DELL EMC

Technical White Paper

Dell EMC PowerMax: Data Reduction

Inline compression and deduplication

Abstract

Dell EMC™ PowerMax Data Reduction is a storage efficiency feature that combines inline compression and inline deduplication. Using both storage efficiency features together enhances capacity savings while maintaining great performance and reliability.

September 2019

H17072.1

Revisions

Date	Description
May 2018	Initial release
September 2019	Content and template update

Acknowledgements

Author: Robert Tasker Jr.

Table of contents

Executive summary

Data reduction with the Dell EMC PowerMax system delivers a boost to system efficiency by combining inline compression with inline deduplication. Paring two techniques of capacity savings creates a system where they share resources and components allowing users to realize an average capacity savings of 3:1. Data Reduction not only compresses data it also eliminates redundant copies of compressed data and delivers great all flash performance. The contents of this technical white paper are intended to inform the reader how Data Reduction functions within the Dell EMC™ PowerMax systems.

1 Data Reduction overview

In PowerMax data storage systems data reduction combines the proven Adaptive Compression Engine (ACE) and inline deduplication to provide a highly performing space efficient platform. Data Reduction allows users to present more front-end effective capacity to lower back end usable capacity. Compression and Dedupe are two different functions that work together. Compression reduces the size of data sets and dedupe identifies identical data sets and stores a single instance. Performing both functions at the same time allows the system to be capacity efficient and deliver an average savings of 3:1.

1.1 Adaptive Compression Engine overview

The Adaptive Compression Engine (ACE) is the combination of multiple components to deliver the performance expected from an all flash storage system and maintain data storage efficiency. Incoming data is compressed inline using hardware-based compression with software compression in place and used as needed. Intelligent algorithms learn from the incoming workload to dynamically create a customized backend, catering to the incoming workload. The Adaptive Compression Engine changes the backend compression pool layout as needed to ensure the system operates at optimal levels for both performance and space efficiency. Using internal statistics, algorithms identify the busiest data in the system allowing it to skip the compression process. The result minimizes de-compression overhead on data that is accessed the most. Working together, these functions allow ACE to deliver great performance and efficiently manage back end capacity usage. ACE is covered in more detail in section 4.

1.2 Deduplication overview

Deduplication (dedupe) is a capacity savings method that identifies identical copies of data and stores a single instance. There are a few facets of deduplication that are needed for it to function properly and provide efficient capacity savings.

- Hash ID
- Hash Table
- Dedupe Management Object (DMO)

The Hash ID is a unique identifier for incoming data that is used to determine if a dedupe relationship is needed. The system uses a SHA-256 algorithm to generate the Hash ID. The Hash ID is stored in a Hash Table and used for future comparison. Hash Tables are an allocation of system memory distributed between the system directors. The Dedupe Management Object (DMO) is where the Hash ID is stored when a dedupe relationship exists. The DMO also manages the required pointers between the front-end devices and the data stored on disk.

2 Terminology

Data Reduction: The use of compression and deduplication together to reduce capacity usage and the cost of physical storage. (In systems prior to PowerMax Data Reduction is compression only)

Data Reduction Reservation: A system resource reservation relative to system sizing used to determine the maximum savings that can be achieved when Data Reduction is enabled.

Storage Group Compression Ratio: The compression ratio displayed for allocations related to a specific storage group. This value may be greater than or less than the system (SRP) Data Reduction ratio displayed in the Efficiency Report.

Compressibility: The maximum compression ratio that may be achieved for either a storage group or a device. This value may be presented as a higher value than current savings due to the design of ACE, more specifically the activity-based compression (ABC) function.

Data Reduction Ready: The state of the system when the default Storage Resource Pool (SRP) can store compressed data. For a system to be able to compress data it must have one compression I/O module per director, have compression enabled and have a system Data Reduction Reservation ratio set.

Data Reduction Capable: A system installed with at least the Q1 2018 PowerMaxOS where the Data Reduction Reservation applied to the system IMPL is 1.0:1.

Compression Pool: The collection of data devices configured within the drives where the track size is the same. For example, the 64K pool is made up of data devices where all of the device’s tracks are 64K in size.

Terabytes Usable (TBu): The backend usable storage capacity in the absence of compression referring to the amount of physical storage in the system.

Example: 50 TBu is 50 Terabytes of usable physical storage.

Terabytes Effective (TBe): The frontend effective storage capacity in the presence of data reduction. This represents the maximum amount of host or application data that can be written to the array.

Example: 50 TBu of physical storage with a data reduction reservation of 3:1 translates to 150 TBe.

3 Configuration details

PowerMaxOS is supported on PowerMax and VMAX All Flash data storage arrays. There are a few different scenarios for the two storage arrays. See the table 1 below for details.

Table 1 Configuration details by storage platform

	PowerMax	**VMAX All Flash**
Adaptive Compression Engine (ACE)	Yes	Yes
Extended Data Compression (EDC)	Yes	No
Inline Deduplication	Yes	No
Data Reduction I/O Module	Yes	No
Compression I/O Module	No	Yes
DEFLATE Compression Algorithm	Yes	No
LZS Compression Algorithm	No	Yes
FBA Storage Resource Pool (SRP)	Yes	Yes
CKD Storage Resource Pool (SRP)	No	No

4 Adaptive Compression Engine

The Adaptive Compression Engine (ACE) is the combination of multiple core components that work together to achieve maximum system efficiency and deliver optimized performance. These core components are:

4.1 Hardware acceleration

Each system is equipped with data reduction hardware that handles the actual compressing and de-compressing of data. The arrays are configured with one module per director which equates to 2 for each engine. The use of the modules reduces data reduction processing overhead. As a secondary function, there is built in software compression that is automatically applied in the event of a fault or failure with one or more of the data reduction modules.

4.2 Optimized Data Placement

Optimized Data Placement is a function within the system that is responsible for dynamically changing the compression pools as needed. Optimized Data Placement dynamically alters the backend by creating a variety of compression pools that cater to the incoming data. Compression pools represent usable capacity on multiple solid-state drives (SSD). Most data can be compressed, however not all data compresses to the same degree. To maximize compression efficiency, the system needs to accommodate multiple sizes of compressed data. To support this variety of compression sizes, multiple compression pools are used to create an optimal backend. The result is a layout of compression pools that cater the data sent to the system.

Compression pools are identified by the label which represents the track size for the data devices within the pool. For example, the 128K pool is made up of data devices where the tracks are all 128K in size. The 8K pool is made up of data devices where the tracks are all 8K in size. In comparison the capacity of the data devices between the pools is the same, however, the 8K pool has 16 times the amount of tracks. Here is a complete list of possible compression pools; 8K, 16K, 24K, 32K, 40K, 48K, 56K, 64K, 72K, 80K, 88K, 96K, 104K, 112K, and 128K. Each compression-enabled system may have a different combination of compression pools that compressed data populates.

4.3 Activity Based Compression

ABC aims to prevent constant compression and decompression of data that is accessed frequently. The ABC function allows the busiest data in the SRP to avoid being compressed regardless of the storage group compression setting. This function differentiates busy data from less busy data and accounts for up to 20% of the allocations in the SRP. Allowing the busiest allocations to skip compression is a benefit to the system as well as to end users. This ensures optimal performance and reduced overhead that can result from constantly decompressing frequently accessed data. The mechanism used to determine the busiest data does not add additional load on the system. ABC leverages statistics collected from the frontend devices to determine what data sets are the busiest and the best candidates to skip compression. It allows the system to maintain balance across the resources providing an optimal environment for both compression savings and performance.

4.4 Fine Grain Data Packing

The Adaptive Compression Engine utilizes data reduction hardware to process incoming data which is divided into four sections. Each section is compressed individually and in parallel which maximizes the efficiency of the data reduction module. The sum of the four compressed sections is the final compressed size and, determines where the data is to be stored. The compression process includes a non-zero allocate function. This function prevents the allocation of any of the 4 sections that contain all zeros. This behavior results in an efficient compression process that has minimal cost to performance.

Another benefit of dividing the extents into four sections comes when there are partial read or write operations. In this case only the sections that contain the requested data are processed. This means each section can be handled independently even though they are still part of the initial full data.

The efficiency of data compression is measured in terms of the compression ratio. This is the ratio between the original size of the data and its size after being compressed. For example, a 128K data is compressed to 64K, resulting in a compression ratio of 2:1.

4.5 Extended Data Compression

PowerMax systems include an additional function that will compress already compressed data to gain further capacity savings. The goal of Extended Data Compression (EDC) is to apply additional compression savings to already compressed data. This is accomplished by identifying data that has not been accessed for a set amount of time. The factors that make data a candidate for EDC are the following:

- The data belongs to a data reduction enabled storage group.
- The data has not been accessed for 30 days.
- The data is not already compressed by EDC.

Data that qualifies for EDC is compressed using the Def9_128_SW algorithm and moved to the appropriate compression pool. This is an automated background process within the system. Additional savings is included in the storage group level achieved compression ratio. This function of ACE is only available with the PowerMax data storage arrays.

5 Deduplication

Deduplication (dedupe) is the process of reducing redundant copies of data that consume storage capacity. The redundant copies are replaced with pointers. The pointers provide the access for the subsequent requests of that shared data by multiple sources. In PowerMax systems, dedupe is accomplished through a series of functions and components including hardware acceleration, dedupe algorithm, Hash Table and Dedupe Management Object (DMO).

5.1 Hardware acceleration

Dedupe is an inline process that uses the same data reduction hardware as ACE. All incoming data is passed through the data reduction hardware to generate a unique identifier called a hash ID. This is done before the data is stored on the physical disks and is performed simultaneously to compression.

5.2 Deduplication algorithm

PowerMax systems use the powerful SHA-256 hashing algorithm implemented in hardware to find duplicate data. The data is then stored as a single instance of data for multiple sources to share. This provides for enhanced data efficiency while maintaining a long history of data integrity.

For data reduction, PowerMax handles each 128k data track as 4 blocks of 32k. Each block can be updated independently, and the new 32k block is passed for data reduction on the way to the backed solid-state storage. The data reduction hardware does both deflate compression and deduplication hash generation in a single pass, producing compressed data with a unique Hash ID. This offload allows the CPUs to be focused on customer I/O rather than background operations to perform these tasks.

The SHA-256 algorithm generates a 32-byte code for each 32k block of data. Consider a system with 1PB of written data with 5% updated per day. In 1 million years of operation, there is a 20% likelihood of a hash collision. As each 128k track is handled as 4 blocks of 32k there would need to be a hash collision on all 4 blocks in the same 128k track to have an actual hash collision. The odds of having all 4 collide makes this only theoretical (less than a 1% chance in a trillion years of operation).

5.3 Hash table

The Hash Table uses system memory to store the unique Hash IDs that are used for comparison as part of the dedupe process. Hash IDs stored in the table are a unique representation of data in a dedupe relationship. Hash IDs generated by the data reduction hardware and the SHA-256 algorithm for new writes are compared against the IDs already populating the Hash Table. If a matching Hash ID already exists in the Hash Table, then a dedupe relationship is generated for the newly written data.

5.4 Dedupe management object

The dedupe management object (DMO) is a 64-byte object within system memory. DMOs only exist when dedupe relationships exist. These objects store and manage the pointers between front end devices and the single instance of data that consumes backend capacity in the array.

5.5 Data reduction I/O flow

All I/O in PowerMax arrays is passed through front end cache and then processed by the system. One of the primary reasons to implement a data reduction method is to reduce the amount of disk space that is consumed. The PowerMax data reduction feature is an inline process. This means dedupe and compression actions are performed after the data is received by the system, but before it is placed on disk. Using an inline process requires some changes to the I/O flow where data reduction applies. Compression aims to limit the amount of disk capacity used by reducing its actual size. Dedupe achieves reduced capacity usage by limiting identical data sets consuming disk space to a single instance.

There are a few different I/O types to consider, Read, Write, Write-update.

- Read - A request to access data that is already populating the array.
- Write - Incoming I/O that will consume disk space.
- Write-update - Incoming I/O that can change data that is allocated to disk space on the array.

Figure 1 below describes the path the I/O will follow which is determined by characteristics of the data set and/or the related storage group.

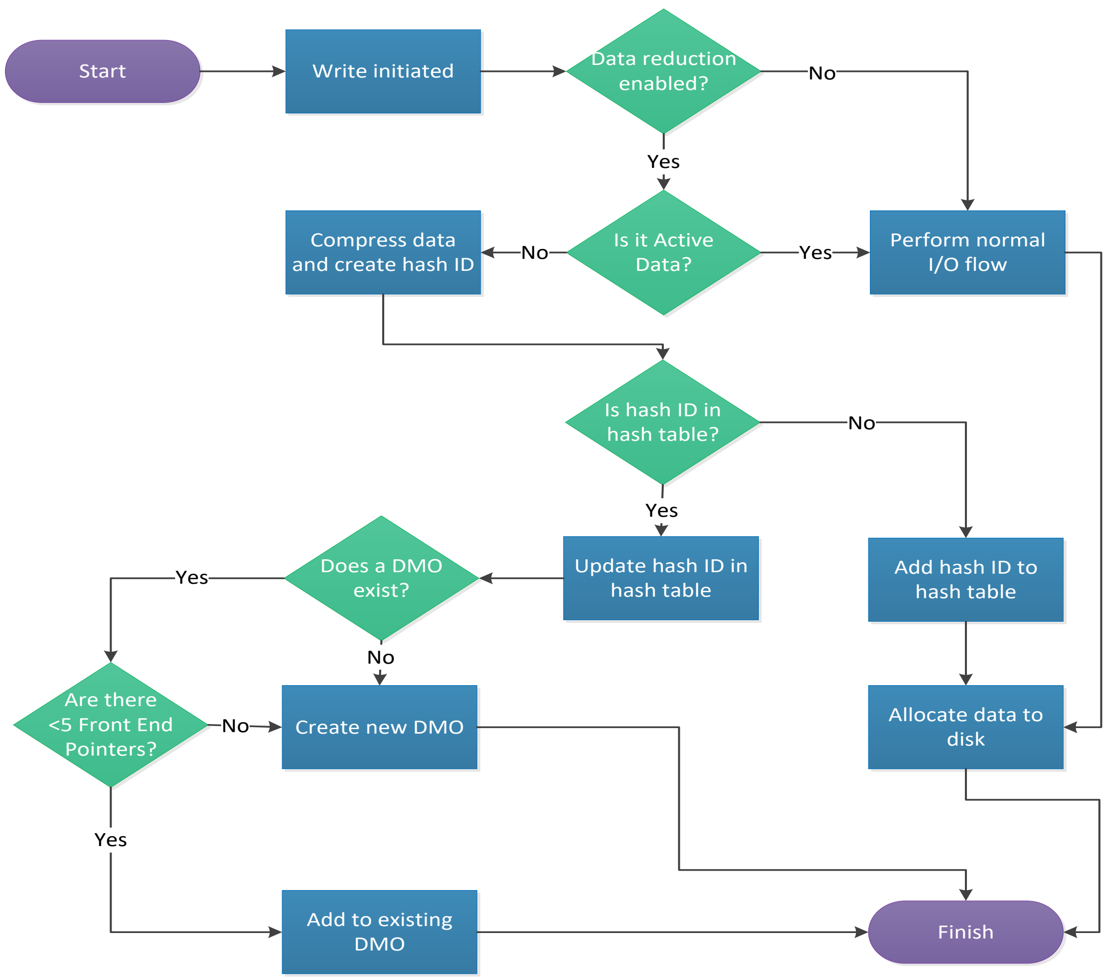

Figure 1 Data Reduction I/O flow for PowerMax enterprise storage systems

6 Capacity usage

Data reduction is an inline feature intended to offer long-term space savings. Data Reduction is machine learning which results in efficient use of the available system resources. The use of statistics collected from the incoming data determines what is active and what is idle. Therefore, the activity-based compression function does not apply to net new writes or data considered active. They are compressed and allocated to a compression pool. This also applies to dedupe as net new writes may not be consuming drive capacity yet. This would be the first entry of a Hash ID into the Hash Table. Continued access to data generates statistics that are used to differentiate active data from idle data. Adding dedupe to this process enhances the space savings as only one instance of the data will consume physical storage.

There are three basic phases of capacity usage to recognize when running workloads on a system using Data Reduction. Throughout the three phases, the achieved compression ratio fluctuates. The three phases as related to used capacity are the initial load phase, climbing phase and optimal running phase.

6.1 Initial load phase

The initial load phase is identified when capacity usage is 0%-40%. During this time, it is possible that the achieved data reduction ratio varies. In most cases the capacity usage growth is due to new data being written to the system. As new allocations are passed through the data reduction hardware, this is likely to change due to compression and dedupe. Over time, data may become more active. In this case, it could be moved to the uncompressed pool (identified as the 128K pool), however, it may remain deduped as a single instance of data consuming the physical capacity. This alters both achieved system compression and storage group compression. This is a learning stage for the system as workload profiles and I/O patterns are identified. In addition, active and idle statistics are collected. The dynamic design of ACE uses the information gathered to change the compression pool layout online to one that best accommodates the workload.

6.2 Climbing phase

When capacity usage is between 41%-70% the system is in the climbing phase. The changing of the compression pool layout may be less active during this phase. The system’s back end layout will have stabilized and the system will only be expanding compression pools when needed. Regardless of the current achieved compression or data reduction ratio it will be moving towards the systems data reduction ratio reservation. This means the achieved system ratio could be increasing or reducing to meet the maximum supported level.

6.3 Optimal phase

Once the used capacity reaches 70% or greater the system is at the optimal operating state. The system achieved data reduction ratio starts to change moving closer to the expected value. As more capacity is consumed, the system becomes more capacity efficient. In this phase the backend configuration is less dynamic as the learning phase has created a compression pool layout that caters to the workload. If the system reaches 100% full, the achieved ratio is close to the expected value, however it may slightly vary. As an example, if the system’s data reduction reservation is 3.0:1 the achieved ratio may be 2.9:1. As capacity is consumed the system will react to achieve maximum expected savings.

7 Managing Data Reduction

Data reduction is enabled at the system level and includes both dedupe and compression together. Systems are installed with a data reduction reservation ratio. The reservation ratio provides the maximum achievable data reduction ratio that can be supported by the configured system resources.

Managing data reduction at the storage group level is supported with Unisphere™ and Solutions Enabler. In both cases this is done using the compression setting, which is enabled by default. When using this option, dedupe and compression are enabled or disabled together. The default SRP is applied as well, as this is required to make the compression option available. When using Solutions Enabler, it is required to include the SRP in the command syntax to successfully create a storage group with data reduction enabled. Although the function is enabled using the compression setting option, it enables both compression and dedupe at the same time. Setting either enable or disable simply turns the function on, telling the system that all I/O for that storage group is sent through the data reduction flow. Both compression and dedupe are limited by the boundaries of the SRP. While multiple SRPs are supported, each SRP will be treated as independent storage resources, neither compression nor dedupe will cross the SRP boundary.

7.1 Viewing Data Reduction

Data reduction savings are presented in two forms: system efficiency and storage group compression ratio. System savings are displayed in the capacity report and present savings for the overall system as well as savings specific to data reduction. Data reduction savings displayed for storage groups only report on compression savings and are presented as a compression ratio.

7.1.1 System efficiency in Unisphere

The data reduction ratio presented in the system capacity report includes both compression and dedupe savings. There are multiple views within Unisphere where the system level savings are displayed. There are a few values presented that relate to system savings. Overall efficiency values represent savings that include data reduction, snapshot savings and virtual provisioning. The savings presented at the system level is not exclusive to any one capacity savings feature. It is intended to deliver the capacity savings for the entire system. The default page presented offers a high-level view of the array capacity usage and efficiency.

Array usage:

- **Subscribed Capacity –** The bar presented represents the total provisioned capacity. The dark shaded portion of the bar indicates host allocations of the presented capacity.
- **Snapshot Capacity –** The total represents the sum of all existing snapshots. The dark shaded portion of the bar represents the amount of existing snapshot capacity that has been modified. The modified capacity also represents additional usable capacity that is consumed by snapshot data.
- **Usable Capacity –** The total amount of usable disk space available. The dark shaded portion represents the amount of disk space that is consumed.
- **Subscribed Usable Capacity –** The percent displayed represents the amount of subscribed capacity in relation to the amount of usable capacity. When this value is over 100% it is an indication that data reduction will be needed to accommodate allocations for the subscribed capacity. Users should be aware that the system is configured to accommodate data with a predetermined data reduction value. This means the subscribed usable capacity can exceed the maximum amount of data the system can physically accommodate even with data reduction applied.

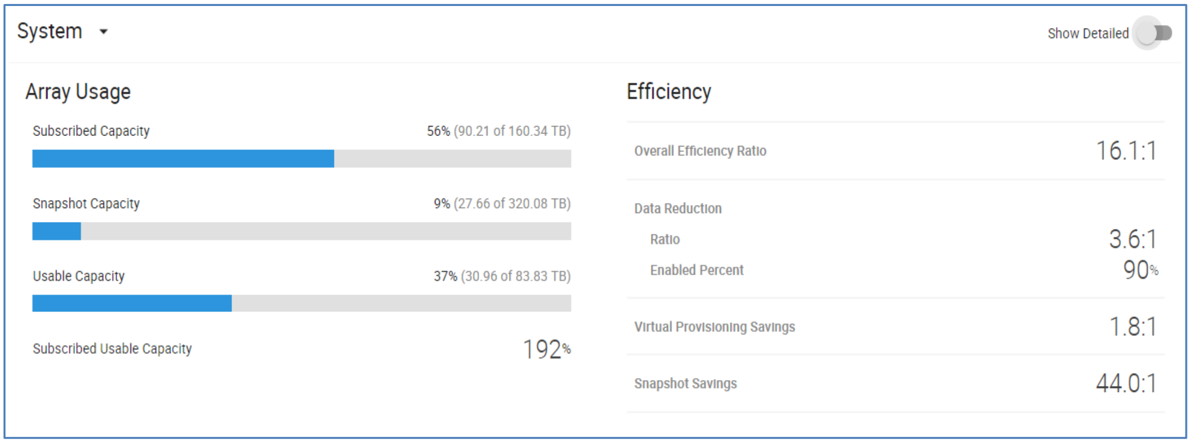

Figure 2 System efficiency report as seen in Unisphere for PowerMax

Efficiency:

- **Overall Efficiency Ratio –** The range of values that describe the capacity space savings that a user may experience regarding data reduction and/or other data services that offer capacity savings, such as Data Reduction, non-zero allocation, over provisioning and SnapVX.

$$\frac{Subscribed\ Total + Snapshot\ total}{User\ Used}$$

- **Data Reduction Ratio –** Savings that represents the combination of inline compression and inline deduplication presented as a ratio. To calculate the data reduction ratio the user needs to toggle to the detailed view shown below in figure 3. Additional information needed is revealed by hovering over the Usage bars. When calculating the Data Reduction ratio using the values presented in the usage portion of the capacity report. The ratio may reflect a different value due to the performance optimization leaving compressible data uncompressed and the enabled percent being under 100.

$$\frac{Subscribed\ Allocated + Modified\ Non\ Shared}{User\ Used}$$

- **Enabled Percent –** The amount of subscribed host allocations that have Data Reduction enabled.

- **Virtual Provisioning Savings –** Savings achieved relative to provisioned capacity and total usable capacity. This may exceed the maximum usable capacity.

$$\frac{Subscribed\ Total\ Capacity}{Subscribed\ allocated}$$

- **Snapshot Savings –** A representation of savings because of using SnapVX to create local replication data.

$$\frac{SnapShot\ Capacity\ Total}{Modified\ NonShared}$$

There is additional data within the report that can be presented to by using the Show Detailed option. The image displayed below presents the detailed view available in Unisphere.

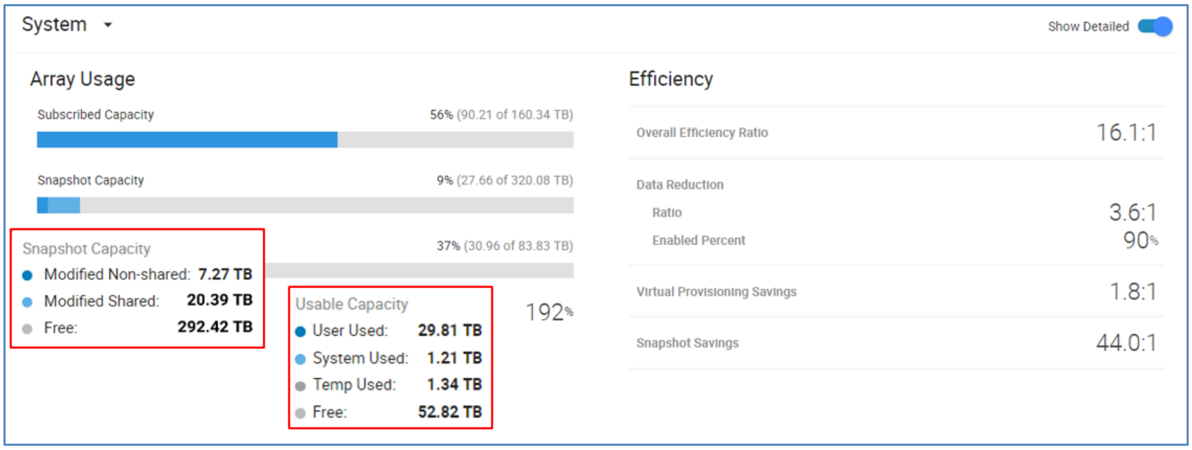

Figure 3 System Efficiency report detailed view as seen in Unisphere for PowerMax

7.1.2 Storage group compression in Unisphere

Data reduction savings presented at the storage group level represents only compression savings. Storage group compression ratios that are displayed represent the savings for data specific to the storage group being viewed. This can be seen in the storage group list as well as within the storage group details section as shown below by figures 4 and 5.

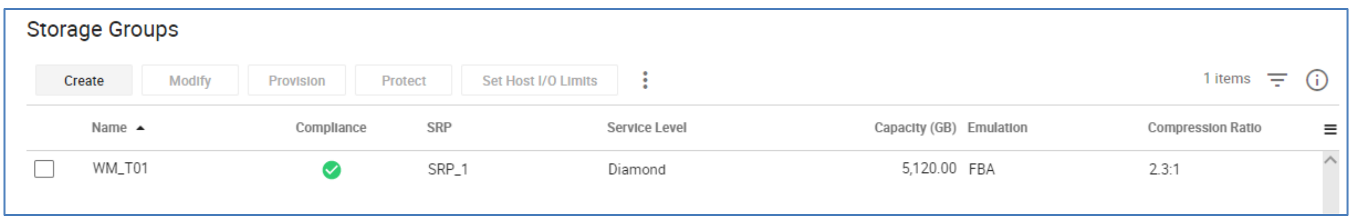

Figure 4 Compression ratio seen in the storage group list in Unisphere.

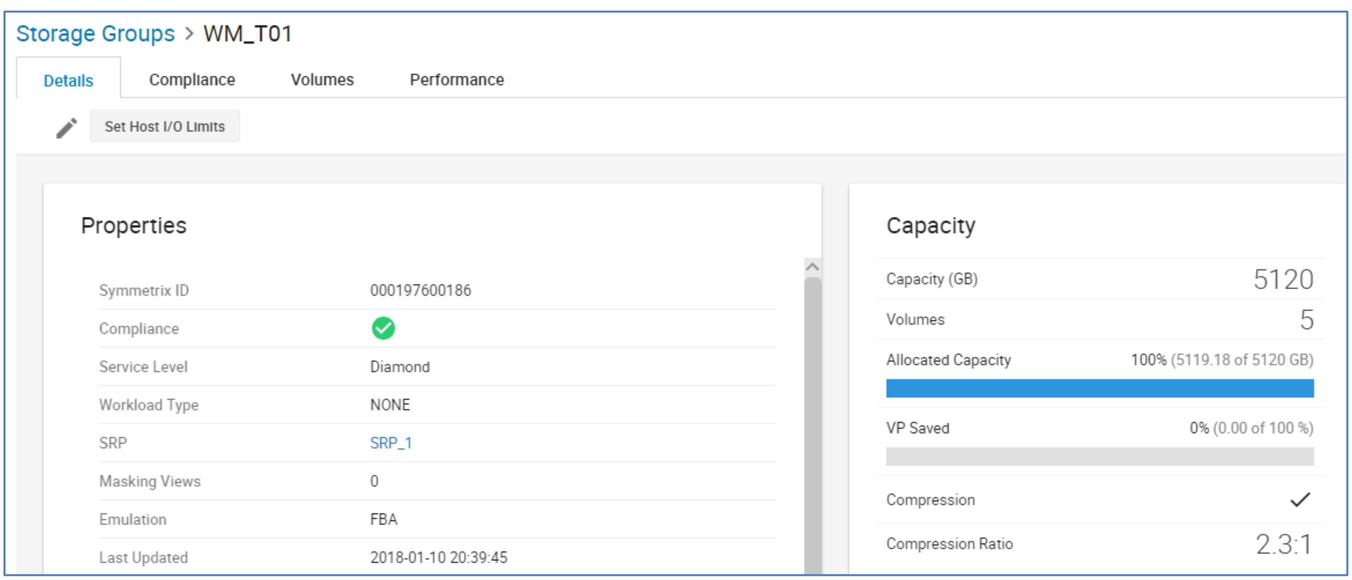

Figure 5 Compression data in storage group details view in Unisphere.

8 Supported data services

All data services offered in both the PowerMax and VMAX™ All Flash systems are supported with data reduction. This includes local replication (SnapVX), remote replication (SRDF™), D@RE, and VMware® vSphere® Virtual Volumes (vVols). Data reduction is supported for FBA. Mixed systems are supported, however data reduction will only apply to the FBA SRP.

8.1 Local replication (SnapVX)

Data reduction is supported with the use of local replication features; there are multiple variations and use cases for local replication. Below are the details regarding the different local replication sessions that can exist. For more detail regarding local replication and SnapVX see the TimeFinder™ and HYPERMAX OS Local Replication Technical Note available at DellEMC.com.

8.1.1 Nocopy sessions (SnapVX, VP Snap)

Uncompressed source data remains uncompressed when becoming snapshot data, and may be compressed later as it becomes less active. Activity to snapshot data through a linked target may prevent uncompressed data from being compressed. Compressed source data remains compressed when becoming snapshot data. Read activity to a snapshot through a linked target may cause the compressed data to be uncompressed.

The compression setting of a linked target only affects data written directly to the linked target and does not affect the snapshot data.

8.1.2 Copy sessions (SnapVX Full Copy Linked Targets, Clone, Mirror)

The compression settings for both the source and target are taken into account for copy sessions.

When compression is enabled on the source the data is decompressed before copying to the target. When compression is enabled on the target the data is compressed before being allocated to the target. Likewise, when compression is enabled on both the source and target the data is decompressed before the copy and then compressed to allocate for the target.

Copy times may vary due to decompression and compression of the data. It is not recommended to change the compression settings in between differential operations (i.e. disabling compression before each differential operation and then again after the copy completes) as this causes data to go through needless compress/decompress cycles.

8.2 Remote replication (SRDF)

Compression for SRDF is already supported and known as SRDF compression. SRDF compression is a feature designed to reduce bandwidth consumption while sending data to and from systems connected using remote replication. SRDF compression and the Adaptive Compression Engine (ACE) both use the same compression module, however, they serve different purposes. Data that has been compressed using ACE is uncompressed before being sent across the SRDF link. In the event that SRDF compression and inline compression apply, the data is uncompressed by the module and then compressed using the SRDF compression function and then sent across to the remote site.

8.3 Data at Rest Encryption

Data reduction is performed as a post-process function. This means that the data being encrypted has already been passed through the data reduction I/O module to be deduped and compressed. The encryption process occurs against deduped and compressed data.

8.4 Virtual Volumes

Data reduction is supported for the allocation of data to VMware® vSphere® Virtual Volumes™ (vVols) and follows the same I/O path as all data as shown in figures 2 to 4. Compression as a feature is not included as a vVols resource to be configured at the host.

9 Conclusion

The use of physical storage capacity is a common concern of storage administrators across the storage industry. The constant and ever growing amounts of data have created the need for more efficiency in the use of physical capacity. Dell EMC PowerMax and VMAX All Flash data storage systems take this to the next level. Combining inline compression with inline dedupe provides exceptional capacity savings with negligible cost to performance. This delivers on capacity savings, which leads to a smaller data center footprint and an overall reduction in TCO. In addition to the savings, using data reduction is as simple as a single click to enable or disable. The system handles all the work.

A Technical support and resources

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

Storage technical documents and videos provide expertise that helps to ensure customer success on Dell EMC storage platforms.