WHITEPAPER

Dell PowerVault ME5 Series ADAPT Software

This white paper identifies and explains the working architecture and operational components of the ADAPT software with PowerVault ME5 storage arrays.

February 2022

Introduction

Dell Technologies continues to innovate its PowerVault ME5 (ME5) controller software to meet the demands of its customers by improving its ADAPT data protection technology from its initial release in ME4. ADAPT – Autonomic Distributed Allocation Protection Technology - is Dell Technologies' erasure encoding solution which is an alternative to traditional RAID types (RAID 0, 1, 5, 6, etc.) with a protection scheme that distributes the parity across a larger set of HDDs or SSDs - providing better data protection, scalability, and other advantages. ME5 includes a wider stripe and therefore reduces parity overhead. This feature brief provides an overview of this software and highlights its key features and benefits including differences between ME4 and ME5.

What does ADAPT software solve?

One of the biggest problems today with RAID systems is that with the ever-increasing capacity of drives, rebuild times are also increasing. This increases the window of time that further drives may fail within the rebuild process. At this point the drive group is open to losing integrity. Therefore, as good practice we want to keep rebuild times down to a minimum.

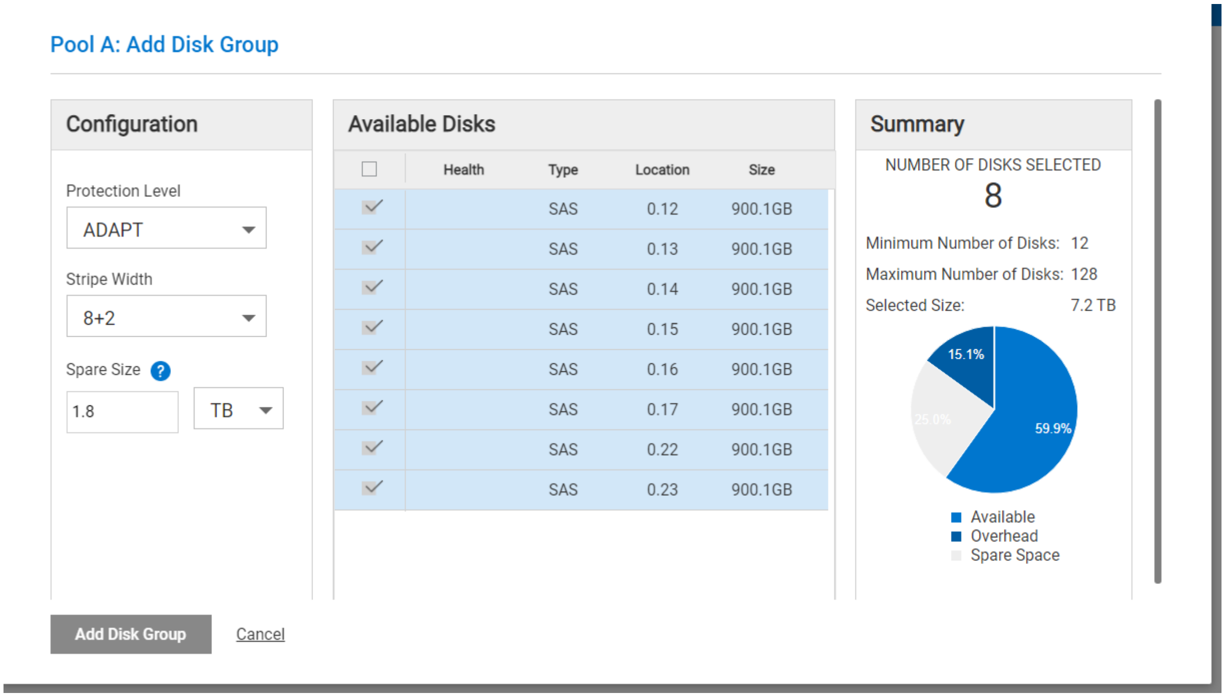
Traditionally, spare drives are kept in standby for when a drive group member fails. However spare drives are passive and sit in the system until the event happens. They are therefore a cost to ownership of the enclosure, but don't contribute to performance. ADAPT resolves this problem by reserving spare capacity (rather than spare drives) within the array. This means the drive group has reserved capacity and uses this like a spare drive if one or more drive members fails. ADAPT's default spare capacity is two times the largest capacity drive in the drive group.

ADAPT has also been developed in consideration of larger enclosures like the ME5 that starts in the 5U chassis with 84 drives behind the RAID controller. This array type removes the restriction of more than 16 members of the array; however, it still operates with stripes comprised of 8 data chunks and 2 parity chunks preventing performance degradation from wide stripes requiring large amounts of data before generating a full stripe write to the array. ADAPT's 8+2 configuration is also ideal for virtual pools because it provides more optimized sequential write throughput as compared to RAID6 with the same number of drives. Now with ME5, additional RAID 16+2 configuration can be supported further reducing the parity overhead.

ADAPT also allows more flexible use of the physical storage by potentially using each drives space and not the minimum space across the drive group as would be seen in traditional RAID. With the greater width of the ADAPT drive group we have given more flexibility in supporting expansion of the drive group by adding a broader set of options to expand the array.

What is ADAPT?

ADAPT is a data protection scheme that maximizes flexibility, provides built in spare capacity, allows for very fast rebuilds, large storage pools, and simplified expansion. ADAPT is an alternative to traditional RAID types (R0,1,5,6) of the ME5 controllers. All disks in the ADAPT disk group must be the same type in the same tier (Tiers on ME4 and ME5 are performance = SSD, standard = 10K/15K and archive = 7.2K) but can have different capacities. ADAPT is shown as a RAID level in the management interfaces (see Figure 1 below).



Pool A: Add Disk Group

Configuration

Protection Level: ADAPT

Stripe Width: 8+2

Spare Size: 1.8 TB

Available Disks

□	Health	Type	Location	Size
✓		SAS	0.12	900.1GB
✓		SAS	0.13	900.1GB
✓		SAS	0.14	900.1GB
✓		SAS	0.15	900.1GB
✓		SAS	0.16	900.1GB
✓		SAS	0.17	900.1GB
✓		SAS	0.22	900.1GB
✓		SAS	0.23	900.1GB

Summary

NUMBER OF DISKS SELECTED

8

Minimum Number of Disks: 12

Maximum Number of Disks: 128

Selected Size: 7.2 TB

Figure 1: Adding a Disk Group in Web Management Interface – ADAPT is a Data Protection option along with RAID-1, 5, 6 and 10

Physical Arrangements of an ADAPT array

ADAPT disk group's 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 automatically moves the data to maintain balance across the disk group. This section provides further information on how this is achieved.

A stripe is constructed of 10 members of the drive group with 8 data chunks and 2 parity chunks on ME5. Additionally, ME5 can support 18 members of the drive group with 16 data chunks and 2 parity chunks if there are greater than 18 members of the drive group. A chunk is the smallest unit per drive of contiguous space that makes the components of the stripe. This may contain data or parity information to protect the other data chunks of the stripe. This uses the same protection technique as RAID 6 with dual parity P and Q Reed Solomon encoding to protect the data. Having more data members in the stripe means an increase in the data to parity overhead. A chunk is 512KiB, therefore making the stripe width 4MiB. Stripes are aggregated in a linear fashion within a stripe zone. Each stripe zone contains 2048 contiguous RAID 6 stripes and is therefore 8GB of user data stored within the stripe zone. See Figure 3 to show how this is accomplished

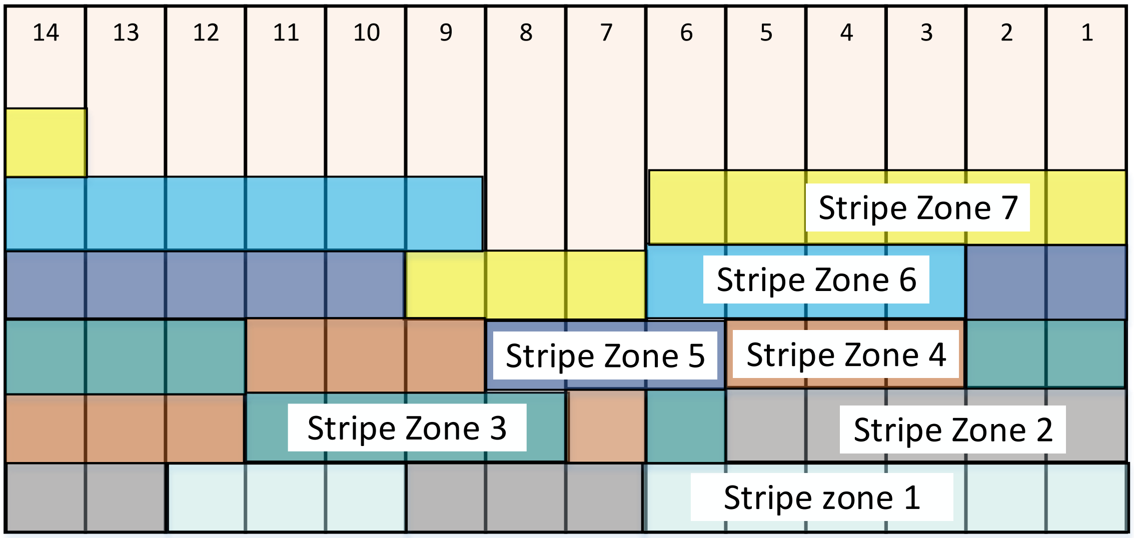


Figure 2: Stripes zones rotated across a drive group

ADAPT handles spare drives in a different way to traditional R6. In traditional R6, you allocate several spares that get used if a drive fails. The problem with this method is that they do not get used until the failure occurs. The method used in an ADAPT array is that all drives are used, and an amount of space is reserved through the disk group such that the array can be rebuilt. The process of rebuild is discussed later in this document. A default of 2 drives of spare capacity is reserved although this can be changed. This is the reason why the minimum number of drives for an ADAPT group is 12 with 2 drives of spare capacity and 80% of the remaining 10 drives used for user data and 20% used for Reed-Solomon redundancy ("parity"). Spare capacity is uniformly distributed across the LBA space of all drives. The spare space is reserved in units of spare space (1MiB). The number of spare spaces is the same as 2 drives worth capacity; however, this is distributed evenly across the drive group members and down the LBA of each member of the drive group. The following diagram shows how the spare space is evenly distributed across the space of the drive group.

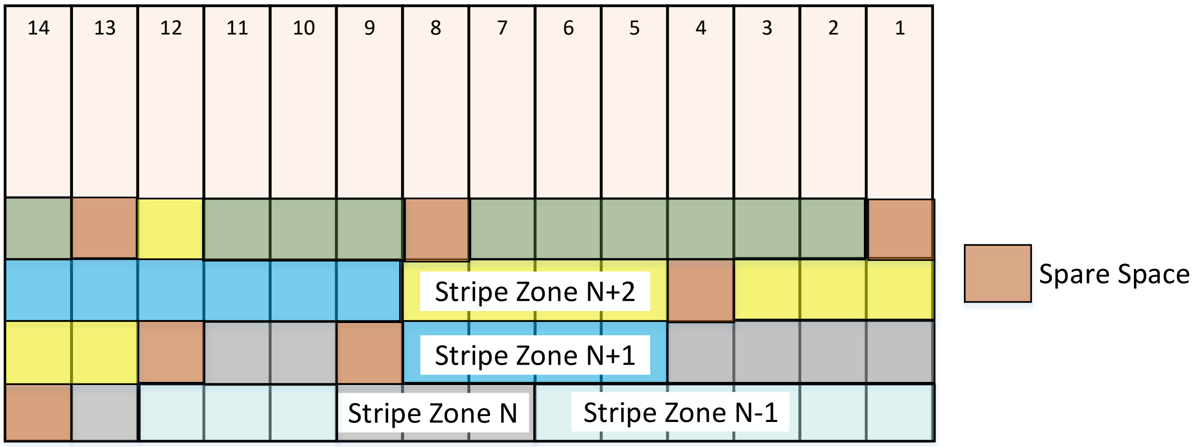


Figure 3: Allocation of spare capacity within the drive group

To handle the case of mixed drive capacities, stripes zones are allocated through the physical LBA range of the drives. When one or more of the drives capacities are exceeded, they no longer participate in zone allocation. In this manner all the physical space of drives of mixed capacity within the array can be used. Figure 4 indicates how this is accomplished.

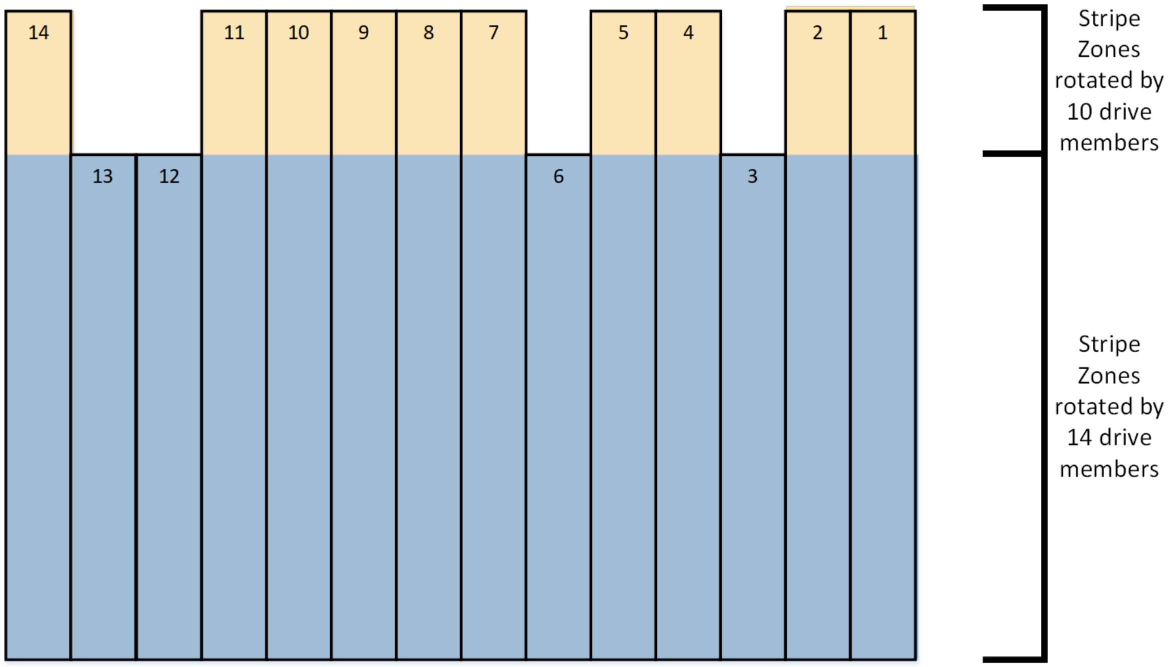

14
11
10
9
8
7
5
4
2
1
13
12
6
3
Stripe Zones rotated by 10 drive members
Stripe Zones rotated by 14 drive members

Figure 4: Stripe zones allocation when drive capacities differ

ADAPT disk group can be expanded to either replenish current target spare capacity or to increase usable capacity. For example, a 14 drive ADAPT array can have an additional 6 drive members added and will yield a 20 drive ADAPT array. If this is actioned, then the controller will conduct a balancing operation. In this operation, stripes zones that are currently rotated through the initial members of the array will be now rotated through all the members of the new array. This process will take some time and is conducted in the background. The following diagram shows how several members can be added to the disk group and how it is expanded, and the stipes is re striped over the controller.

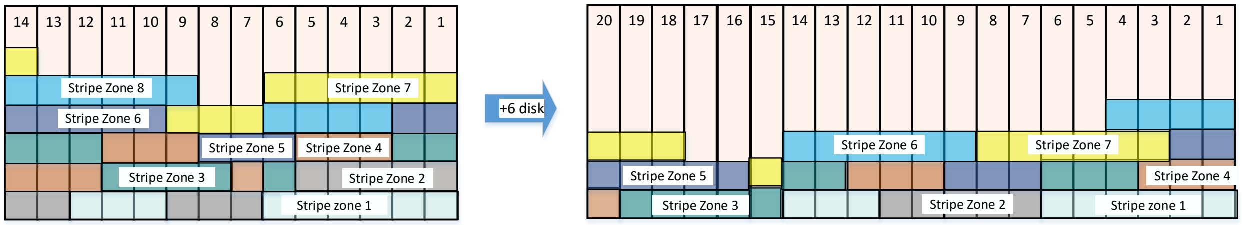


Figure 5: Restriping on disk group expansion

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 transferred to all disks in the disk group, allowing for quick rebuilds and minimal disruption to I/O.

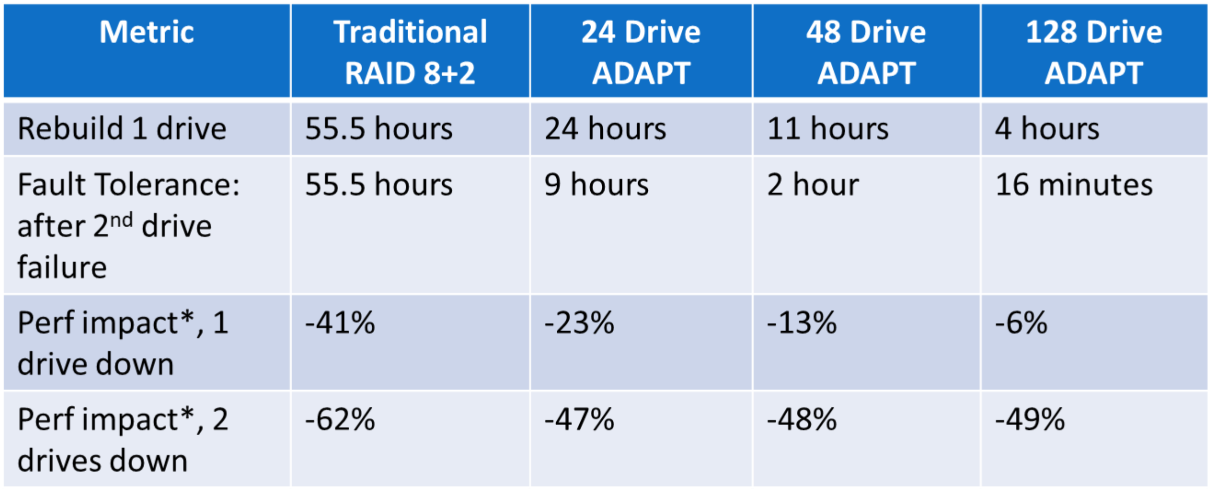
A reserve of spare drive capacity is taken from the available space that is constructed. This allows rebuilds to be conducted within the array. The advantages of this approach are considerable as it allows faster rebuilds because all members of the disk group array are used in the rebuild process. The next section goes on to look at this in more detail.

What makes ADAPT important?

Problem: With increased capacities come additional challenges on how to manage for failures in a High Availability RAID system. Although drives are very reliable, RAID systems are engineered to ensure protecting against drive failures. These larger drive sizes by nature require longer times to rebuild in the event of a failure. The PowerVault ME5 series ADAPT takes advantage of the larger drives sizes to reduce these rebuild times and increase overall system reliability and performance.

RAID systems are engineered to ensure protecting against drive failures. These larger drive sizes by nature require longer times to rebuild in the event of a failure. With the PowerVault ME5 Series, it can take advantage of the larger drives sizes, while offering additional system software capability via ADAPT to reduce these rebuild times and increase overall system reliability and performance.

The table below shows performance impacts and rebuild times for a RAID6 (8+2) disk group vs an ADAPT disk group of 24, 48, and 128 drives. Note that these data points are based on the assumptions of 10TB drives and rebuild rates of 50MB/sec/drive.



Metric	Traditional RAID 8+2	24 Drive ADAPT	48 Drive ADAPT	128 Drive ADAPT
Rebuild 1 drive	55.5 hours	24 hours	11 hours	4 hours
Fault Tolerance: after 2nd drive failure	55.5 hours	9 hours	2 hour	16 minutes
Perf impact*, 1 drive down	-41%	-23%	-13%	-6%
Perf impact*, 2 drives down	-62%	-47%	-48%	-49%

Figure 6: performance impacts and rebuild times for a RAID6 (8+2) disk group vs ADAPT disk groups

When more than 1 drive member of the array has failed, some stripes will have 1 to 2 lost chunks, depending on how they fall within the layout of the array. These stripes with 2 lost members are more at risk than ones with one lost member as any further loss of data within that stripe can no longer be corrected or protected. Therefore, priority is given to stripes that have 2 lost chunks in the rebuild process. This means that there are 2 critical measurements of rebuild, one to get back redundancy, the other to complete the rebuild and get back to optimal state. The following diagram captures the different priority of rebuilds due to the loss of 2 members of the disk array.

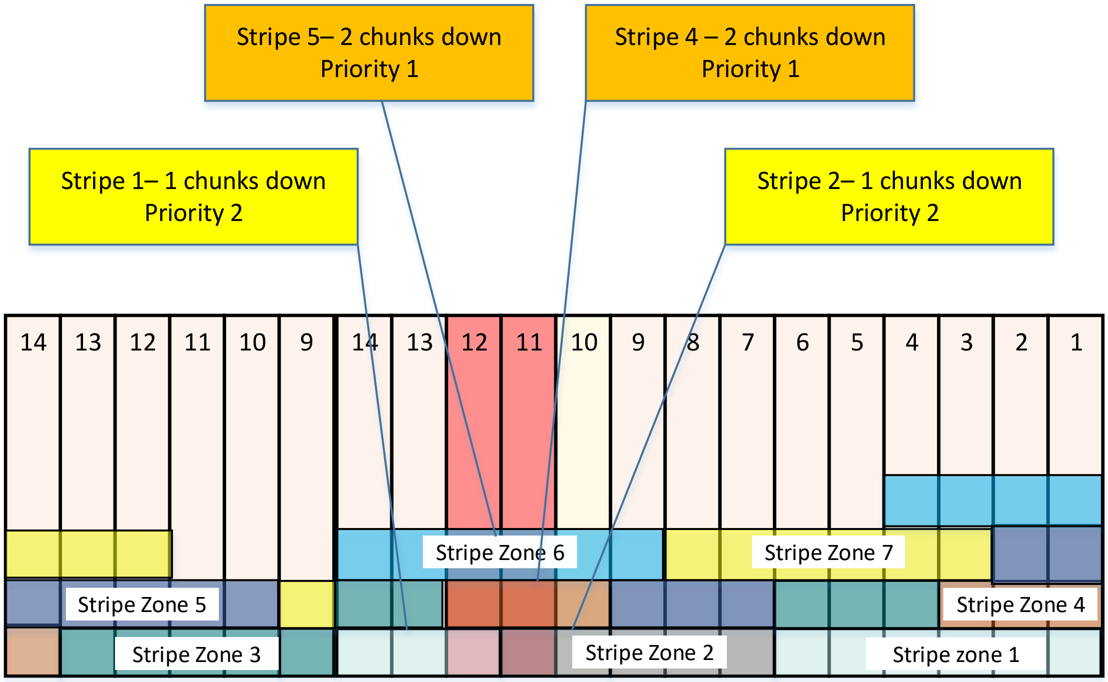


Figure 2: Stripe rebuild priority

ADAPT only needs to do this process for the stripes that are impacted by the loss of the drives. This leads to less data requiring to be rebuilt. Further as the array gets wider the contribution to that amount of data reduces per member and thus as the arrays get wider the improvement in rebuild time increases.

The other key metric is the impact to performance when the drive is lost. In traditional RAID, the parity information will be needed to re construct the data that will have been lost due to the drive loss. In ADAPT with wider array widths, the impact of any 1 drive loss is lessened. Further due to the reserved capacity, not all stripes are impacted. Thus, an ADAPT array when compared to several traditional arrays will have less impact in performance for the cases of 1 or 2 drives down.

ADAPT use case: practical examples and expectations

ADAPT is an additional option that is provided at the time of generating the disk group. An ADAPT disk group can be created in Linear or Virtual modes. In Linear the space is allocated across all members of the array. In the virtual case, it is allocated as needed on a page-by-page basis as new space within the Volume is created. The concept of a page is a fundamental facilitation for several the ME5 Series advanced features. More can be read about this in the white paper “Introduction to PowerVault Auto-Tiering and SSD Read Cache Technology”. When virtual mode is selected, other features are enabled by this underlying concept of the page such as:

- Tiering
- Snapshot
- Remote Replication

The minimum unit of allocation of any virtual mode volume is a page (4MiB). In the case of ADAPT, this is the same as a single stripe.

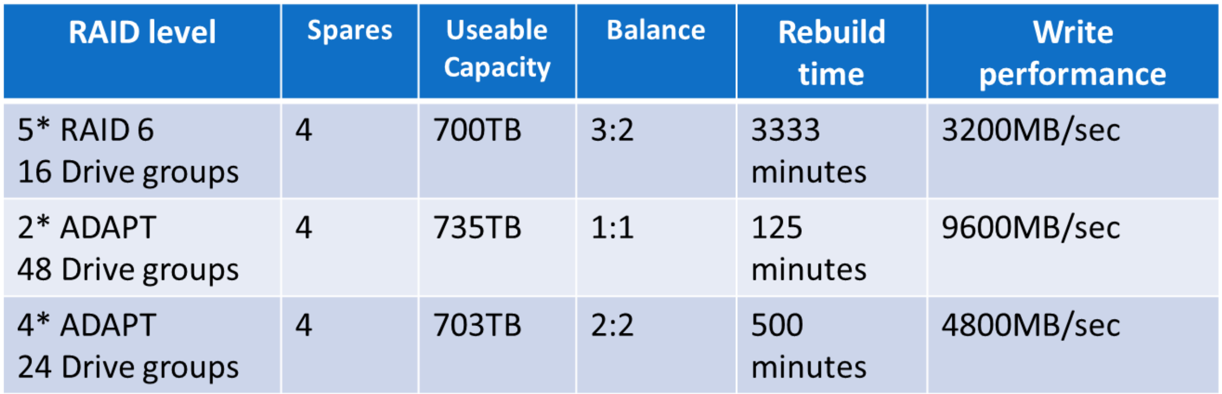
ADAPT is therefore a companion technology to help create modern flexible use of storage. An example of this would be in tiering the disk groups. On creation of one disk group when ADAPT is used for the bulk storage the option to associate this with the standard or archive tier should be made. The higher performance SSD devices should then be selected into their own drive group and have the option of performance tier selected. See companion white paper on Tiering solution.

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 * 48 disk groups or 4 * 24 disk groups would be considered optimal

*Note: 1*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*

The following table summarizes some of the key attributes to arrays built on the 5U enclosure



RAID level	Spares	Useable Capacity	Balance	Rebuild time	Write performance
5* RAID 6 16 Drive groups	4	700TB	3:2	3333 minutes	3200MB/sec
2* ADAPT 48 Drive groups	4	735TB	1:1	125 minutes	9600MB/sec
4* ADAPT 24 Drive groups	4	703TB	2:2	500 minutes	4800MB/sec

Figure 8: Key attributes to arrays built on the 5U enclosure

As can be seen here, the 4*ADAPT 24D has the same useable space as the R6 solution. However, it has twice the spare capacity. This means that there is much higher availability and reliability of the enclosure configured in this manner.

Because the ME5 controller has ownership at the drive group level, it is advised to have a balanced number of drive groups per controller to balance workload at the controller. Having to balance between five drive groups may lead to different performance at each drive group level as each controller will limit performance differently on its respective number of drive groups. An imbalance is not a problem, but it does make it harder to balance the load and therefore achieve the best performance per controller.

Wide drive groups lead to faster bandwidth performance per drive group. However, with the ME5 controller and current drive technologies, the write limit of the controller pair is approached at a single disk group level if we consider sequential read or write IO. Therefore, if more disk groups are added or wider drive groups are used incremental performance may not be observed when considering sequential IO.

For Random IO wider drive groups have more performance on reads as the IO's will land on more spindles. With performance limited by the seek time of the spindle, therefore more drives mean increased performance. The same is true for writes; however, like R6 there is a significant Read modify Write overhead.

Summary: Features and Benefits

Feature	Benefit
Significantly reduce rebuild time via parallel architecture	Data protection especially with large devices
Self-healing capability	Automatically allocates spare capacity to recover device failures. No spare drives
Mixed drive capacities	Maximize usable capacity, reduce $/TB
Eliminate enclosure geometry restrictions	Simplified user configurations
No loss of data with any two devices failed	Fault tolerance
Support pools from 12 to 128 devices and online pool expansion	Excellent scalability (and performance scaling)
Good sequential I/O performance	HDD streaming applications

Conclusion

As Dell Technologies continues to bring higher density disk drives to market, it is imperative to develop systems which can leverage these larger drives and provide increased reliability and performance. ADAPT is the latest system software enhancement from Dell Technologies and is available by default with every PowerVault ME5 array in addition to more traditional RAID data protection schemes.

Glossary and References

Term	Definition
Chassis	The sheet metal housing of the enclosure and all that is contained within it.
Chunk	The amount of contiguous data that is written to a disk group member before moving to the next member of the disk group. For ADAPT this is 512KiB
Disk Group	A group of disks that is configured to use a specific RAID level and provides storage capacity for a pool. Minimum for ADAPT = 12 Maximum for ADAPT = 129 Maximum ADAPT groups per controller is 4
Linear (pool/mode)	The storage class designated for logical components such as volumes that do not use paged storage to virtualize data storage. The linear method stores user data in sequential, fully allocated physical blocks of fixes (static) mapping between the logical data presented to the hosts and the physical storage where it is stored.
Drives per ADAPT drive group	12 Minimum , 128 Maximum
Page	A range of contiguous LBA's of a virtual disk group 4MiB in size.
Paged storage	A method of mapping logical host requests to physical storage that maps the request to virtualized "pages" of storage that are in turn mapped to physical storage. This provides more flexibility for expanding capacity and automatically moving data than the traditional linear method in which request are directly mapped to storage devices. Paged storage is also called virtual storage
Pool	A container for volumes that contains 1 or more disk pools For ME5 maximum virtual pool size if 1PB
Storage system	A controller enclosure with 1 or more connected expansion enclosure
Stripe	A number of chunks spread across different disks. A complete stripe containing data and parity information to protect the data within this space For ADAPT this is 8 Data chunks plus 2 parity chunks on ME4 and ME5 and 16 data chunks plus 2 parity chunks on drive groups greater than 18 drives.
Tier	A homogeneous group of disks, typically of the same capacity and performance level that comprise one or more virtual disk groups in the same disk pool. Tiers differ in their performance, capacity and cost characteristics, which forms the basis for the choices that are made with respect to which data is placed on which tier. The predefined tiers are • Performance which uses SSDs • Standard which uses enterprise class spinning disks (10/15K RPM) • Archive which uses midline spinning disks For ADAPT disk groups they can be built of drives within the same tier
Virtual (Pool/mode)	The storage class designation for logical components such as volumes that use paged storage technology to virtualize disk storage

References: "PowerVault ME5 Series Tiering and Cache Software White Paper" located on www.delltechnologies.com/powervault