process can take a significant amount of time as each zero is written from the server to the array. Then, the array sends an acknowledgment of each write to the server, taking more time. With the block zeroing primitive, the ESXi host offloads to the PowerVault ME5 array the task of zeroing out the blocks. The primitive also 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 PowerVault ME5 array, you can create fault-tolerant virtual machines much faster.

Hardware-assisted locking

To protect Virtual Machine File System (VMFS) metadata, the hardware-assisted locking primitive provides a more granular method than traditional SCSI reservations. Previously, virtual-machine tasks would cause the ESXi host to issue a SCSI reservation lock to the underlying volume of the datastore. These VM tasks could include powering on or off a virtual machine, growing a thin-provisioned virtual disk, or moving a VM with vMotion to another host. The resulting SCSI reservation lock 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 many virtual machines simultaneously, as typically observed in a virtual desktop infrastructure (VDI) environment. The hardware-assisted locking primitive resolves this issue by working with the PowerVault ME5 array to lock only the necessary blocks rather than the entire volume. This feature enables other hosts to perform similar operations against that same volume simultaneously.

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 action frees up space on the array that was 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 behavior results because ESXi did not inform the array that distinct 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 communicated 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 ability is dependent upon arrays using 1 MB or smaller pages. The PowerVault ME5 array uses 4 MB pages, and is incompatible with automatic UNMAP.

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 issue was 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, vSphere conducts the reactive monitoring task across all ESXi hosts and performs the reactive action automatically and instantaneously, enabling administrators to use their storage environments more efficiently.

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 must be oversized to cover occasional I/O peaks. Rather, SIOC levels out these peaks.

SIOC works by monitoring the I/O latency of a datastore. When that latency exceeds the threshold that has been set, SIOC engages and enforces the assigned disk shares. By default, all VMDKs receive the same number of shares, and during times of contention, excessive consumers are restricted. SIOC achieves this result by restricting the number of queue slots available to the VMDKs that are consuming more than their assigned share. This action also improves the storage performance of the previously deprived VMDKs. Alternatively, a VMDK may be assigned more or fewer 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. This result enables the storage administrator to deal with more important tasks. If SIOC is engaging for significant periods of time, the administrator may have to add additional I/O capacity or relocate I/O-intensive VMDKs.

When using datastores backed by a PowerVault ME5 array configured with a single tier of disks, we recommend using SIOC to balance the I/O needs of VMDKs that share the same datastore.

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 PowerVault ME5 storage can spread a volume across multiple tiers of storage with varying performance characteristics. Automated tiered storage (ATS) moves data between the different tiers of storage depending upon the performance requirements of the data. This feature enables the PowerVault ME5 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 can slow highly active data. This action prevents ATS from repositioning the data to a higher performing tier of storage.

As a best practice, when using datastores backed by an PowerVault ME5 array configured with ATS, leave SIOC at the default setting of **Disabled**.

References

Dell Technologies documentation

The following Dell Technologies documentation provides other information related to this document. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- Dell Technologies Info Hub
- PowerVault ME5 product page
- Dell PowerVault ME5 Administrator’s Guide

VMware documentation

For more information related to VMware vSphere, see VMware vSphere Documentation.