Dynamic Disk Pools Technical Report

Dell PowerVault MD3 Series of Storage Arrays

Dell Storage Engineering
September 2017

A Dell Technical White Paper

Table of contents

Introduction

The Dynamic Disk Pool (DDP) feature dynamically distributes data, spare capacity, and protection information across a pool of disk drives. These pools may range in size from a minimum of 11 drives to potentially as large as of all of the drives in the PowerVault MD3 storage system including the expansion units. In addition to creation of a single DDP, storage administrators may opt to create traditional disks in conjunction with DDPs or even multiple DDPs, providing an unprecedented level of flexibility. This technical report provides a technical overview of the DDP feature as well as providing best practice guidance for usage.

Technical Overview

In essence, DDP functions as effectively another RAID level offering in addition to the previously available RAID 0, 1, 10, 5, and 6 traditional RAID Disk Groups. DDP greatly simplifies storage administration because there is no need to manage Idle spares or RAID groups.

The following discusses the intrinsic characteristics and architecture of the DDP feature.

DDPs are composed of several lower level elements.

1. **D-Piece** : The first of these is known as a D-Piece. A D-Piece consists of a contiguous 512MB section from a physical disk containing 4,096 128KB segments. Within a pool, 10 D-Pieces are selected using an intelligent optimization algorithm from selected drives within the pool.
2. **D-Stripe** : The ten associated D-Pieces are then considered a D-Stripe which is 5GB in size. Within the D-Stripe itself, the contents are similar to a RAID 6, 8+2 scenario whereby 8 of the underlying segments potentially contain user data, 1 segment contains parity (P) information calculated from the user data segments, and the final segment containing the Q value as defined by RAID 6.

From the virtual disk or LUN perspective, they are essentially created from an aggregation of multiple 4GB D-Stripes as required to satisfy the defined virtual disk size up to the maximum allowable within a DDP.

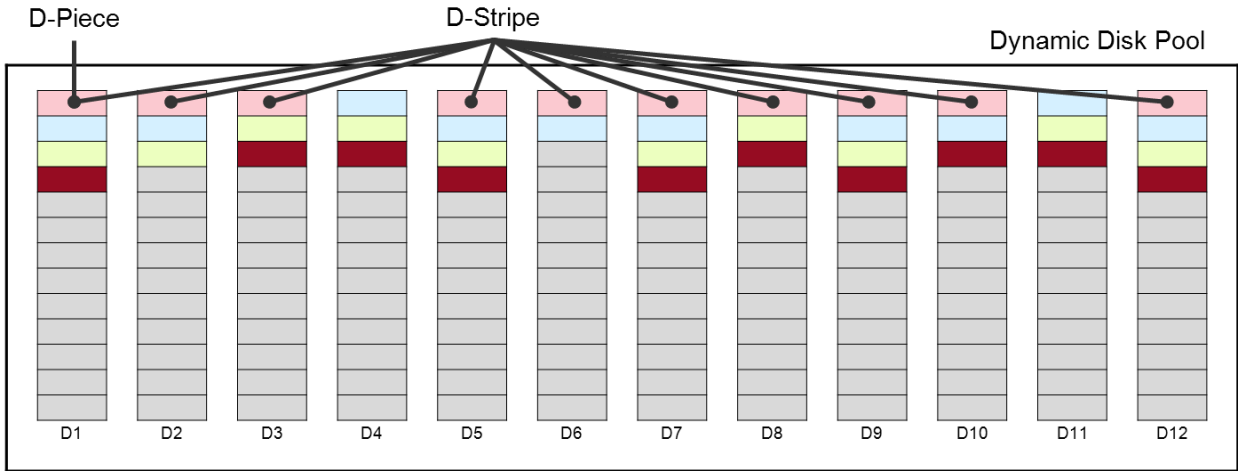


Figure 1 D-Piece and D-Stripe

In **Figure 1** above, a simplified diagram is shown of a single DDP containing 12 disk drives. As shown above, there are ten D-Pieces organized into a D-Stripe located pseudo-randomly across ten of the disks within the pool and this continues similarly for each 5GB D-Stripe. It is important to note that while distribution of D-Pieces and D-Stripes is approximately equal across all disks, as seen in the above example it is possible that this will not completely be the case.

Once a storage administrator has completed the action of defining a DDP, which largely consists of simply defining the number of desired drives in the pool, the D-Piece and D-Stripe structures are created, similar to how traditional RAID stripes are created during Virtual disk creation. After the DDP has been defined, a virtual disk may be created within the pool. This virtual disk will consist of some number of D-Stripes located across all the drives within the pool up to the defined value for the virtual disk capacity, such that **Number of D-Stripes = Usable capacity / 4GB.**

As an example, a 500GB virtual disk would consist of 125 D-Stripes. Allocation of D-Stripes for a given virtual disk is performed starting at the lowest available range of LBAs for a given D-Piece on a given disk drive.

Multiple virtual disks may be defined within a DDP and as noted previously there may be multiple pools created within the supported storage system. Alternatively, the storage administrator could also choose to create traditional RAID Disk Groups in conjunction with DDP or any combination thereof.

As an example, with a 60-drive PowerVault MD3X60x array with all drives being of the same capacity, the following combinations would all be supported:

- 1x 60 drive DDP
- 2x 30 drive DDPs
- 1x RAID 5 (30 Disk Group) and 1x DDP (30 drive pool)

Similarly to traditional RAID Disk Groups, DDPs can be expanded by the addition of disk drives to the pool through the Dynamic Capacity Expansion (DCE) process and up to 12 disks may be added to a defined disk pool concurrently. When a DCE operation is initiated, a small percentage of the existing D-Pieces are effectively migrated to the new disks.

Data Availability

Another major benefit of a DDP is that rather than using dedicated “stranded” hot spares the pool itself contains integrated preservation capacity to provide rebuild locations for potential drive failures. This simplifies management as it is no longer required to plan or manage individual hot spares as well as greatly improving both the time of rebuilds if they need to occur and enhancing the performance of the virtual disks themselves while under a rebuild as opposed to traditional hot spares.

When a drive in a DDP fails, the D-Pieces from the failed drive are reconstructed using the same mechanisms as would be used by RAID 6 normally to potentially all other drives in the pool. During this process, it is ensured that no single drive contains two D-Pieces from the same D-Stripe. The individual D-Pieces are reconstructed at the lowest available LBA range on the selected disk drive.

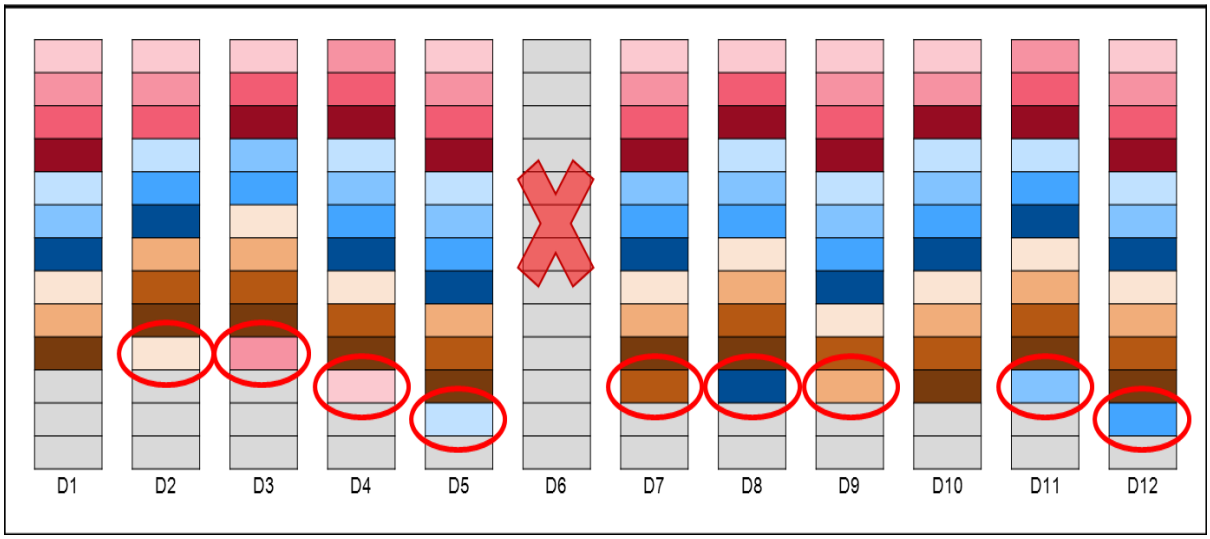


Figure 2 DDP Reconstruction

In **Figure 2,** disk drive 6 is shown to have failed. Subsequently, the D-Pieces that were previously resident on that disk are recreated across several other drives in the pool simultaneously. Since there are multiple disks participating in the effort, the overall performance impact of this situation is lessened as well as the length of time needed to complete the operation being reduced dramatically.

In the event of multiple disk failures within a DDP, priority reconstruction is given to any D-Stripes which are missing two D-Pieces in order to minimize any data availability risk. After those critically affected D-Stripes are reconstructed, the remainder of the necessary data will continue to be reconstructed.

From a controller resource allocation perspective, there are two user-modifiable reconstruction priorities within DDP:

- Degraded reconstruction priority is assigned for instances where only a single D-Piece needs to be rebuilt for affected D-Stripes. The default is 'high'.
- Critical reconstruction priority is assigned for instances where a D-Stripe has two missing D-Pieces which need to rebuilt. The default is 'highest'.

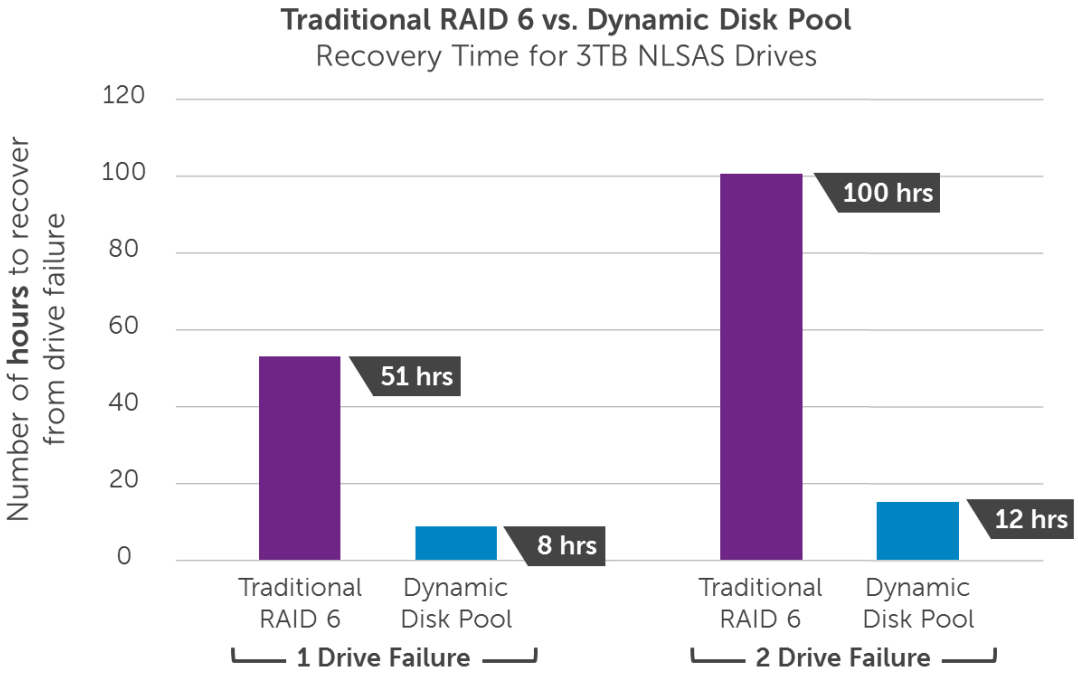
It is recommended to use Low or Medium priority settings for NL-SAS drives, this will increase the drive reconstruction time but will also lessen the impact of I/O performance during rebuild.

For very large disk pools with two simultaneous disk failures, only a relatively small number of D-Stripes are likely to encounter the critical situation where two D-Pieces need to be reconstructed. As discussed previously, these critical D-Pieces will be identified and reconstructed first at the highest priority thus returning the DDP to a 'degraded' state from 'critical' state very quickly such that a further drive failure could be tolerated.

As an example, assume that a DDP consisting of 192 disk drives has been created and has a dual disk failure. In that event, it is likely that the critical D-Pieces would be reconstructed in less than a minute and, after that minute, a further disk failure could be tolerated. From a mathematical perspective given the same 192 drive pool, only 5.2% of D-Stripes would have a D-Piece on one drive in the pool and only 0.25% of D-Stripes would have two D-Pieces on those two particular drives, meaning that only 5.2GB of data would have to be reconstructed to exit the critical stage. A very large disk pool can continue to maintain multiple sequential failures without data loss up until there is no additional preservation capacity to continue the rebuilds.

After the reconstruction, the failed drive or drives may be subsequently replaced, although this is not specifically required. Fundamentally, this replacement of failed disk drives is treated in much the same way as a Dynamic Capacity Expansion (DCE) of the DDP. Failed drives may also be replaced prior to the DDP exiting from a critical or degraded state.

Aside from the reduced time to move from a critical state to that of simply degraded, the general rebuild process for a DDP can be up to 10x faster than a traditional RAID Disk Group.



In the above example, the rebuild time for several different configurations was determined empirically. The tests were conducted with 3TB 7200RPM NL-SAS drives on a Dell PowerVault MD3860f storage system. The RAID 6, 10+2 baseline with a single drive failure resulted in almost a 51 hour rebuild time, while a similar size DDP completed in less than ten hours.

Management

Management of DDPs is substantially simplified versus that of traditional RAID Disk Groups. Due to the nature of a DDP, there is fundamentally only one decision required from a storage administrator- how many disks are preferred to be in the disk pool.

As shown in the above, creation of a new disk pool is a simple matter of selecting the desired number of disk drives. Preservation capacity and many other attributes of the disk pool are predefined. The amount of preservation capacity to be utilized by a DDP is fixed depending on the number of drives that were created within the pool.

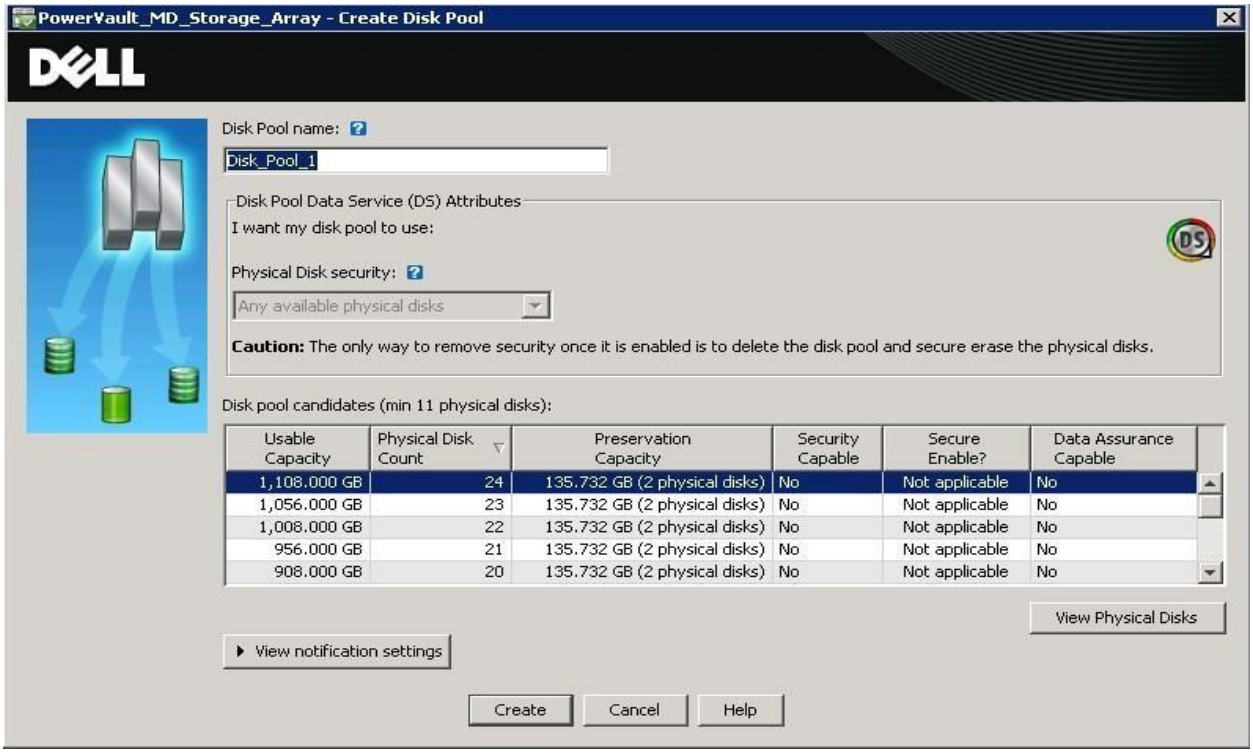


Figure 3 Create Disk Pool

After a DDP has been created, virtual disks may be created with the pool in much the same manner as was previously used for traditional RAID Disk Groups with the addition that in a pool thinly-provisioned virtual disks are also available.

Figure 4 Create Virtual Disk (resize)

Given the nature of D-Stripe distribution within a DDP, performance is largely going to depend on the frequency of I/O requests to specific disks within the pool. For environments where potential disk contention is not a limiting factor for performance but instead the effective bottleneck is at the controller level for items such as CPU resources and available host-side bandwidth, then very little difference would be expected between a DDP and a similarly configured traditional RAID Disk Group(s).

Restrictions and Limitations

DDPs do have some restrictions that are not in place for traditional RAID Disk Groups.

- The segment size for all virtual disks in a DDP is 128KB and this may not be modified. Therefore, the Dynamic Segment Sizing feature is not available.
- DDPs do not support the Dynamic RAID Migration (DRM) feature and the only supported protection mechanism is that of RAID 6.
- The pre-read redundancy check feature is not available for use with virtual disks in a DDP.
- Tray and drawer loss protection is not considered as part of the disk pool creation.
- DDPs cannot be migrated data intact to another PowerVault MD3 Storage system.
- Maximum virtual disk size within a DDP is 64TB.
- Maximum combined DDP capacity, except for MD32x0i or MD36x0i, is 1024TB.
- Maximum combined DDP capacity for MD32x0i or MD36x0i is 256TB.

Performance

Dynamic Given the nature of D-Stripe distribution within a DDP, performance is largely going to depend on the frequency of I/O requests to specific disks within the pool. For environments where potential disk contention is not a limiting factor for performance but instead the effective bottleneck is at the controller level for items such as CPU resources and available host-side bandwidth, then very little difference would be expected between a DDP and a similarly configured traditional RAID Disk Group(s).

Sequential Performance

Within a DDP, D-Stripes are pseudo-randomly distributed across groupings of ten disk drives. This can have the effect of randomizing access patterns to the underlying physical disks even in cases where the host read requests are sequential in nature and is particularly observable in disk pools where the spindle count is not sufficient to counter the potential for randomization. For a minimally-sized pool with a single virtual disk, the sequential read performance can be fairly similar to a traditional RAID 6 Group with a single virtual disk as the underlying architecture is similar. As with RAID Groups, if additional virtual disks exist within a DDP and concurrent I/O access is expected, sequential performance may suffer as a result of head movement with the physical disk.

Due in part to caching mechanisms, sequential writes on a DDP are generally more similar to that of traditional RAID Disk Groups.

Random Performance

Given the nature of DDPs, random performance, both read and write, on a disk pool tends to be relatively similar in nature to that of traditional RAID Groups of comparable size.

In general, if a PowerVault storage system is expected to drive maximum performance for a sustained periods of time, then generically traditional RAID Disk Groups will outperform DDPs for both sequential and random workflows. This is primarily due to the allocation of resources within the disk pool itself. For a traditional RAID group, the configuration may be such that a given I/O stream has unique access to the disk resources within that group. In contrast, all I/O streams within a DDP share access to the disks within the pool. This can create the potential for resource contention with the underlying physical media depending on the workload.

Summary

As disk capacities continue to increase, the rebuild times for disk failures within RAID Groups will likewise increase leaving storage systems potentially at risk for additional drive failures and prolonging the performance impact to applications during the rebuild process. DDPs offer an exciting new approach to traditional RAID sets by substantially improving rebuild times, limiting critical exposure during dual drive failures, reducing the performance penalty suffered during a rebuild, as well as significantly simplifying storage administration. In addition, given that virtual disks within DDPs can potentially span very large numbers of disk drives versus traditional RAID 5 or RAID 6 Disk Groups, in environments with mixed or non-concurrent workflows, there can be a tremendous advantage in terms of performance as all pool resources are available to all hosts

Dell PowerVault MD3 Storage arrays support both DDPs and traditional RAID Disk Groups, the following table could be used as reference for comparison between the two technologies.

Table 1 DDP vs. Traditional RAID group Summary

Component	Dynamic Disk Pool	Traditional RAID Group
Ease of configuration and management	✓	
Performance during drive failure	✓	
Maximum performance for large, sequential workloads	✓	✓
Maximum performance for small, random workloads	✓	✓
Shortest drive reconstruction time	✓	
Least time required to add drives (Dynamic Capacity Expansion)	✓	
Greater protection of data from multiple drive failures occurring over time	✓	
Simplified administration of spare capacity	✓	
Flexible and efficient capacity utilization	✓	
Ability to fine-tune performance for specific workloads		✓

The following flow chart can be used as a guide to help with deciding whether a DDP or traditional RAID Disk Group may be the proper choice for a given environment.

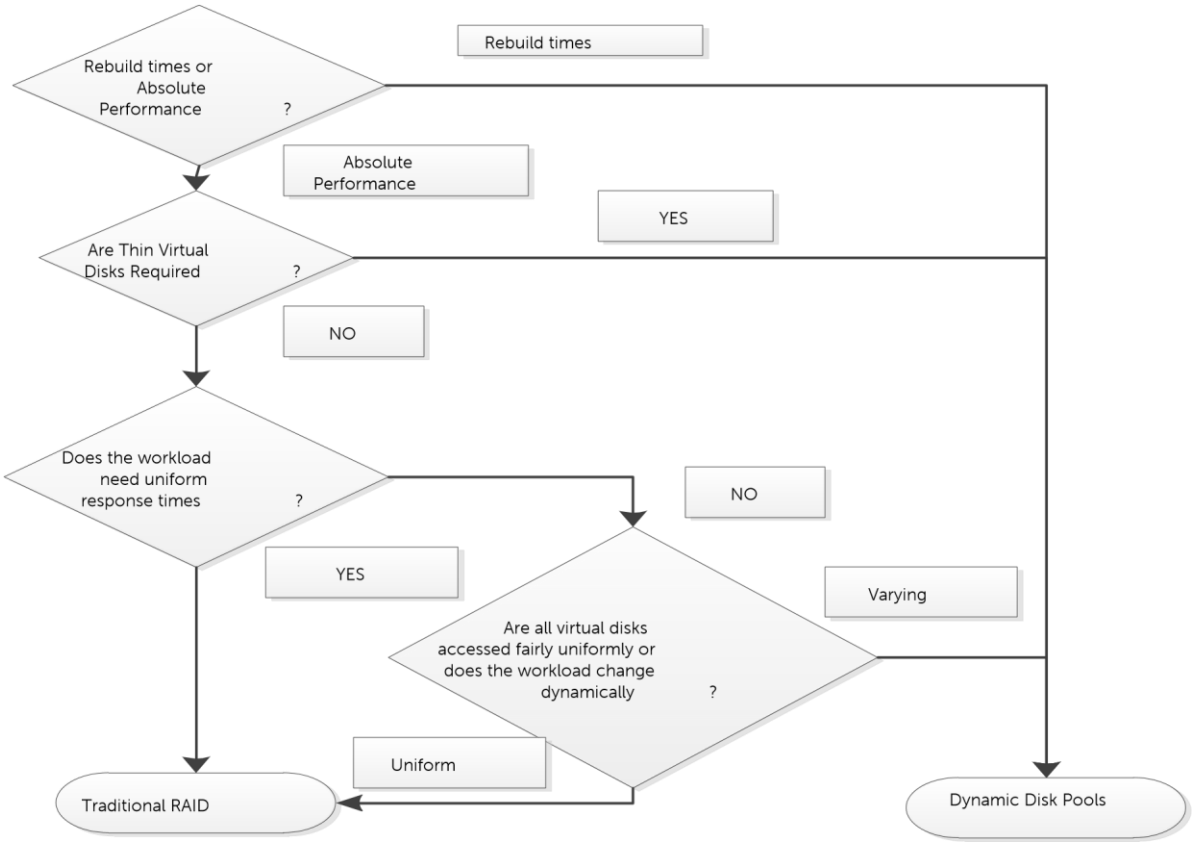


Figure 5 Traditional RAID vs. DDP Decision Tree