White Paper

Dell EMC PowerVault ME4 Series Storage System Best Practices

Abstract

This white paper highlights best practices for optimizing and deploying Dell EMC™ Powervault ME4 series (ME4012/4024/4084) and should be used in conjunction with other Powervault ME4 manuals (Deployment guide, Admin Guide, Support Matrix etc.)

April 2020

Revisions

Date	Description
February 2020	Initial XenServer 7.6 release
April 2020	Minor Revisions

Acknowledgements

Authors: Selim Selveroglu

Contents

Executive Summary

This white paper highlights best practices for optimizing and deploying Dell EMC Powervault ME4 series (ME4012/4024/4084) and should be used in conjunction with other Powervault ME4 manuals (Deployment guide, Admin Guide, Support Matrix etc.) All manuals available from Dell Support web page.

Note

Features and recommendations in this document reflect current software functionality as of the latest firmware available at the time of publication. Features, functionality and GUI might vary with different storage system firmware levels.

Intended Audience

This best practice document is intended for Powervault ME4 series storage administrators and presales & deployment engineers with previous SAN infrastructure and SAN storage knowledge.

Prerequisites

Prerequisites for using this product include knowledge of:

- Storage system configuration
- SAN management
- Connectivity methods such as direct attached storage (DAS), Fibre Channel, and serial attached SCSI (SAS)
- Networking
- iSCSI and Ethernet protocols

Related documentation

In addition to this guide, other documents or materials for this product include:

- Dell EMC PowerVault ME4 Series Storage System Administrator's Guide
- Dell EMC PowerVault ME4 Series Owner's Manual
- Dell EMC PowerVault ME4 Series Storage System CLI Guide
- Dell EMC PowerVault ME4 Series Storage System Deployment Guide
- Dell EMC PowerVault ME4 Series Support Matrix

Introduction

Powervault ME4 models referenced in this paper are Powervault ME4012 / ME4024 and ME4084.

The affordable, simple, and fast Dell EMC PowerVault ME4 Series SAN/DAS Storage Series is optimized to run a variety of mixed workload applications – physical and virtual – for small businesses. Whether you need to consolidate your block storage, support the demands of data intensive applications, take advantage of intelligent data management, or optimize your virtual environments, the ME4 Series has been designed to meet your growing business needs. The flexibility of the ME4 Series lets you decide the protocol, supports a wide range of mixed drive types (including SED), scales to 4PB raw, is highly aligned with Dell PowerEdge Servers, and is delivered to you with all-inclusive software – everything you'll need to store, manage, and protect your data.

Based on the family of Intel processors and ASIC Chipsets, Dell EMC PowerVault ME4 Series storage implements a block architecture with VMware virtualization integration and concurrent support for native iSCSI, Fibre Channel, and SAS protocols. Each system leverages dual storage processors (single storage processor systems are available) and a full 12Gb SAS back-end. Additional storage capacity is added via Disk Array Enclosures (DAEs) while Distributed RAID (ADAPT) delivers faster drive re-build times. And all ME4 Series arrays are managed by an integrated HTML5 web-based GUI.

The two non-dense ME4 base arrays start at 2U and the dense ME4 array starts at 5U. Both models include dual controllers with, 8GB per controller and 4x10Gb iSCSI, 4x12Gb SAS, and 4x16Gb FC network connections (auto-negotiation supported on iSCSI and FC).

Powervault ME4012

- Up to 3.1PB capacity
- 12 x 3.5" drive bays
- Up to 264 drives
- Multi-protocols – SAS, iSCSI, Fibre Channel
- Single/Dual Controller
- 12Gb SAS Backend
- All premium software features included

PowerVault ME4024

- Up to 3.0PB capacity
- 24 x 2.5” drive bays
- Up to 276 drives
- Multi-protocols – SAS, iSCSI, Fibre Channel
- Single/Dual Controller
- 12Gb SAS Backend
- All premium software features included

PowerVault ME4084

- Up to 4.0PB capacity
- 84 x 3.5" drive bays
- Up to 336 drives
-
- Multi-protocols – SAS, iSCSI, Fibre Channel
- Dual Controller
- 12Gb SAS Backend
- All premium software features included

All Inclusive Softwares

PowerVault ME4 Series provides simplified, all-inclusive software licensing, including:

- **ADAPT (Distributed RAID):** (like Dynamic Disk Pooling) reduces drive rebuild times
- Thin Provisioning: Allocate and consume physical storage capacity as needed in disk pools. Thin is virtual mode only.
- SSD Read Cache: Increase execution speed of applications by caching previously read data
- IP & FC Remote Replication: Safely replicate data to any global location that includes mirroring thin provisioned pools
- Snapshots: Easily recover files after accidental deletion or alteration with Redirect on Write snaps
- 3 Level Tiering: Get great performance with less hardware expense
- Volume Copy/Clones: Enable seamless volume relocation and disk-based backup and recovery with a full, replicated copy of source data
- Encryption (SEDs): Render data useless to unauthorized users with drive-level encryption, even if the drive has been removed from the enclosure (internal key management included)
- Virtualization Integrations: VMware vSphere, vCenter SRM, Microsoft Hyper-V

Resources

PowerVault ME4 Series Data Sheet
PowerVault ME4 Series Spec Sheet

System Concepts

Virtual and Linear Storage

This product uses two different storage technologies that share a common user interface. One uses the virtual method while the other one uses the linear method.

Virtual storage (system default) is the most common selection and is recommended for most environments. Virtual storage allocates space in pages and allows data to be moved to improve system performance of the storage system. Virtual storage can support thin provisioning (user selectable), tiering, replication, and many other features not available to linear configurations. However, you cannot exceed 2 PB usable capacity in Virtual Storage.

Virtual storage is a method of mapping logical storage requests to physical storage (disks). It inserts a layer of virtualization such that logical host I/O requests are mapped onto pages of storage. Each page is then mapped onto physical storage. Within each page the mapping is contiguous, but there is no direct relationship between adjacent logical pages and their physical storage.

Linear storage is used in applications where performance and data workloads dictate that data be allocated on disks in a contiguous fashion with more predictable performance. Users in streaming media and video editing, High Performance Computing environments may prefer the performance and raw capacity available to a linear storage configuration. Features such as thin provisioning, snapshot, read cache, tiering, and replication are not available in a linear storage environment. Linear storage is similar to thick provisioning.

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. This linear method of mapping is highly efficient. The negative side of linear mapping is lack of flexibility. This makes it difficult to alter the physical layout after it is established.

Notes:

- Virtual Mode is required for Thin Provisioning, Tiering, SSD Read Cache, Snapshots and Replication
- Linear mode is required to support 4PB of usable capacity (see spec sheet for details)
- Virtual Mode cannot exceed 2PB usable capacity (see admin guide & CLI Guide for details
- ADAPT is supported in either Virtual or Linear Mode
- Once you select Virtual or Linear Mode you can’t change it online.

Pools

A pool is an aggregation of one or more disk groups that serves as a container for volumes. Virtual and linear storage systems both use pools. Dual controller systems consist of two pools. Each storage controller has ownership of a one pool.

In both virtual and linear storage, if the owning controller fails, the partner controller assumes temporary ownership of the pool and resources owned by the failed controller. If a fault-tolerant cabling configuration, with appropriate mapping, is used to connect the controllers to hosts, LUNs for both controllers are accessible through the partner controller so I/O to volumes can continue without interruption.

Disk Groups

A disk group is an aggregation of disks of the same type, using a specific RAID level that is incorporated as a component of a pool, for storing volume data. Disk groups are used in both virtual and linear storage environments. You can add virtual, linear, or read-cache disk groups to a pool.

Volumes

A volume is a logical subdivision of a virtual or linear pool and can be mapped to host-based applications. A mapped volume provides addressable storage to a host (for example, a file system partition you create with your operating system or third-party tools). For more information about mapping,

Volume groups

You can group a maximum of 1024 volumes (standard volumes, snapshots, or both) into a volume group. Doing so enables you to perform mapping operations for all volumes in a group at once, instead of for each volume individually. A volume can be a member of only one group. All volumes in a group must be in the same virtual pool. A volume group cannot have the same name as another volume group but can have the same name as any volume. A maximum of 256 volume groups can exist per system. If a volume group is being replicated, the maximum number of volumes that can exist in the group is 16.

Thin Provisioning

Thin provisioning is a virtual storage feature that allows a system administrator to overcommit physical storage resources. This allows the host system to operate as though it has more storage available than is actually allocated to it. When physical resources fill up, the administrator can add physical storage by adding additional disk groups on demand.

Automated Tiered Storage

Automated Tiered Storage is a virtual storage feature that 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 data can move to disks with higher performance.
- Infrequently accessed 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, which provide the best performance but also the highest cost.
- **Standard**—This middle tier uses enterprise-class spinning SAS disks, which provide good performance with mid-level cost and capacity.
- **Archive**—This lowest tier uses midline spinning SAS disks, which provide the lowest performance with the lowest cost and highest capacity.

SSD read cache

Unlike tiering, where a single copy of specific blocks of data resides in either spinning disks or SSDs, the Read Flash Cache (RFC) feature uses one SSD read-cache disk group per pool as a read cache for frequently accessed data only. Each read-cache disk group consists of one or two SSDs with a maximum usable capacity of 4TB. A separate copy of the data is also kept in spinning disks. Read-cache content is lost when a controller restart or failover occurs.

ADAPT (Autonomic Distributed Allocation Protection)

ADAPT is a RAID-based data protection level that maximizes flexibility, provides built in spare capacity, and allows for very fast rebuilds, large storage pools, and simplified expansion. All disks in the ADAPT disk group must be the same type (enterprise SAS, for example), and in the same tier, but can have different capacities. ADAPT is shown as a RAID level in the management interfaces.

General Best Practices

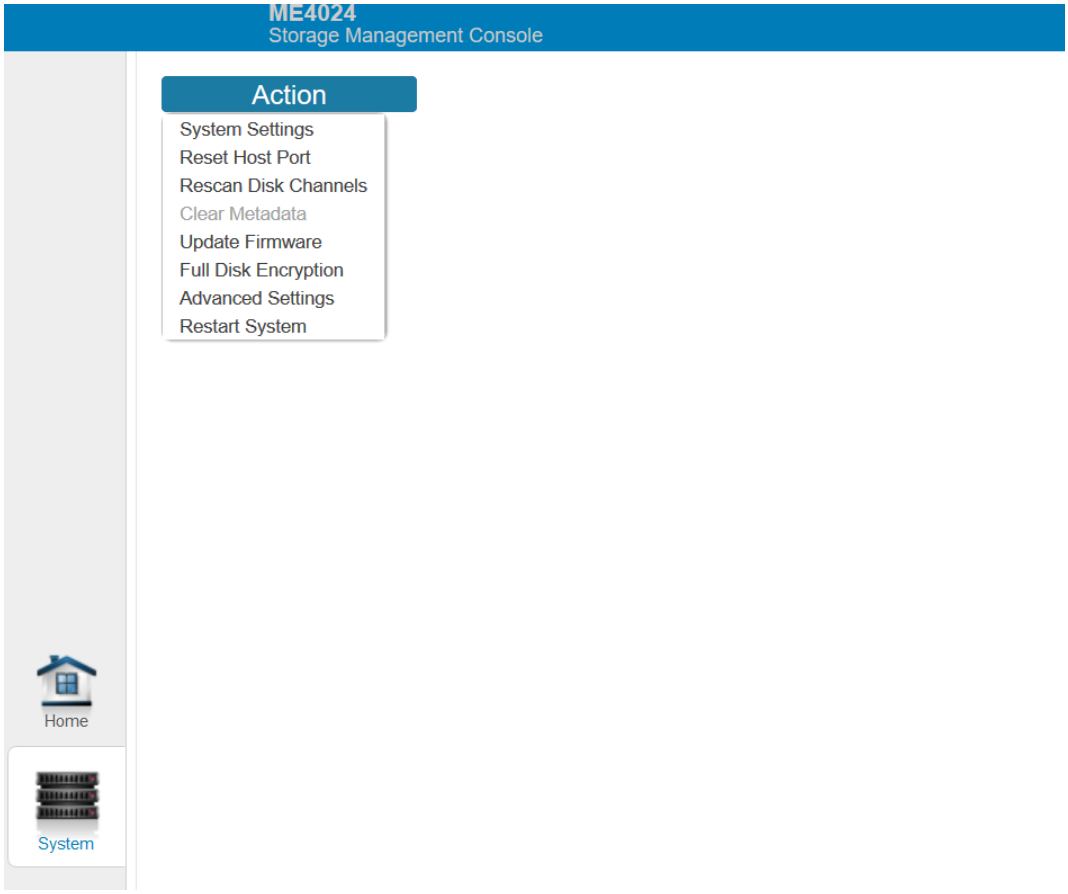
This section has some general practices when administering Powervault ME4 storage arrays

Become Familiar with Manuals

This document includes some information from manuals, for become familiar to your array reading all manuals are highly recommended. You can access Powervault ME4 public support page and documentation from here

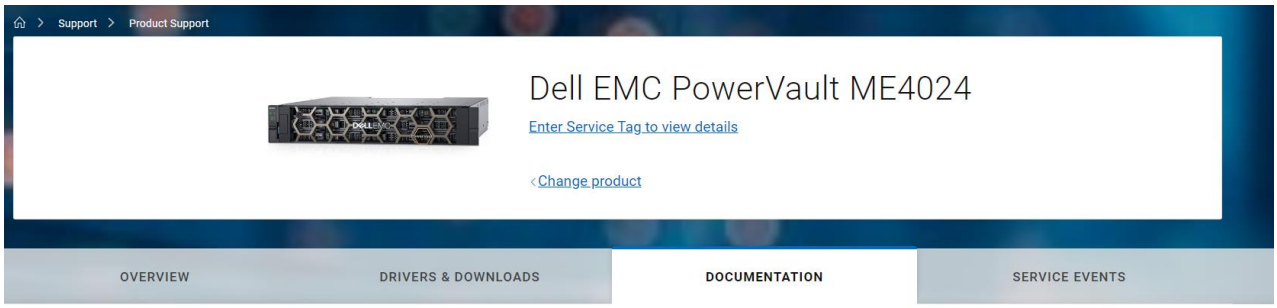
Stay with up-to-date Firmware

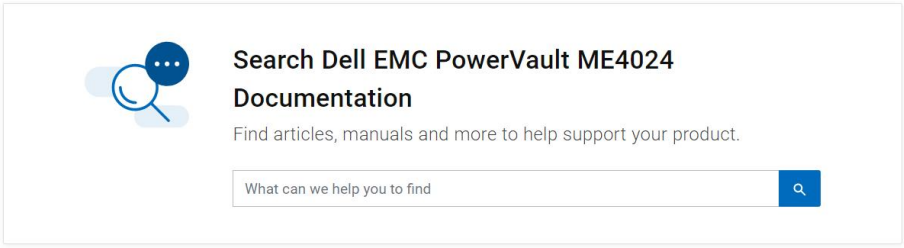
For better performance, reliability and gaining new futures, it's important to update your storages firmware regularly. You can find most up to date firmware at https://www.dell.com/support web site. You can update your array's firmware in MESM (ME Storage Manager) System Menu -> Action -> Update Firmware. For more details please check Admin Guide or this guide's "Best Practices for firmware update" section.

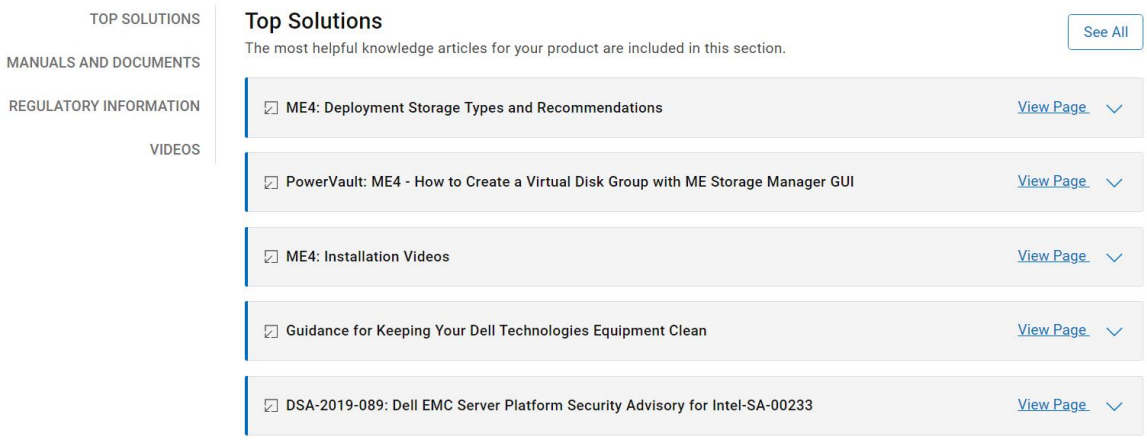


Always use supported configurations

Always check ME4 Support Matrix for supported configurations, Operating systems and array rules. Do not risk your data, critical applications with unsupported configurations. Dell EMC does not recommend or provide support for unsupported configurations. You can access latest support matrix from https://www.dell.com/support web site with your arrays array's service tag number or searching with your arrays model.







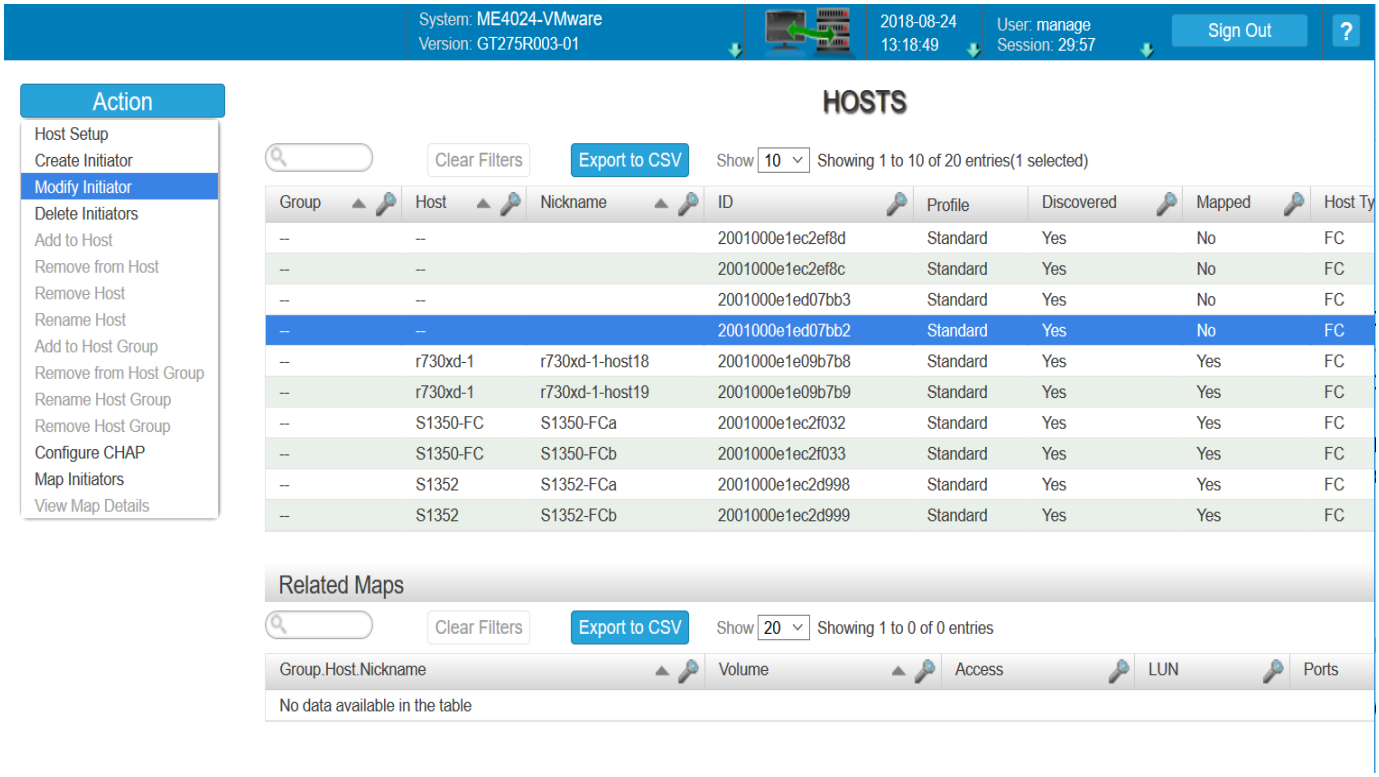
Host Information

The Hosts block shows how many host groups, hosts, and initiators are defined in the system. An initiator identifies an external port to which the storage system is connected. The external port may be a port in an I/O adapter in a server, or a port in a network switch. A host is a user-defined set of initiators that represents a server. A host group is a user-defined set of hosts for ease of management. If the external port is a switch and there is no connection from the switch to an I/O adapter, then no host information will be shown.

Identifying Your Hosts Easily

For easily identifying your hosts, it's highly recommended that using nicknames. recommended method for acquiring and renaming World Wide Names (WWNs) is to connect one cable at a time and then rename the WWN to an identifiable name. You can change it via MESM.

Select "***Hosts"*** on left pane. -> Click WWN you want to give nickname -> From "***Action***" menu -> select "***Modify Initiator***" -> Enter a Nickname for Initiator ->Click "***ok***"

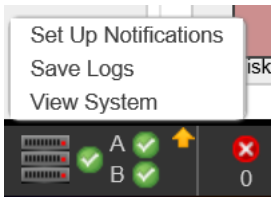


How to Monitor Array Health?

The health panel in the footer shows the current health of the system and each controller.

Hover the cursor over this panel to display the System Health panel, which shows the health state. If the system health is not OK, the System Health panel also shows information about resolving problems with unhealthy components.

The icon indicates that the panel has a menu. Click anywhere in the panel to display a menu to change notification settings, save log data, and view system information.



The Powervault ME4 storage systems can report their status through SNMP and/or Email.

Configuring E-mail and SNMP Notifications

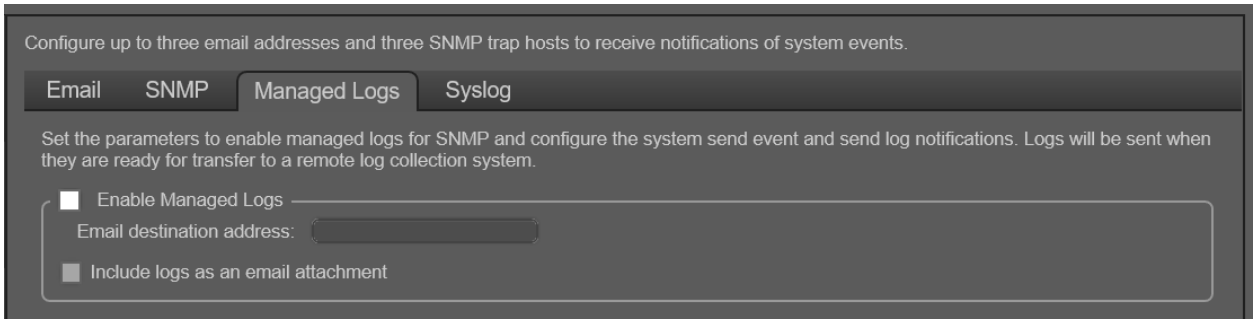
The Notifications tab provides options for you to set up and test several types of system notifications. These include:

- Configuring SMTP settings.
- Sending notifications to email addresses when events occur in the system.
- Enabling managed logs settings, which transfers log data to a log-collection system.
- Setting remote syslog notifications to allow events to be logged by the syslog of a specified host computer. Syslog is a protocol for sending event messages across an IP network to a logging server. This feature supports User Datagram Protocol (UDP) but not Transmission Control Protocol (TCP).
- Testing notifications.

You should enable at least one notification service to monitor the system. Email notifications can be sent to as many as three different email addresses. In addition to the normal email notification, Dell EMC recommends enabling managed logs with the Include logs as an email attachment option enabled. When this feature is enabled, the system automatically attaches the system log files to the managed log's email notifications that are sent. The managed log's email notification contains the logs for future diagnostic investigation.

Powervault ME4 systems has limited space for logs. When this log space is full, the oldest entries in the log are overwritten. For most systems, this space is adequate to allow for diagnosing issues seen on the system. The managed logs feature notifies the storage administrator that the logs are nearing a full state and that older information will soon get overwritten. The storage administrator can then choose to manually save the logs. If the Include logs as an email attachment check box is selected, the segment of logs that is nearing a full state is attached to the email notification. Managed logs attachments can be multiple megabytes in size.

Enabling the managed logs feature allows log files to be transferred from the storage system to a log-collection system to avoid losing diagnostic data. The Include logs as an email attachment option is disabled by default.



Send E-mail notifications

1. Perform one of the following to access the options in the Notifications tab:

- In the Home topic, select **Action -> System Settings**, then click **Notifications.**
- In the System topic, select **Action -> System Settings**, then click **Notifications.**
- In the footer, click the events panel and select **Set Up Notifications**.
- In the Welcome panel, select **System Settings**, and then click the **Notifications** tab

2. Select the Email tab and ensure that the SMTP Server and SMTP Domain options are set, as described in To configure SMTP settings in Administrator's Guide

3. Set the email notification:

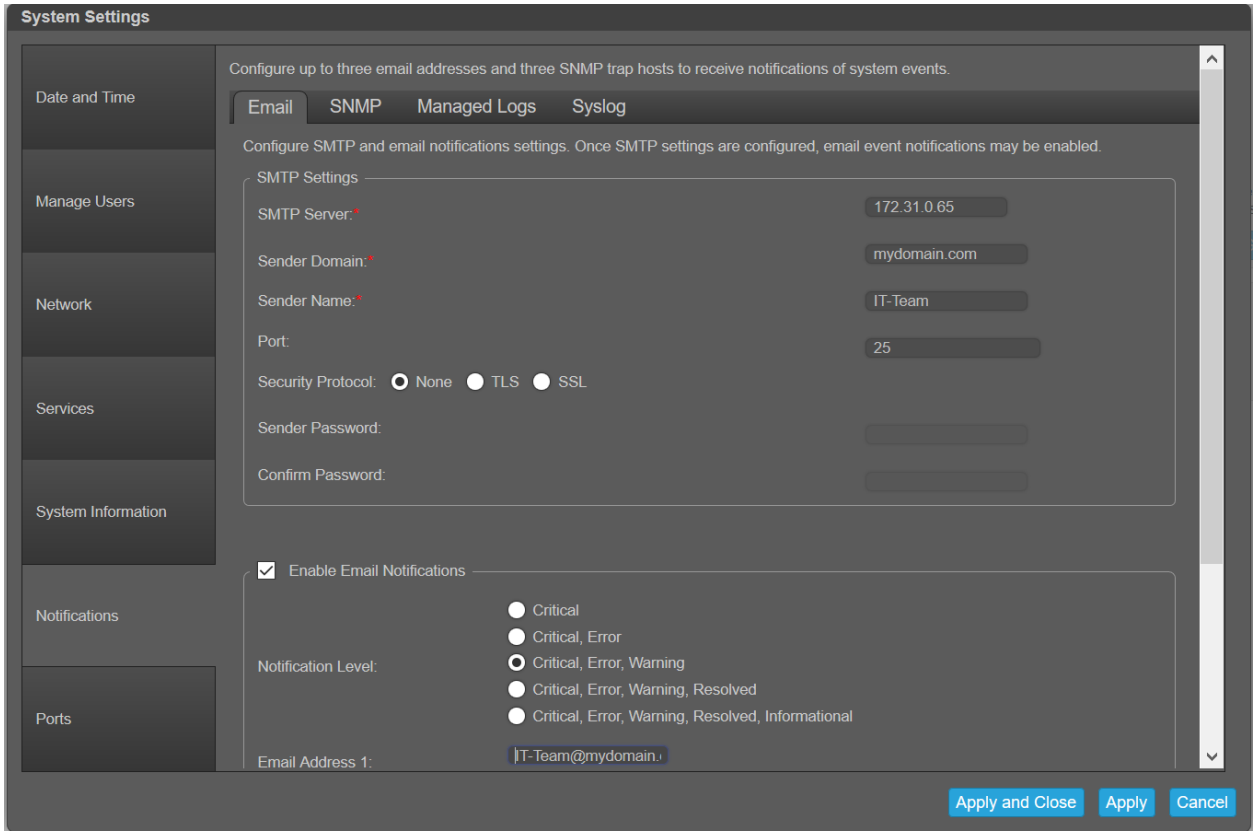
- To enable email notifications, select the **Enable Email Notifications** check box. This enables the notification level and email address fields.
- To disable email notifications, clear the **Enable Email Notifications** check box. This disables the notification level and email address fields.

4. If email notification is enabled, select the minimum severity for which the system should send email notifications: **Critical** (only); **Error** (and Critical); **Warning** (and Error and Critical); **Resolved** (and Error, Critical, and Warning); Informational (all).
5. If email notification is enabled, in one or more of the Email Address fields enter an email address to which the system should send notifications. Each email address must use the format *user-name@domain-name.* Each email address can have a maximum of 320 bytes. For example**: Admin@mydomain.com** or **IT-team@mydomain.com**.
6. Perform one of the following:

- To save your settings and continue configuring your system, click **Apply.**
- To save your settings and close the panel, click **Apply and Close.**

  A confirmation panel appears.

7. Click **OK** to save your changes. Otherwise, click **Cancel.**



Send notifications to SNMP trap hosts

1. Perform one of the following to access the options in the Notifications tab:

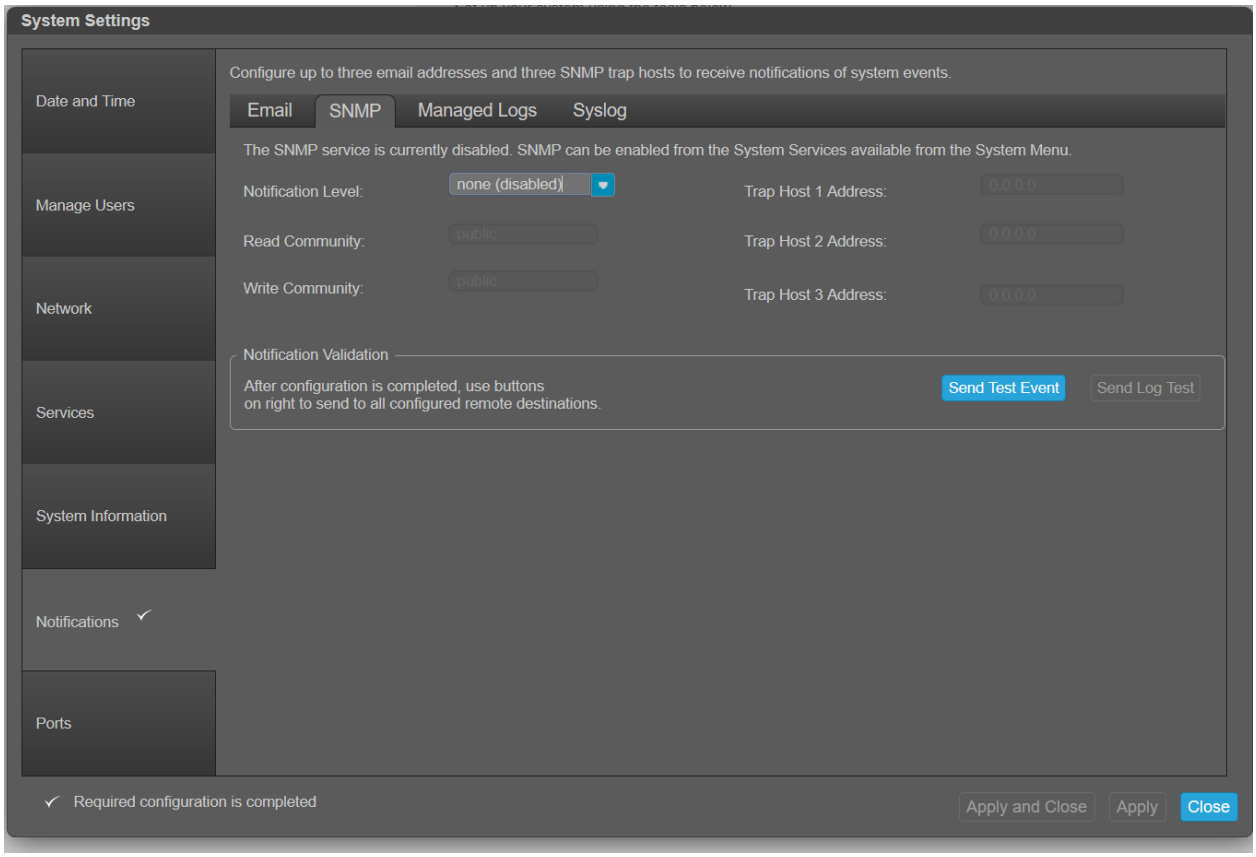
- In the Home topic, select **Action -> System Settings**, then click **Notifications.**
- In the System topic, select **Action -> System Settings**, then click **Notifications.**
- In the footer, click the events panel and select **Set Up Notifications**.
- In the Welcome panel, select **System Settings**, and then click the **Notifications** tab

2. Select the **SNMP** tab. If a message near the top of the panel informs you that the SNMP service is disabled, enable the service.
3. Select the minimum Notification Level severity for which the system should send email notifications: **Critical** (only); **Error** (and Critical); **Warning** (and Error and Critical); **Informational/Resolved** (all); or **none.**
4. In the **Read community** field, enter the SNMP read password for your network. This password is included in traps that are sent. The value is case sensitive and can have a maximum of 31 bytes. It can include any character except for the following: " < >
5. In the **Write community** field, enter the SNMP write password for your network. The value is case sensitive and can have a maximum of 31 bytes. It can include any character except for the following: " ' < >
6. In the **Trap Host Address** fields enter the network addresses of hosts that are configured to receive SNMP traps. The values can be IPv4 addresses, IPv6 addresses, or FQDNs.
7. Perform one of the following:

- To save your settings and continue configuring your system, click Apply.
- To save your settings and close the panel, click Apply and Close.

A confirmation panel appears.

8. Click **OK** to save your changes. Otherwise, click **Cancel.**



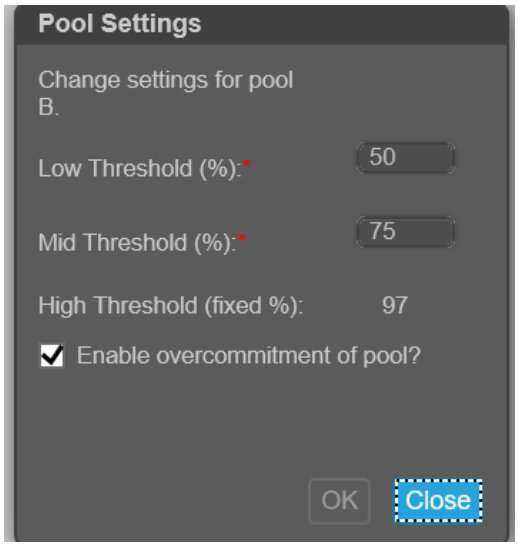
How to Provision Storage?

This section outlines the best methods for optimizing virtual storage features such as Thin Provisioning, automated tiering for Powervault ME4 series

Thin Provisioning

Thin provisioning is a virtual storage feature that allows a system administrator to overcommit physical storage resources. This allows the host system to operate as though it has more storage available than is actually allocated to it. When physical resources fill up, the administrator can add physical storage by adding additional disk groups on demand.

Overcommit is enabled by default. The overcommit setting lets you oversubscribe the physical storage (that is, provision volumes in excess of physical capacity). If you disable overcommit, you can provision virtual volumes only up to the available physical capacity. Overcommit is performed on a per-pool basis by using the Change Pool Settings option.



Each virtual pool has three thresholds for page allocation as a percentage of pool capacity. You can set the low and middle thresholds. The high threshold is automatically calculated based on the available capacity of the pool minus 200 GB of reserved space.

You can view and change settings that govern the operation of each virtual pool:

- Low Threshold. When this percentage of virtual pool capacity has been used, informational event 462 will be generated to notify the administrator. This value must be less than the Mid Threshold value. The default is 50 percent.
- Mid Threshold. When this percentage of virtual pool capacity has been used, event 462 will be generated to notify the administrator to add capacity to the pool. This value must be between the Low Threshold and High Threshold values. The default is 75 percent. If the pool is not overcommitted, the event will have Informational severity. If the pool is overcommitted, the event will have Warning severity.
- High Threshold. When this percentage of virtual pool capacity has been used, event 462 will be generated to alert the administrator to add capacity to the pool. This value is automatically calculated based on the available capacity of the pool minus 200 GB of reserved space. If the pool is not overcommitted, the event will have Informational severity. If the pool is overcommitted, the event will have Warning severity and the system will use write-through cache mode until virtual pool usage drops back below this threshold.
- Enable overcommitment of pools. This check box controls whether thin provisioning is enabled, and whether storage-pool capacity may exceed the physical capacity of disks in the system.

Note: *If the pool size is 500 GB or smaller, or the middle threshold is relatively high or both, the high threshold may not guarantee 200 GB of reserved space in the pool. The controller will not automatically adjust the low and middle thresholds in such cases.*

For a thin-provisioned volume mapped to a host, when data is deleted from the volume not all of the pages, or space associated with that data will be deallocated, or released. This is especially true for smaller files. To deallocate the pages in Windows, select the mapped volume and do either Perform a “Quick Format” or View its properties, select the Tools tab and Under **Defragmentation** click **Optimize.**

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 utilizing 1 MB or smaller pages. The ME4 Series array utilizes 4 MB pages, and therefore is incompatible with automatic unmap.

The command for executing unmap is a part of the esxcli command set of the ESXi OS. 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

Pool Balancing

In a storage system with two controller modules, try to balance the workload of the controllers. Each controller can own one virtual pool. Having the same number of disk groups and volumes in each pool will help balance the workload, increasing performance.

Quick Rebuild (ADAPT only)

Quick rebuild is a method for reconstructing virtual disk groups that reduces the time that user data is less than fully fault-tolerant after a disk failure in a disk group. Taking advantage of virtual storage knowledge of where user data is written, quick rebuild only rebuilds the data stripes that contain user data.

Typically, storage is only partially allocated to volumes, so the quick-rebuild process completes significantly faster than a standard RAID rebuild. Data stripes that have not been allocated to user data are scrubbed in the background, using a lightweight process that allows future data allocations to be more efficient.

After a quick rebuild, a scrub starts on the disk group within a few minutes after the quick rebuild completes. Quick rebuild is only usable in ADAPT not in traditional RAID

Modifying Virtual Volumes

A virtual disk group requires the specification of a set of disks, RAID level, disk group type, pool target (A or B), and a name. If the virtual pool does not exist at the time of adding the disk group, the system will automatically create it. Multiple disk groups (up to 16) can be added to a single virtual pool.

You can expand a volume. If a virtual volume is not a secondary volume involved in replication, you can expand the size of the volume but not make it smaller. If a linear volume is neither the parent of a snapshot nor a primary or secondary volume, you can expand the size of the volume but not make it smaller. Because volume expansion does not require I/O to be stopped, the volume can continue to be used during expansion.

The recommended method to expand the volume size is to add a new virtual disk group with the same RAID level, capacity disks, and physical number of disks as the existing virtual disk group in the same tier.

Automated Tiered Storage

Automated Tiered Storage is a virtual storage feature that 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 data can move to disks with higher performance.
- Infrequently accessed 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, which provide the best performance but also the highest cost.

Standard—This middle tier uses enterprise-class spinning SAS disks, which provide good performance with mid-level cost and capacity.

Archive—This lowest tier uses midline spinning SAS disks, which provide the lowest performance with the lowest cost and highest capacity.

When the status of a disk group in the Performance Tier becomes critical (CRIT), the system will automatically drain data from that disk group to disk groups using spinning disks in other tiers providing that they can contain the data on the degraded disk group. This occurs because similar wear across the SSDs is likely, so more failures may be imminent.

If a system only has one class of disk, no tiering occurs. However, automated tiered storage rebalancing happens when adding or removing a disk group in a different tier.

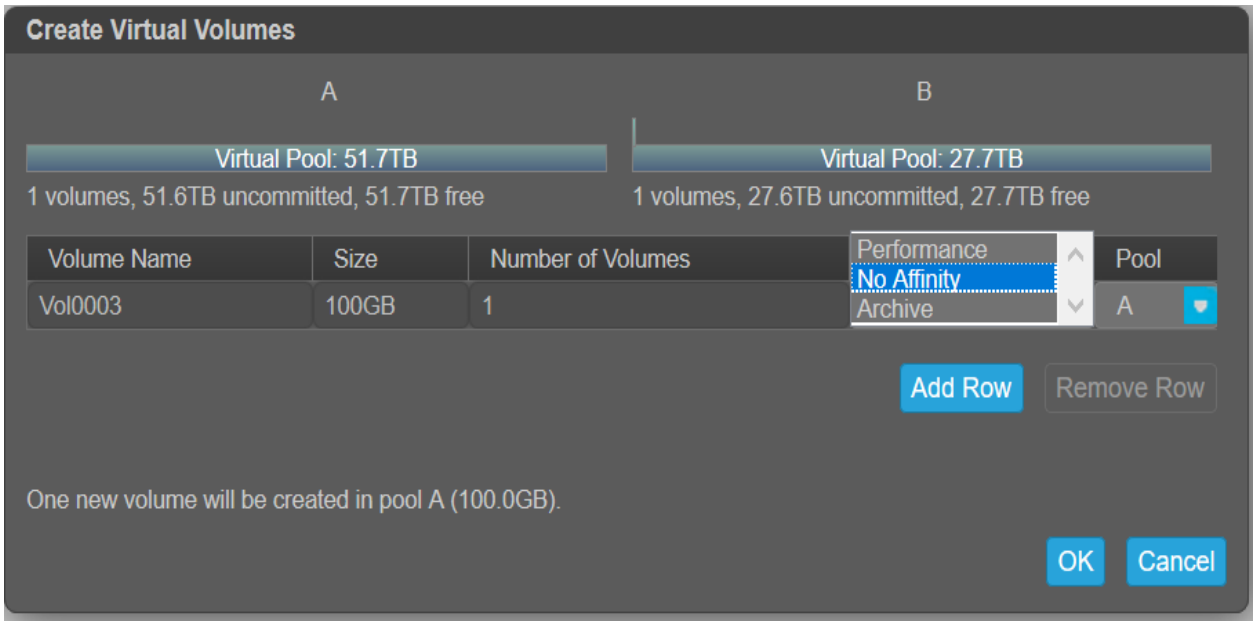
How Tiering Works?

Auto-tiering uses a concept called “paging”. User volumes are logically broken down into small, 4MB chunks called pages. Pages are ranked based upon an algorithm. The page rank is used to very efficiently select good pages to move between tiers. The result is that pages can be migrated between tiers automatically such that I/O’s are optimized in real-time.

- Tiering algorithm runs every 5 seconds.
- Only 80 MB of data is migrated every five seconds to avoid degrading system throughput.
- Frequently accessed data moved up to higher performance disks
- Infrequently access data moved down to lower performance disks
- Pages are only migrated down if room needed for highly ranked page
- Single copy of specific blocks of data resides in either spinning drives or SSDs

Volume Tier Affinity Feature

It is not possible to pin a volume to any tier. However, you can change volume's affinity settings when you are creating it or after you created.



No Affinity - The default strategy in Powervault ME4 is to prefer the highest spinning disk (non-SSD) tiers for new sequential writes and the highest tier available (including SSD) for new random writes. As data is later accessed by the host application it will be moved to the most appropriate tier based on demand with 'hot' data being promoted up towards the highest performance tier and 'cold' data being demoted downwards to the lower spinning disk-based tiers. This standard strategy will be followed for data on volumes set to 'No Affinity'.

Performance **-** For data on volumes set to the 'Performance' affinity the standard strategy will be followed for all new writes however subsequent access to that data will have a lower threshold for promotion upwards making it more likely for that data to be available on the higher performance tiers. Preferential treatment will be provided to 'hot' data that has performance affinity at the SSD tier making it more likely for archive or no affinity data to be demoted out of the SSD tier to make room. This is useful for volumes where you know the data will be in demand and want to ensure that it has priority treatment for promotion to and retention in your highest performance tier.

Archive - For volumes that are set to the 'Archive' affinity all new writes will be initially placed in the archive tier so long as space is available – if no space is available they will be placed on the next higher tier available. Subsequent access to that data will allow for its promotion to the performance tiers as it becomes 'hot' however it will have a lower threshold for demotion and will be moved out of the highest performance SSD tier if there is a need to promote 'hot' data up from a lower tier.

Best Practice in Tier Setup

In general, it is best to have two tiers instead of three tiers. The highest tier will nearly fill before using the lowest tier. The highest tier must be 95% full before the controller will evict cold pages to a lower tier to make room for incoming writes.

Typically, you should use tiers with SSDs and 10K/15K disks, or tiers with SSDs and 7K disks. An exception may be if you need to use both SSDs and faster spinning disks to hit a combination of price for performance, but you cannot hit your capacity needs without the 7K disks; this should be rare.

Recommended setting for Volume Tier Affinity is "***No Affinity***" for most configurations. This setting attempts to balance the frequency of data access, disk cost, and disk availability by moving the volume's data to the appropriate tier.

If the virtual volume uses mostly random or burst low-latency workloads such as online transaction processing (OLTP), virtual desktop infrastructure (VDI), or virtualization environments, recommended setting is "***Performance".*** This setting keeps as much of the volume's data in the performance tier for as long as possible.

If the virtual volume contains infrequently accessed workloads such as backup data or email archiving, recommended setting is "***Archive***". This option keeps as much of the volume's data in the archive tier for as long as possible.

Best Practices for High Availability

High availability is always advisable to protect assets in the event of a device failure. This section gives you some options and information to help you in the event of a failure.

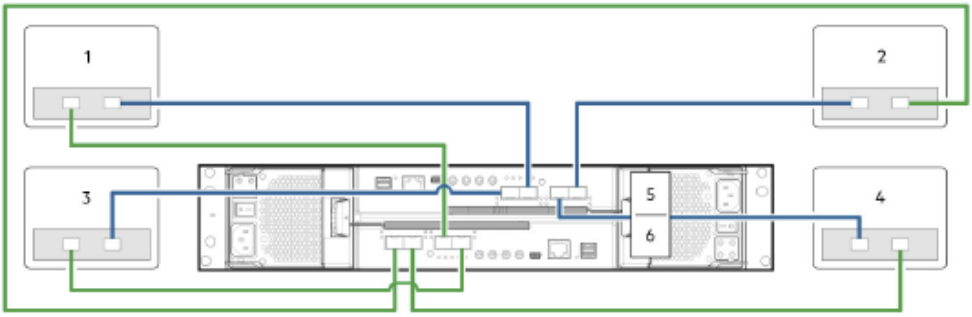
Volume Mapping

In both virtual and linear storage, if the owning controller fails, the partner controller assumes temporary ownership of the pool and resources owned by the failed controller. If a fault-tolerant cabling configuration, with appropriate mapping, is used to connect the controllers to hosts, LUNs for both controllers are accessible through the partner controller so I/O to volumes can continue without interruption.

The best practice is to map volumes to two ports on each controller to take advantage of load balancing and redundancy.

To avoid multiple hosts mounting the volume and causing corruption, the hosts must be cooperatively managed, such as by using cluster software.
If multiple hosts mount a volume without being cooperatively managed, volume data is at risk for corruption. To control access by specific hosts, you can create an explicit mapping. An explicit mapping can use different access mode, LUN, and port settings to allow or prevent access by a host to a volume, overriding the default mapping. When an explicit mapping is deleted, the volume's default mapping takes effect.

Direct Attach Cabling example



ME4 Series 2U direct attach- four servers /one HBA per server / dual path

1 – Server 1	2- Server 2
3 - Server 3	4 – Server 4
5 – Controller Module A	6 – Controller Module B

For more detailed information about host cabling please check the System Deployment Guide

Physical port selection

In a system configured to use either all FC or all iSCSI ports, use the ports in the following order:

A0, B0
A2, B2
A1, B1
A3, B3

The reason for doing so is that each pair of ports (A0, A1 or A2, A3) are connected to a dedicated CNC chip. If you are not using all four ports on a controller, it is best to use one port from each pair (A0, A2) to ensure better I/O balance on the front end.

Multipath Configuration and Multipath Software

ME4 Series storage systems comply with the SCSI-3 standard for Asymmetrical Logical Unit Access (ALUA). ALUA-compliant storage systems provide optimal and non-optimal path information to the host during device discovery. To implement ALUA, you must configure your servers to use multipath I/O (MPIO).

The Microsoft MPIO feature needs to be installed prior to connecting to Dell EMC ME4 Series Storage arrays.

For Red Hat and Suse Linux Device Mapper multipath is required for multipath support.

Please check Powervault ME4 Support Matrix for details.

Snapshots

The system can create snapshots of virtual volumes up to the maximum number supported by your system. Snapshots provide data protection by enabling you to create and save source volume data states at the point in time when the snapshot was created. Snapshots can be created manually, or you can schedule snapshot creation. After a snapshot has been created, the source volume can be expanded.
To view the maximum number of snapshots for your system, see System configuration limits in Administrator's Guide. When you reach the maximum number of snapshots for your system, before you can create a new snapshot, you must delete an existing snapshot.

You need to determine how you manage snapshots on virtual volumes on pools that have overcommit enabled. Powervault ME4 has two options for setting the frequency of snapshot management:

1) If you already maintain snapshots, then you probably do not need to change anything. The system automatically sets the limit at 10% of the pool and only notifies you if a threshold is crossed.

2) The `set snapshot-space` CLI command enables you to set the percent of the pool that can be used for snapshots (the snapshot space). Optionally, you can specify a limit policy to enact when the snapshot space reaches the percentage. You can set the policy to either notify you via the event log that the percentage has been reached (in which case the system continues to take snapshots, using the general pool space), or to notify you and trigger automatic deletion of snapshots. If automatic deletion is triggered, snapshots are deleted according to their configured retention priority. For more information, see the CLI documentation

Dual Power Supplies

All Powervault ME4 storages ship with redundant power supplies. Each redundant power supply module requires power from an independent source or a rack power distribution unit with Uninterruptible Power Supply (UPS). 2U enclosures use standard AC power and the 5U84 enclosure requires high-line (high-voltage) AC power

Fault Tolerance (Reverse) Cabling

Reverse cabling allows any drive enclosure to fail—or be removed—while maintaining access to other enclosures. Fault tolerance and performance requirements determine whether to optimize the configuration for high availability or high performance when cabling.

Reverse cabling connections example between a 5U controller enclosure and 5U expansion enclosures

1 Controller module A (0A)
2 Controller module B (0B)
3 Expansion module 1A
4 Expansion module 1B
5 Expansion module 2A
6 Expansion module 2B
7 Expansion module 3A
8 Expansion module 3B

Please check Deployment Guide for other reverse cabling examples for 2U enclosures.

SMART

Self-Monitoring Analysis and Reporting Technology (SMART) provides data that enables you to monitor disks and analyze why a disk failed. When SMART is enabled, the system checks for SMART events one minute after a restart and every five minutes thereafter. SMART events are recorded in the event log.

Please check Administrator's Guide for how to enable SMART.

Scrubbing

The system-level Disk Group Scrub option automatically checks all disk groups for disk defects. If this option is disabled, you can still perform a scrub on a selected disk group. Scrub analyzes the selected disk group to find and fix disk errors. It will fix parity mismatches for RAID 3, 5, 6, 50, and ADAPT; find but not fix mirror mismatches for RAID 1 and 10; and find media errors for all RAID levels.

Scrub can last over an hour, depending on the size of the disk group, the utility priority, and the amount of I/O activity. However, a manual scrub performed by Scrub Disk Group is typically faster than a background scrub performed by Disk Group Scrub. You can use a disk group while it is being scrubbed. When a scrub is complete, event 207 is logged and specifies whether errors were found and whether user action is required.

Autonomic Distributed Allocation Protection (ADAPT)

ADAPT is a RAID-based data protection level that maximizes flexibility, provides built in spare capacity, and allows for very fast rebuilds, large storage pools, and simplified expansion. All disks in the ADAPT disk group must be the same type (enterprise SAS, for example), and in the same tier, but can have different capacities. ADAPT is shown as a RAID level in the management interfaces.

ADAPT disk groups use all available space to maintain fault tolerance, and data is spread evenly across all the disks. When new data is added, new disks are added, or the system recognizes that data is not distributed across disks in a balanced way, it moves the data to maintain balance across the disk group.
Reserving spare capacity for ADAPT disk groups is automatic since disk space dedicated to sparing is spread across all disks in the system. In the case of a disk failure, data will be moved to many disks in the disk group, allowing for quick rebuilds and minimal disruption to I/O.

One of the key differences between ADAPT and traditional RAID groups is the width that arrays can be constructed. RAID 5 and 6 can be applied up to a width of 16 drives. However, whilst ADAPT widths are a minimum of 12, the maximum is 128 making the potential drive group width and therefore size much bigger than traditional R6. This has significant implications especially when potential topologies of a ME4084 are considered. With this ability one can consider how to layout the drive groups on a ME4084. For example, R6 would give 5 * 16 disk groups and 4 spares.

With ADAPT: 2 x 48 disk groups or 4 x 24 disk groups would be considered optimal

Note 1 x 84 whilst technically possible is not optimal as controllers own disk groups and therefore this topology does now allow the performance of both controllers to be contributed to the solution

For more information about ADAPT please check *PowerVault ME4 Series ADAPT Software Whitepaper*

Hot Spares

Spare disks are unused disks in your system that you designate to automatically replace a failed disk, restoring fault tolerance to disk groups in the system. Types of spares include:

Dedicated spare: Reserved for use by a specific linear disk group to replace a failed disk. Most secure way to provide spares for disk groups, but expensive to reserve a spare for each disk group.

Global spare. Reserved for use by any fault-tolerant disk group to replace a failed disk.

Dynamic spare. Available compatible disk that is automatically assigned to replace a failed disk in a fault-tolerant disk group.

When a disk fails, the system looks for a dedicated spare first. If it does not find a dedicated spare, it looks for a global spare. If it does not find a compatible global spare and the dynamic spares option is enabled, it takes any available compatible disk

Note: You cannot designate spares for ADAPT disk groups. For information on how ADAPT disk groups manage sparing Please check Administrator's Guide . A best practice is to designate spares for use if disks fail. Dedicating spares to disk groups is the most secure method, but it is also expensive to reserve spares for each disk group. Alternatively, you can enable dynamic spares or assign global spares.

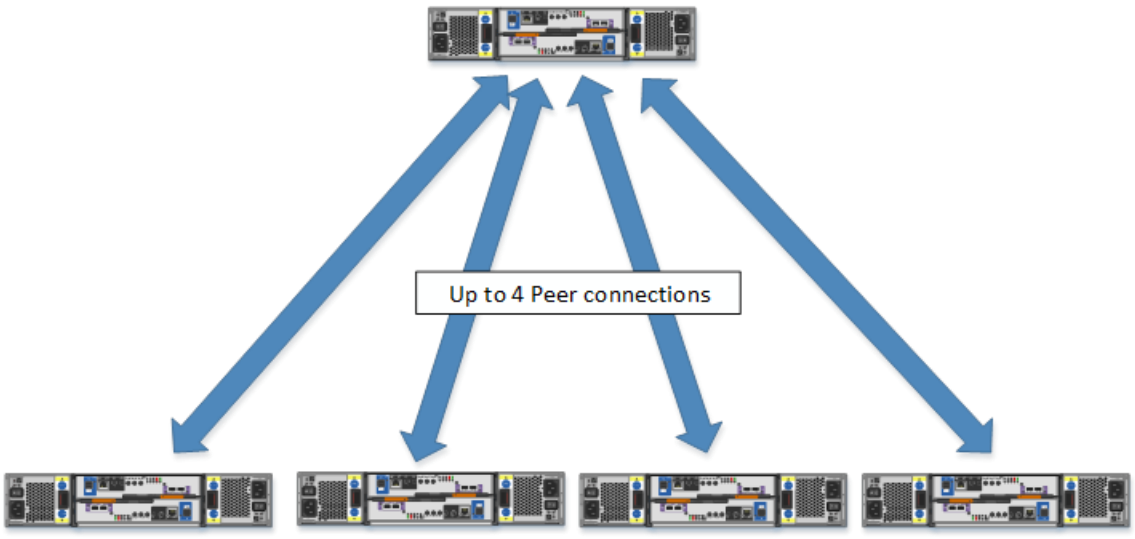
Virtual Volume Replication

Replication is a feature for disaster recovery. This feature performs asynchronous replication of block-level data from a volume in a primary system to a volume in a secondary system by creating an internal snapshot of the primary volume and copying the changes to the data since the last replication to the secondary system via FC or iSCSI links.

The two associated standard volumes form a replication set, and only the primary volume (source of data) can be mapped for access by a server. Both must be connected through switches to the same fabric or network (no direct attach). The server accessing the replication set need only be connected to the primary system. If the primary system goes offline, a connected server can access the replicated data from the secondary system.

The Powervault ME4 allows for up to four peer connections (replication partner), these are bi-directional connections, e.g.: a primary system can have up to four replication destinations (DR-sites), or the other way around, up to four prod systems replicate to the same DR site(many-to-one or one-to-many).

However, on a volume basis, a particular volume can only be part of one replica set, hence can only be replicated to one replication destination (one-to-one, not one-to-many).



For Cabling considerations of replication Please check Deployment Guide. For other Replication prerequisites Please check Administrator's Guide .

Best Practices for SSDs

The use of SSDs (solid-state drives) can greatly enhance the performance of a system. Since the SSDs do not have moving parts, data that is random in nature can be accessed much faster.

Use SSDs for Randomly Accessed Data

You can use SSDs for virtual disk groups. When combined with virtual disk groups that consist of other classes of disks, improved read and write performance is possible through automated tiered storage. Alternatively, you can use one or two SSDs in read-cache disk groups to increase read performance for pools without a Performance tier. The application workload of a system determines the percentage of SSDs of the total disk capacity that is needed for best performance.

Database indexes and Temporary Database files are good examples for SSD usage and you can get benefit of SSDs. Another great example for SSD usage is Virtual Desktop Infrastructures (VDI). For More details about VMware Horizon View VDI environments with Powervault ME4; please check *" Dell EMC PowerVault ME4024 and VMware Horizon View with 1300 Persistent VDI Users "* document at support web site.

Best Practices for Performance

Volume Cache Options

You can set options that optimize reads and writes performed for each volume. It is recommended that you use the default settings.

You can enable and disable the write-back cache for each volume. By default, volume write-back cache is enabled. Because controller cache is backed by supercapacitor technology, if the system loses power, data is not lost. For most applications, this is the preferred setting.

CAUTION: Only disable write-back caching if you fully understand how the host operating system, application, and adapter move data. Used incorrectly, write-back caching can hinder system performance.

Using write-back or write-through caching

When modifying a volume, you can change its write-back cache setting. Write-back is a cache-writing strategy in which the controller receives the data to be written to disks, stores it in the memory buffer, and immediately sends the host operating system a signal that the write operation is complete, without waiting until the data is actually written to the disk. Write-back cache mirrors all of the data from one controller module cache to the other. Write-back cache improves the performance of write operations and the throughput of the controller.

When write-back cache is disabled, write-through becomes the cache-writing strategy. Using write-through cache, the controller writes the data to the disks before signaling the host operating system that the process is complete. Write-through cache has lower write throughput performance than write-back, but it is the safer strategy, with minimum risk of data loss on power failure. However, write-through cache does not mirror the write data because the data is written to the disk before posting command completion and mirroring is not required. You can set conditions that cause the controller to change from write-back caching to write-through caching. For more information, see Changing system cache settings in Administrator's Guide .

In both caching strategies, active-active failover of the controllers is enabled.

NOTE: The best practice for a fault-tolerant configuration is to use write-back caching.

Cache optimization mode

You can also change the optimization mode.

Standard: This controller cache mode of operation is optimized for sequential and random I/O and is the optimization of choice for most workloads. In this mode, the cache is kept coherent with the partner controller. This mode gives you high performance and high redundancy. This is the default.

No-mirror: In this mode of operation, the controller cache performs the same as the standard mode with the exception that the cache metadata is not mirrored to the partner. While this improves the response time of write I/O, it comes at the cost of redundancy. If this option is used, the user can expect higher write performance but is exposed to data loss if a controller fails.

CAUTION: Changing the cache optimization setting while I/O is active can cause data corruption or loss. Before changing this setting, quiesce I/O from all initiators.

Optimizing read-ahead caching

You can optimize a volume for sequential reads or streaming data by changing its read-ahead cache settings.

You can change the amount of data read in advance. Increasing the read-ahead cache size can greatly improve performance for multiple sequential read streams.

- The **Adaptive** option works well for most applications: it enables adaptive read-ahead, which allows the controller to dynamically calculate the optimum read-ahead size for the current workload.
- The **Stripe** option sets the read-ahead size to one stripe. The controllers treat NRAID and RAID-1 disk groups internally as if they have a stripe size of 512 KB, even though they are not striped.
- Specific size options let you select an amount of data for all accesses.
- The **Disabled** option turns off read-ahead cache. This is useful if the host is triggering read ahead for what are random accesses. This can happen if the host breaks up the random I/O into two smaller reads, triggering read ahead.

NOTE: Only change read-ahead cache settings if you fully understand how the host operating system, application, and adapter move data so that you can adjust the settings accordingly.

Other methods to enhance array performance

There are other methods to maximize performance of the Powervault ME4 series beside optimizing cache settings, the performance of the array can be maximized by using the following techniques.

Power-of-2 Method

You can configure max “2” pools in dual controller systems and each controller owns a pool. You need to balance disks between two pools. In other words, you need to use magical number “2” which means divide your disk quantity “2” for load balancing in each controller.

For example;

If you have 13 x SSD drives and 25 x SAS drives and 49 NL-SAS drives in ME4 storage array;

Pool 1 : 6 x SSD; 12 x SAS ; 24 x NL-SAS

Pool 2 : 6 x SSD; 12 x SAS ; 24 x NL-SAS

And assign at least 1 x global spares for each type of drives.

Disk groups in a pool

For better efficiency and performance, use similar disk groups in a pool.

- Disk count balance: For example, with 20 disks, it is better to have two 8+2 RAID-6 disk groups than one 10+2 RAID-6 disk group and one 6+2 RAID-6 disk group.
- RAID balance: It is better to have two RAID-5 disk groups than one RAID-5 disk group and one RAID-6 disk group.
- In terms of the write rate, due to wide striping, tiers and pools are as slow as their slowest disk groups.
- All disks in a tier should be the same type. For example, use all 10K disks or all 15K disks in the Standard tier.
- Create more small disk groups instead of fewer large disk groups.
- Each disk group has a write queue depth limit of 100. This means that in write-intensive applications this architecture will sustain bigger queue depths within latency requirements.
- Using smaller disk groups will cost more raw capacity. For less performance-sensitive applications, such as archiving, bigger disk groups are desirable.

Which RAID Level should you use?

There is not only one answer to this question. It depends environment. The following table describes the characteristics and use cases of each RAID level.

RAID level	Protection	Performance	Capacity	Application use cases	Suggested disk speed
RAID 1 RAID 10	Protects against up to one disk failure per mirror set	Great random I/O performance	Poor: 50% fault tolerance capacity loss	Databases, OLTP, Exchange Server	10K, 15K, 7K
RAID 5	Protects against up to one disk failure per RAID set	Good sequential I/O performance, moderate random I/O performance	Great: One-disk fault tolerance capacity loss	Big data, media and entertainment (ingest, broadcast, and past production)	10K, 15K, lower capacity 7K
RAID 6	Protects against up to two disk failures per RAID set	Moderate sequential I/O performance, poor random I/O performance	Moderate: Two disk fault tolerance capacity loss	Archive, parallel distributed file system	High capacity 7K

Disk count per RAID level

The following table shows recommended disk counts for RAID-6 and RAID-5 disk groups. Each entry specifies the total number of disks and the equivalent numbers of data and parity disks in the disk group. Note that parity is actually distributed among all the disks.

RAID Level	Total Disks	Data disks (equivalent)	Parity disks (equivalent)
RAID 6	4	2	2
	6	4	2
	10	8	2
RAID 5	3	2	1
	5	4	1
	6	8	1

To ensure best performance with sequential workloads and RAID-5 and RAID-6 disk groups, use a power-of-two data disks.

The controller breaks virtual volumes into 4-MB pages, which are referenced paged tables in memory. The 4-MB page is a fixed unit of allocation. Therefore, 4-MB units of data are pushed to a disk group. A write performance penalty is introduced in RAID-5 or RAID-6 disk groups when the stripe size of the disk group isn't a multiple of the 4-MB page.

- Example 1: Consider a RAID-5 disk group with five disks. The equivalent of four disks provide usable capacity, and the equivalent of one disk is used for parity. Parity is distributed among disks. The four disks providing usable capacity are the data disks and the one disk providing parity is the parity disk. In reality, the parity is distributed among all the disks, but conceiving of it in this way helps with the example.

  Note that the number of data disks is a power of two (2, 4, and 8). The controller will use a 512-KB stripe unit size when the data disks are a power of two. This results in a 4-MB page being evenly distributed across two stripes. This is ideal for performance.

- Example 2: Consider a RAID-5 disk group with six disks. The equivalent of five disks now provides usable capacity. Assume the controller again uses a stripe unit of 512-KB. When a 4-MB page is pushed to the disk group, one stripe will contain a full page, but the controller must read old data and old parity from two of the disks in combination with the new data in order to calculate new parity. This is known as a read-modify-write, and it's a performance killer with sequential workloads. In essence, every page push to a disk group would result in a read-modify-write.

  To mitigate this issue, the controllers use a stripe unit of 64-KB when a RAID-5 or RAID-6 disk group isn't created with a power-of-two data disks. This results in many more full-stripe writes, but at the cost of many more I/O transactions per disk to push the same 4-MB page.

Gauging the percentage of life remaining for SSDs

An SSD can be written and erased a limited number of times. Through the SSD Life Left disk property, you can gauge the percentage of disk life remaining. This value is polled every 5 minutes. When the value decreases to 20%, an event is logged with Informational severity. This event is logged again with Warning severity when the value decreases to 5%, 2% or 1%, and 0%. If a disk crosses more than one percentage threshold during a polling period, only the lowest percentage will be reported. When the value decreases to 0%, the integrity of the data is not guaranteed. To prevent data integrity issues, replace the SSD when the value decreases to 5% of life remaining.

All-flash array

The all-flash array feature, enabled by default, allows systems to run exclusively with disk groups that consist of SSDs, providing the ability to have a homogeneous SSD-only configuration. Systems using an all-flash array have one tier that consists solely of SSDs. If a system includes disk groups with spinning disks, the disk groups must be removed before the all-flash array feature can be used.

SSD read cache

Unlike tiering, where a single copy of specific blocks of data resides in either spinning disks or SSDs, the Read Flash Cache (RFC) feature uses one SSD read-cache disk group per pool as a read cache for frequently accessed data only. Each read-cache disk group consists of one or two SSDs with a maximum usable capacity of 4TB. A separate copy of the data is also kept in spinning disks. Read-cache content is lost when a controller restart or failover occurs. Taken together, these attributes have several advantages:

- 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 does not need to be fault tolerant, potentially lowering system cost.
- Controller read cache is effectively extended by two orders of magnitude, or more.

When a read-cache group consists of one SSD, it automatically uses NRAID. When a read-cache group consists of two SSDs, it automatically uses RAID 0.

Full Disk Encryption (FDE)

A system and the FDE-capable disks in the system are initially unsecured but can be secured at any point. Until the system is secured, FDE-capable disks function exactly like disks that do not support FDE.

Enabling FDE protection involves setting a passphrase and securing the system. Data that was present on the system before it was secured is accessible in the same way it was when it was unsecured. However, if a disk is transferred to an unsecured system or a system with a different passphrase, the data is not accessible.

CAUTION: Do not change FDE configuration settings while running I/O. Temporary data unavailability may result. Also, the intended configuration change might not take effect.

IMPORTANT: Be sure to record the passphrase as it cannot be recovered if lost.

Best practices for firmware updates

Controller modules, expansion modules, and disk drives contain firmware that operate them. As newer firmware versions become available, they may be installed at the factory or at a customer maintenance depot or they may be installed by storage-system administrators at customer sites.

- In the health panel in the footer, verify that the system health status is OK. If the system health status is not OK, view the Health Reason value in the health panel in the footer and resolve all problems before you update firmware. For information about the health status. Please check this document's "How to Monitor Array Health? "part.

- Run the `check firmware-upgrade-health` CLI command before upgrading firmware. This command performs a series of health checks to determine whether any conditions exist that need to be resolved before upgrading firmware. Any conditions that are detected are listed with their potential risks. For information about this command, see the CLI Reference Guide.

- If any unwritten cache data is present, firmware update will not proceed. Before you can update firmware, unwritten data must be removed from cache. See information about the clear cache command in the CLI Reference Guide.

- If a disk group is quarantined, resolve the problem that is causing the component to be quarantined before updating firmware.

- To ensure success of an online update, select a period of low I/O activity. This helps the update complete as quickly as possible and avoids disruption to host and applications due to timeouts.

Attempting to update a storage system that is processing a large, I/O-intensive batch job may cause hosts to lose connectivity with the storage system.

When planning a firmware upgrade;

- Online firmware upgrades are performed while host I/O being processed, and this time frame performance of array can impact. Select appropriate time frame for upgrade operation especially when host I/O activity is low.
- Spare at least 30 minutes for firmware upgrades.
- Please ensure that both controller's management ports' ethernet connection available before start upgrade process.

CAUTION: Do not perform a power cycle enclosures or restart a controller during the firmware update. If the update is interrupted or there is a power failure, the disk drive might become inoperative. If this occurs, contact technical support.

NOTE: Expansion firmware is updated automatically with controller updates.

Updating disk-drive firmware

You can update disk-drive firmware by loading a firmware file obtained from Dell Support Web Site

A dual-ported disk drive can be updated from either controller. Disk Drive Firmware update is an OFFLINE process. Stop I/O to the storage system. During the update all volumes will be temporarily inaccessible to hosts. If I/O is not stopped, mapped hosts will report I/O errors. Volume access is restored after the update completes.

Please check Administrator's Guide for more details.