



Hewlett Packard
Enterprise

Technical white paper

HPE PRIMERA DATA REDUCTION



CONTENTS

HPE Primera Data Reduction technologies.....	4
Thin provisioning.....	4
Compression.....	4
Deduplication.....	5
Deduplication and compression.....	5
Data reduction and encryption.....	5
Choosing the volume type.....	6
Thin.....	6
Data reduction.....	6
Estimating data reduction space savings.....	6
HPE Storage Assessment Foundry.....	7
Common provisioning groups.....	7
Reasons to create multiple CPGs.....	7
CPG automatic growth.....	8
Backup and copy data management.....	8
Creating volumes.....	8
Capacity efficiency.....	9
Understanding capacity efficiency ratios.....	14
Managing overprovisioning.....	14
Tracking volume space usage.....	15
Volumes.....	15
CPGs.....	15
System space.....	16
Monitoring and alerting.....	17
Allocation warnings.....	17
Remaining physical capacity alerts.....	18
Remaining CPG free space alerts.....	19
View logged warning and limit alerts.....	19
User recommendations.....	19
Migrate between volume types on the same array.....	19
HPE Primera Thin Persistence software.....	20
HPE Primera Thin Persistence reclamation.....	20
HPE Primera Thin Persistence methods.....	21
Appendix A—HPE Primera Thin Persistence methods.....	21
HPE Primera Thin Reclamation for Microsoft Windows Server® 2012.....	21
HPE Primera Thin Reclamation for Microsoft Windows Server 2003 and 2008.....	22
HPE Primera Thin Reclamation for VMware vSphere.....	22
HPE Primera Thin Reclamation for Linux.....	25
HPE Primera Thin Reclamation for HP-UX.....	26
HPE Primera Thin Reclamation for UNIX®.....	26
HPE Primera Thin Reclamation for Symantec Storage Foundation.....	26
HPE Primera Thin Reclamation for Oracle Databases.....	27



Appendix B—File systems and deduplication.....28

 Microsoft Windows.....28

 Microsoft Hyper-V.....30

 VMware vSphere.....31

 Oracle Solaris.....31

 Linux.....32

 Symantec Storage Foundation.....33

 HP-UX.....33

 IBM AIX.....34



HPE PRIMERA DATA REDUCTION TECHNOLOGIES

Data is being generated at a rate, which exceeds storage capacity growth; therefore, the storage of data has become increasingly challenging. Also, there is an industry-wide transition of primary storage to solid-state drives (SSDs), which has a higher per-byte cost compared to hard disk drives (HDDs). For these reasons, price capacity or cost per byte, in addition to price performance, is emerging as a critical factor for enterprise storage. With HPE Primera Data Reduction, HPE Primera storage sets a new standard for total system efficiency that not only reduces the cost of flash but extends flash media endurance and offers the lowest TCO of any all-flash array.

HPE Primera Data Reduction technologies including Compression, Deduplication, Thin Provisioning, Thin Conversion, Thin Persistence, and Thin Copy Reclamation achieve advanced data reduction through leveraging built-in hardware capabilities and HPE Primera Express Indexing technology.

While HPE Primera Data Reduction is extremely easy to deploy and use, a certain amount of planning is advantageous to help maximize its benefits. This paper documents best practices for data reduction on HPE Primera storage and is intended for administrators looking to get the most out of their HPE Primera deployment. Also, it describes other HPE Primera Thin Technologies that you can use in conjunction with HPE Primera Data Reduction to help maximize its effectiveness.

Like all data reduction technologies, HPE Primera Data Reduction is designed to reduce the cost of flash, making it more affordable. However, some workloads have compression at the application layer—examples of this would be compressed video and databases that apply compression before the data is written. In these scenarios, data reduction in the storage system will likely provide no additional savings as the efficiencies have already been made. When this is the case, it can be beneficial to not use data reduction at the array level on these volumes to free up resources to be used when compressing other workloads.

To address this, HPE Primera systems support selective data reduction—this allows the storage administrator to selectively enable data reduction on a per-volume basis for complete flexibility. They can choose between thin and data reduction volume types to optimally match the application characteristics or performance requirements.

Thin provisioning

Thin provisioning allows a volume to be created and made available to a host without the need to dedicate physical storage until it is actually used. HPE Primera Thin Provisioning software has long been considered the gold standard in thin provisioning for its simplicity and efficiency. It is the most comprehensive thin provisioning software solution available, allowing enterprises to purchase only the disk capacity they actually need and only when they actually need it.

Thin provisioning breaks the connection between logical and physical capacity and allows the virtual presentation of more capacity than is physically available. This is the enabling technology behind data reduction to further reduce the footprint of data, thereby increasing the effective capacity.

HPE Primera Thin Persistence software enables thin-provisioned storage on HPE Primera arrays to stay thin over time by helping ensure that unused capacity is reclaimed for use by the array on an ongoing basis. At its heart is the HPE Primera Zero Detect technology that identifies and removes repeated zeros from incoming data streams, reducing the amount of capacity required to store data. HPE Primera Zero Detect is implemented as a dedicated hardware engine in each HPE Primera ASIC and is therefore available in all HPE Primera systems and works independently of other data reduction technologies—meaning savings from this technology can be made on all data.

Compression

Compression works by looking inside data streams for opportunities to reduce the size of the actual data. The inline lossless compression algorithm of HPE Primera is specifically designed to work on a flash-native block size to drive efficiency and performance and leverages a series of different technologies to offer the highest possible savings.

HPE Primera compression is based on the LZ4 algorithm. This algorithm is very efficient; it offers not only good compression ratios and compression speed but also exceptional decompression speed, which is crucial to storage performance. Writes are mirrored in cache before being acknowledged back to the host. The actual compression, while an inline process, is performed after the host write is acknowledged and is therefore asynchronous to the host I/O. However, read requests are synchronous, as the host has to wait for the data, so good decompression performance is essential. Decompression is also important to overwrite, as existing data may need to be decompressed and merged with the new incoming data before being compressed and written to the SSDs.

HPE Primera systems use 16 KiB physical pages to store data. When compression is used, HPE Primera Data Packing allows multiple compressed pages (up to eight) to be stored in a single 16 KiB physical page. This data packing technique is part of the inline process that not only optimizes physical space but also creates larger, more media efficient, write sizes than other approaches, which improves both performance and media endurance. If an overwrite occurs, the system will refit the new compressed page in place if there is available space, thus reducing the amount of system garbage collection that is needed, but if the new data will not fit into the existing page it will be queued



for packing with other new writes. HPE Primera Express Indexing technology, common to all HPE Primera thin volume types, is used to track the data in a compressed volume.

HPE Primera compression also features Express Scan, a technology designed to identify incompressible streams of data and store them in their native format instead of wasting CPU cycles attempting to compress incompressible data. This greatly improves the performance of the system, adding a new dimension to efficiency.

Compression compliments HPE Primera Zero Detect and deduplication to further reduce the amount of flash required to store a given amount of data. The combination of these three technologies is key to reducing the effective cost of flash below that of spinning disk and therefore making flash affordable for mainstream applications.

Deduplication

Deduplication works by looking inside data streams for duplicate blocks of information. It has become standard with disk-based backup due to a high degree of data redundancy and less emphasis on performance. Backup and archive workloads have been an ideal target for deduplication technologies. Traditional primary storage workloads have lower data redundancy and hence lower deduplication ratios, and therefore deduplication of primary storage has not been seen as beneficial. However, the landscape around primary storage deduplication is changing as the widespread deployment of server virtualization is driving the demand for primary storage deduplication.

The inline deduplication algorithm of HPE Primera is specifically designed to work on a flash-native block size to drive efficiency and performance and leverages a series of different technologies to offer the highest possible savings. As blocks of a data stream are written to a data reduction volume, they are fingerprinted by the HPE Primera ASIC and the fingerprints are checked for matches with existing data. If a match occurs, the ASIC does a bit-by-bit verification of the incoming data block with the data of matched fingerprint to guarantee the data is identical before storing the block into the Dedup Store. HPE Primera Express Indexing is used to track the blocks of the volume within the local datastore and the Dedup Store.

Deduplication and compression

HPE Primera Data Reduction combines deduplication and compression to help maximize space savings. For data reduction volumes, the data will be checked for duplicate blocks first, and the unique blocks will then be compressed. Note that the dedup occurs between all the volumes in the CPG.

The following should be taken into account before implementing data reduction volumes:

- Only HPE Primera systems with an SSD tier can take advantage of data reduction.
- It is only applicable to virtual volumes residing solely on SSD storage.
- The minimum size of a data reduction volume is 16 GiB and the maximum size is 16 TiB.
- The granularity of HPE Primera Deduplication is 16 KiB, and therefore the efficiency is greatest when I/Os are aligned to this granularity. For hosts that use file systems with tuneable allocation units, consider setting the allocation unit to 16 KiB or a multiple of 16 KiB. For more information on HPE Primera Deduplication and file systems, see [Appendix B](#). For applications that have tuneable block sizes, consider setting the block size to 16 KiB or a multiple of 16 KiB.
- Deduplication is performed not only on the data contained within the virtual volumes but also between virtual volumes in the same CPG. For maximum deduplication, store data with duplicate affinity on virtual volumes within the same CPG.
- Deduplication and compression are space-saving techniques that can reduce the cost differential between SSDs and HDDs. For maximum performance, consider using thin-provisioned volumes.

HPE Primera systems with SSDs allow selective enabling of data reduction on a per-volume basis. This lets the storage administrator choose the optimal volume type, whether it is thin or data reduction to match the application characteristics or performance requirements.

DATA REDUCTION AND ENCRYPTION

Data security is an important consideration for the modern enterprise. Encryption is a key component of a security strategy, but it is crucial in the I/O path. Using host-based encryption will nullify any storage data reduction features and can therefore increase the cost of storage.

HPE Primera arrays offer data-at-rest encryption using FIPS-2 compliant Self-Encrypting Drives (SEDs) and the ability to use an external enterprise key manager (EKM). The SED is a drive with a circuit (ASIC) built into the drive controller's chipset, which encrypts/decrypts all data to and from the drive media automatically and is therefore fully compatible with HPE Primera Data Reduction.



CHOOSING THE VOLUME TYPE

Thin

The use of data reduction technologies has the significant operational benefit of reducing storage consumption. However, there are certain scenarios where data reduction may not be of benefit and regular thin provisioning can be a better choice such as:

- The data is to be stored solely on HDDs.
- Environments that use application or file system based encryption, deduplication, or compression.
- High write workloads—With thin provisioning, metadata is only updated when space is allocated and subsequent overwrites do not generate any metadata updates. Therefore, intensive write workloads can achieve higher total IOPS at lower latency without data reduction.

Data reduction

Data reduction volumes combine deduplication and compression techniques and are ideal for data that has a high level of redundant or compressible nonredundant data. Data sets that are good candidates for data reduction volumes include:

- Virtual machine (VM) images—The OS binaries from multiple VMs can be reduced to a single copy by deduplication and the application data within the VMs can benefit from compression.
- Virtual desktop infrastructure (VDI)—Client virtualization environments with both hosted persistent and nonpersistent desktops can achieve excellent data reduction ratios.
- Databases—Most databases do not contain redundant data blocks but do have redundant data within blocks so they can benefit from compression.
- Home directory and file shares—Storage deduplication and compression can offer significant space savings.

Data with a low level of redundancy should not be stored on data reduction volumes. Data sets that are not good candidates include:

- Compressed data—Compression creates a stream of unique data that will not benefit from storage data reduction. The most common types of compressed data are video and image files.
- Encrypted data—The use of host or SAN encryption will also result in a stream of unique data that will not benefit from storage data reduction.

Use the data reduction estimate tool to check the data reduction ratio of existing volumes before conversion to data reduction volumes.

ESTIMATING DATA REDUCTION SPACE SAVINGS

Data reduction estimate tools are available to show the space-saving benefits of HPE Primera Data Reduction on existing data without the need to convert the volumes. The tools will estimate the amount of space savings that can potentially be achieved by finding common data across specified volumes and provides the data reduction ratio based on the calculated data.

To launch the estimator from the CLI, use the `checkvv` command with the `-reduce_dryrun` option on a group of VVs or a VV set to perform a dry run conversion, which will report the space savings data reduction would achieve if the VVs specified were in the same CPG. The specified VVs can be full- or thin-provisioned, and they can reside on any type of drive not just SSDs. The estimated data reduction ratio will be shown under the task's detailed information, which can be accessed via `showtask -d`. The following example shows a data reduction estimation task being started and its associated task results:

```
cli% checkvv -reduce_dryrun vv2
```

The `-reduce_dryrun` command relies on HPE Primera compression and deduplication technology to emulate the total amount of space savings to be achieved by converting one or more input VVs to data reduction VVs. Please note that this is an estimation and results of a conversion may differ from these results.

The command can be run against live production volumes and will generate a non-intrusive background task. The task may take some time to run, depending on the size and number of volumes, and can be monitored via the `showtask` commands.

Do you want to continue with this operation?



```
select q=quit y=yes n=no: y
```

Task 12092 has been started to validate administration information for VVs: vv2

```
cli% showtask -d 12092
```

```
Id Type      Name      Status Phase Step -----StartTime----- -----FinishTime----- -Priority- -User--
12092 reduce_dryrun checkvv done    ---    --- 2019-07-09 22:16:09 BST 2019-07-09 22:44:27 BST n/a      3parsvc
```

Detailed status:

```
2019-07-09 22:16:09 BST Created      task.
2019-07-09 22:16:09 BST Started      checkvv space estimation started with option -reduce_dryrun
2019-07-09 22:44:27 BST Finished      checkvv space estimation process finished
```

```
-(User Data)- -(Compression)- -----(Dedup)----- -(DataReduce)-
Id Name      Size[MB] Size[MB] Ratio      Size[MB] Ratio Size[MB] Ratio
166 vv2      387807  198876  1.95      --      --  198876  1.95
-----
1 total      387807  198876  1.95      387808  1.00  198876  1.95
```

HPE Storage Assessment Foundry

The HPE Storage Assessment Foundry also provides tools for analyzing an existing storage environment. The assessment will provide a detailed report on performance and capacity including guidance on the storage capacity efficiency ratios, which can be achieved when the data is stored on an HPE Primera array. For a free storage efficiency assessment, contact your HPE sales or channel partner representative.

COMMON PROVISIONING GROUPS

Common provisioning groups (CPGs) are policies for how free chunklets within the HPE Primera should be used when creating volumes. A CPG policy contains parameters such as disk type and availability level. CPGs automatically grow the underlying logical disk (LD) storage, according to calculated growth parameters, on-demand to store data in a thin volume.

Reasons to create multiple CPGs

All TPVVs and data reduction VVs associated with a CPG allocate space from a shared pool of LDs. This means that VVs associated with a particular CPG have identical LD characteristics. VVs that require different characteristics must use a different CPG.

Reasons to create multiple CPGs include:

- To map VVs belonging to different lines of business, departments, or customers onto particular CPGs for reporting and management purposes. This allows a logical separation of resources and may help with chargeback models, as chargeback could be based on CPG space usage rather than usage at an individual VV level.
- When virtual domains are used (a CPG can only belong to one virtual domain).
- When using HPE Primera Data Reduction, there is a limit of 2048 data reduction volumes per CPG on all HPE Primera 600 systems. Since dedup is performed at the CPG level, volumes with similar data sets should be grouped together to increase the likelihood of deduplication on larger systems with more than 2048 data-reduction-enabled VVs.

While there are several reasons to create multiple CPGs, it is recommended that the number of CPGs be kept to a minimum, as each CPG will reserve its own growth space. For data reduction, the recommended maximum number of CPGs is shown in Table 1.

TABLE 1. Recommended maximum data reduction CPGs

System type	Maximum data reduction CPGs
HPE Primera 630	4
HPE Primera 650 2N	6
HPE Primera 650 4N	12
HPE Primera 670 2N	8
HPE Primera 670 4N	16



CPG automatic growth

The CPGs dynamically allocate storage in increments, which are determined by the number of nodes in the system and the RAID set size that has automatically been selected. This on-demand allocation unit determines the automated response to the growth demands of volumes. To grow the volume, the HPE Primera OS may expand existing LDs according to the CPGs growth increment or create additional ones. Growth is triggered when the CPG's available space falls below 85% of the growth increment value. A mechanism with warnings and limits can be configured on the array to control the growth of a CPG. If the CPG cannot grow, then when the free space in the CPG is exhausted, I/Os that require space to be allocated will fail.

BACKUP AND COPY DATA MANAGEMENT

While a true backup should reside on a separate storage platform for disaster recovery purposes, it is common to use backup software to create local point-in-time backup copies of application data for quick recovery.

It is often assumed that these backup copies are of the same format as the database and will therefore be dedupable with the database, but depending on the backup method, that may not be the case. However, the backup copies should be dedupable amongst themselves, so it is good practice to keep them in a separate CPG.

Backup copies to data reduction volumes may not consume significant additional space but they can impact performance as all the application data has to be read from and written to the same array, which increases the loading on the host, SAN, and array.

A superior option that does not involve host data copies is to use snapshots to create virtual copies of the data. The only drawback is the need to quiesce the data before the snap, so the data is consistent. This is usually done by creating a script to quiesce the application and then coordinate with array to perform the snapshot. The HPE solution to copy data management is to use the HPE Recovery Manager Central (RMC) software, which allows users to create and automate application-consistent snapshots. HPE RMC supports multiple applications including VMware vSphere®, Microsoft® SQL Server, Microsoft exchange, Oracle, and SAP HANA® databases.

CREATING VOLUMES

HPE Primera systems support thin-provisioned volumes with optional data reduction. Volumes can be created using the HPE StoreServ Management Console (SSMC), CLI, or Web Services API. Figure 1 shows the HPE SSMC Virtual Volume creation pane.

Create Virtual Volume General ▾ ?

General ☐ Advanced options

Name

System

Domain

Provisioning

Dedup and Compression ☒ Yes

To achieve the best space efficiency, deduplicated virtual volumes should be formatted at the host using a 16 KiB allocation unit.
If the volume will be used for a database, HPE recommends using a Thin Provisioned volume instead of a deduplicated volume.

CPG x 🔍

RAID 6 SSD

Size

App Volume set 🔍

FIGURE 1. Creating a volume in HPE SSMC

CAPACITY EFFICIENCY

The HPE Primera OS has several metrics to measure the capacity efficiency of the system: compaction ratio, dedup ratio, compression ratio, and data reduction ratio.

- The compaction ratio is how much logical storage space a volume consumes compared to its virtual size and applies to all thin volume types.
- The dedup ratio is how much storage space is being saved by deduplication on data reduction volumes.
- The compression ratio is how much storage space is being saved by compression on data reduction volumes.
- The data reduction ratio is how much storage space is being saved by the combination of both deduplication and compression.

The ratios are shown as decimals and the :1 is not displayed, that is, 4.00 is actually 4:1 (4 to 1). The data reduction, dedup, and compression ratios do not include savings from inline zero detection, as these are included in the compaction ratio.

The capacity efficiencies can be shown per volume, volume family, CPG, virtual domain, or system and are in terms of usable storage (that is, not including RAID overhead). The efficiencies displayed will depend on the scope of the request. For example, deduplication is performed at the CPG level, so dedup and data reduction ratios are not displayed for individual volumes.

Base volumes

The capacity efficiencies of a base volume are shown by the `showvv -space` command and are calculated as follows:

- The compaction ratio of a thin volume is the virtual size of the volume divided by its used data space.
- The compression ratio of a compressed or data reduction volume is the size of the data stored in the volume divided by the space used by the volume after compression.
- There are no dedup ratios on a per-volume basis; dedup ratios are shown on a per CPG basis.

In the `showvv` output, the `Snv` columns show the space associated with virtual copies and the `Usr` columns show the space associated with the base volumes. The `Total` columns show the sum of the `Usr` and `Snv` counts. For each grouping, the `Rsvd` column shows the allocated space and `Used` shows how much is in use.

The `HostWr` column shows how much data is currently in use by the host. If data is freed via UNMAPs or is zeroed out, then the `HostWr` total will decrease. A large discrepancy between the `Dedup Store` of the CPGs containing the deduplicated portion of data reduction volumes is shown as a volume starting with `shared` and the volumes associated with a `Dedup Store` are listed following it.

In the following example, there is a thin-provisioned volume named `vv1` and a data reduction volume named `vv2`. Each volume is in its own CPG and contains exactly the same data, which consists of six VMs.

```
cli% showvv -space -cpv CPG?
```

					-----Snp-----					-----Usr-----					-----Total-----									
					--(MiB)-- --(% VSize)--					--(MiB)-- --(% VSize)--					--(MiB)-- --(% VSize)--					---Efficiency---				
Id	Name	Prov	Compr	Dedup	Type	Rsvd	Used	Used	Wrn	Lim	Rsvd	Used	Used	Wrn	Lim	Rsvd	Used	HostWr	VSize	Compact	Compress			
175	.shared.CPG2_0	dds	NA	NA	base	0	0	0.0	--	--	139776	130748	0.2	--	--	139776	130748	--	67108864	--	--			
176	vv2	tdvv	Yes	Yes	base	0	0	0.0	--	--	261120	191965	18.3	0	0	261120	191965	894533	1048576	5.46	2.01			
174	vv1	tpvv	No	No	base	0	0	0.0	--	--	900608	894534	85.3	0	0	900608	894534	894533	1048576	1.17	--			
3 total						0	0				1301504	1217247				1301504	1217247	1789066	69206016					

As the compaction ratio of a thin volume is the virtual size of the volume divided by its used data space, this can be verified by taking the `VSize` of the volume and dividing it by the `Usr Used` size, that is, for `vv1`, the compaction ratio is $1048576/894534 = 1.17$.

The compression ratio of a data reduction volume is the size of the data stored in the volume divided by the space used by the volume after compression. The compression ratio for a data reduction volume such as `vv2` cannot be manually calculated since the data stored in the volume is not readily available as the `HostWr` value includes dedupable data stored in the `Dedup Store`.

The base volume savings can also be seen in the HPE StoreServ Management Console (SSMC) by selecting the particular VV. Figure 2 shows the space savings displayed for `vv1`.



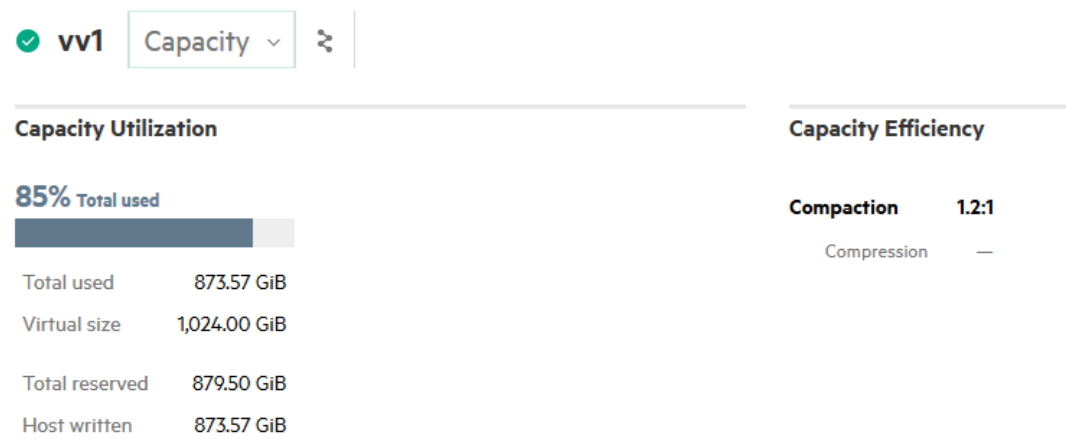


FIGURE 2. Virtual volume space savings in HPE SSMC

Common provisioning groups

The capacity efficiencies of a CPG are shown by the showcpg -space command and are calculated as follows:

- The compaction ratio is the sum of the virtual sizes of all the volumes and snapshots in the CPG divided by the sum of their in-use data space, snapshot space, and used Dedup Store space.
- The dedup ratio of a CPG is the sum of all the data written to the data reduction volumes divided by the sum of the data stored in the data reduction volumes and the data stored in the Dedup Store.
- The compression ratio of a CPG is the sum of all the sizes of the data reduction volumes divided by the sum of the space used by these volumes after compression.
- The data reduction ratio of a CPG is the sum of all the data written to data reduction volumes divided by the sum of the space used by the volumes and the data associated with the volumes in the Dedup Store. The compaction ratio differs from the data reduction ratio in that it incorporates the thin-provisioning savings (including zero detection).

In this example, there are CPGs named CPG1 and CPG2 containing the volumes from the base volume example.

```
cli% showcpg
      ----Volumes---- -Usage- -----[MiB]-----
Id Name  Warn% VVs  TPVVs  TDVVs  Use  Snp   Base  Snp   Free   Total
5 CPG1    -    1    1      0    1    0  900608  0  20480  921088
6 CPG2    -    2    0      1    2    0  261120  0   8448  409344
-----
2 total              3    0 1161728  0  28928 1330432

cli% showcpg -space
      -Private[MiB]- -----[MiB]----- Efficiency -----
Id Name  Warn%   Base   Snp  Shared  Free   Total Compact Dedup Compress DataReduce Overprov
5 CPG1    -   900608    0      0  20480  921088   1.17    -      -      -      0.11
6 CPG2    -   261120    0 139776   8448  409344   3.25  1.72   2.01   2.76   0.02
-----
2 total      1161728    0 139776 28928 1330432   1.72  1.72   2.01   2.76   0.07
```



The base volume savings can also be seen in the management console by selecting the particular CPG. Figure 3 shows the space savings displayed by the HPE StoreServ Management Console (SSMC) for CPG2.

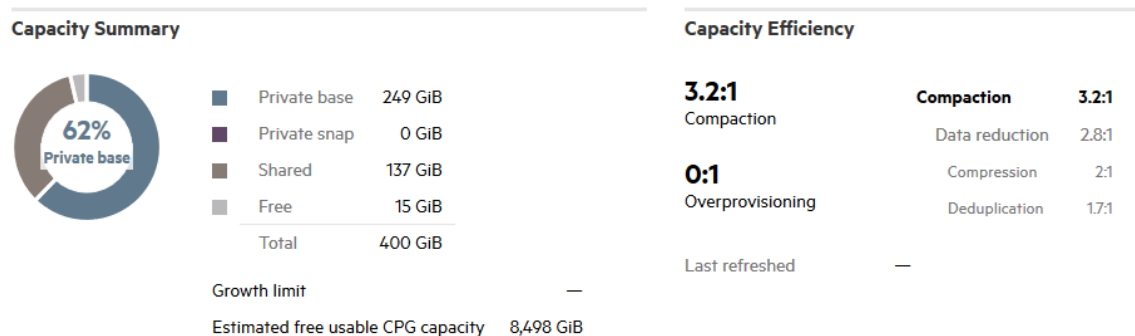


FIGURE 3. CPG space savings in HPE SSMC

Volume families

If a volume has snapshots, then the `showvv` command will display the individual snapshots below their parent volumes. However, the capacity efficiency displayed is that of the entire volume family because all the snaps of a base volume share the same snapshot area. If a block changes on a base volume, the old data is copied to the snapshot area for that volume and all snaps of the volume that were created before that block was updated to that single block. This allows a volume to have hundreds of snaps without requiring additional space or incurring additional performance impact.

The capacity efficiencies of a volume family are shown by the `showvv -space` command and are calculated as follows:

- The compaction ratio of a thin volume family is the sum of the virtual size of the volume and the virtual volume sizes of all its snapshots divided by its used data space of the volume and all its snapshots.
- The compression ratio of a data reduction volume is the size of the data stored in the volume and all its snapshots divided by the data space used by the volume and all its snapshots.

In this example, a snapshot was created from the previous thin volume `vv1` and the data was changed by adding a new VM and deleting an existing one. The changed data now shows in the Snp usage columns of the parent volume. The result is the compaction ratio for all the volume families has increased because the snapshot virtual size is now included, which doubled the virtual size to 2 TiB, but the snapshot only contains an extra 127 GiB of data.

```
cli% showvv -space -cpv CPG1
```

```
-----Snp-----Usr-----Total-----
----[MiB]-----[% VSize]-- --[MiB]----[% VSize]-- -----[MiB]----- ---Efficiency---
Id Name      Prov Compr Dedup Type   Rsvd  Used Used Wcn Lim  Rsvd  Used Used Wcn Lim  Rsvd  Used HostWc  VSize Compact Compress
174 vv1       tppv No    No    base  135680 126975 12.1  0  0  900608 894534 85.3  0  0  1036288 1021509 894534 1048576 3.08  --
180 vv1-snp   snp  NA     NA    vcopy  --    *0 *0.0  0  0  --    --    --    --    --    --    --    -- 1048576  --    --
-----
2 total      135680 126975 900608 894534 1036288 1021509 894534 2097152
```

The space usage columns for the snapshots contain “--” as the space usage of the snapshots is maintained in the base volume.

Virtual domains

The capacity efficiencies of a virtual domain are shown by the `showsys -domainspace` command and are calculated as follows:

- The compaction ratio is the sum of the virtual sizes of all the volumes and snapshots in the virtual domain divided by the sum of their in-use data space, snapshot space, and used Dedup Store space.
- The dedup ratio of a virtual domain is the sum of all the data written to the data reduction volumes divided by the sum of the data stored in the volumes and the data associated with the volumes in the Dedup Store.
- The compression ratio of a virtual domain is the sum of all the sizes of the data reduction volumes divided by the sum of the space used by these volumes after compression.
- The data reduction ratio of a virtual domain is the sum of all the data written to data reduction volumes divided by the sum of the space used by the volumes and the data associated with the volumes in the Dedup Store. The compaction ratio differs from the data reduction ratio in that it incorporates the thin-provisioning savings (including zero detection).

In this example, there are two virtual domains, dom0 and dom1, in addition to the base system domain. The base domain contains just thin volumes and therefore has no values in the Dedup, Compr, or DataReduce columns. Both dom0 and dom1 have data reduction volumes, so they do have data reduction ratios.

```
cli% showsys -domainspace
      -Legacy[MiB]- -Private[MiB]- ---CPG[MiB]--- -----[MiB]----- -----Efficiency-----
Domain  Used    Snp    Base  Snp Shared   Free Unmapped   Total Compact Dedup Compress DataReduce Overprov
-              0      0 1399464 1368 91281  88943      0 1581056    3.1    -      -      -      0.1
dom0      0      0 16022470      0 91281 2809769      0 18923520    5.5  1.4    1.49    1.45    1.7
dom1      0      0   74899      0 91281  227036      0   393216    2.5  1.4    1.53    1.96    0.1
-----
              0      0 17496833 1368 273843 3125748      0 20897792    3.2  1.4    1.50    1.73    1.2
```

System

The capacity efficiencies of the system are shown by the `showsys -space` command and are calculated as follows:

- The compaction ratio is the sum of the virtual sizes of all the volumes and snapshots in the system divided by the sum of their in-use data space, snapshot space, and used Dedup Store space.
- The dedup ratio of a system is the sum of all the data written to the data reduction volumes divided by the sum of the data stored in the volumes and the data associated with the volumes in the Dedup Store.
- The compression ratio of a system is the sum of all the sizes of the data reduction volumes divided by the sum of the space used by these volumes after compression.
- The data reduction ratio of a system is the sum of all the data written to data reduction volumes divided by the sum of the space used by the volumes and the data associated with the volumes in the Dedup Store. The compaction ratio differs from the data reduction ratio in which it incorporates the thin-provisioning savings (including zero detection).

This example shows how the system-wide capacity efficiencies are displayed.

```
cli% showsys -space
----- System Capacity [MiB] -----
Total Capacity      : 20938752
Allocated           : 10613760
  Legacy Volumes    :      0
  User              :      0
  Snapshot          :      0
  CPGs [VVs]        : 6561792
  Shared            : 1710844
  Private           : 4724846
  Base              : 4549224
  Reserved          : 4549224
```



Reserved [vSphere VVols]	:	0
Snap	:	175622
Reserved	:	175622
Reserved [vSphere VVols]	:	0
Free	:	126102
Unmapped	:	0
System	:	4051968
Internal	:	1035264
Admin	:	927744
Space	:	2088960
Used	:	0
Unused	:	2088960
Free	:	10324992
Initialized	:	10324992
Uninitialized	:	0
Unavailable	:	0
Failed	:	0
----- Efficiency -----		
Compaction	:	2.95
Dedup	:	1.36
Compression	:	1.43
Data Reduction	:	1.71
Overprovisioning	:	0.72

Figure 4 shows the system-wide space savings displayed by the HPE StoreServ Management Console (SSMC).

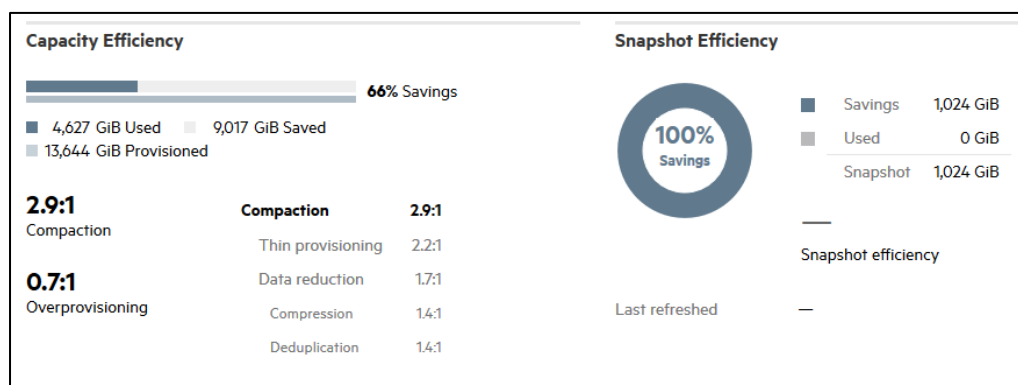


FIGURE 4. System space savings

Not all space on the physical disks is used for storing your data. A small portion of the space on the array is dedicated to volumes with an administrative function.

There is a full-provisioned volume named admin that is used to store system administrative data such as the System Event Log. The logging LDs, starting with the name log, are used to store data temporarily during physical disk failures and disk replacement procedures. There are also preserved data space logical disks (PDSLDs), which start with the name pdsld. Preserved data is the data moved from the system's cache memory to the PDSLD space in the eventuality of multiple disk or cage failures. The HPE Primera System Reporter software is integrated into the OS and executed on the controller nodes, and the database files are stored in a full-provisioned volume called .srdata.

Understanding capacity efficiency ratios

It may not be immediately apparent that even low capacity efficiency ratios indicate significant space savings. As capacity efficiency ratios increase, there are diminishing returns in terms of space savings. The percentage of space reduction obtained is 100% less the inverse of the capacity efficiency ratio. Space savings for selected capacity efficiency ratios are shown in Table 2.

TABLE 2. Space savings by capacity efficiency ratio

Capacity efficiency ratio	Space reduction percentage
1:1	0%
1.5:1	33%
2:1	50%
4:1	75%
10:1	90%

MANAGING OVERPROVISIONING

Overprovisioning is the provisioning of virtual capacity to hosts in excess of the amount of usable storage in the system. Thin provisioning allows volume capacity unused by the host to not consume disk resources. Therefore, to be able to utilize all the system capacity, it is necessary to provision more storage than physically available. The use of data reduction further decreases the storage consumption, which increases the necessity for overprovisioning.

Thus, it is essential to monitor the amount of overprovisioning as well as the actual capacity utilization to avoid a situation where the system runs out of space. HPE Primera systems can show the amount of overprovisioning as a ratio at the CPG, virtual domain, or system level.

The overprovisioning ratio is calculated as:

- Overprovisioning ratio = virtual size/(allocated space + free space)

The virtual size is the sum of the effective virtual size of the volumes and snapshots. The effective virtual size depends on the volume type and is calculated as follows:

- Thin provisioned VVs will use virtual size.
- Data reduction VVs will use virtual size/data reduction ratio.

To account for CPGs with different RAID set sizes all sizes are converted to raw sizes.

To see the CPG overprovisioning ratio, use the commands `showcpg -space` or `showspace -cpg <cpg>`. To see the virtual domain overprovisioning ratio, use the command `showsys -domainspace`. To see the system overprovisioning ratio, use the command `showsys -space` or use `showsys -space -devtype <devtype>` to restrict the view to a particular device type (SSD, FC, or NL).

Organizations often have a standard for the amount of overprovisioning they want to allow, and this policy can be facilitated by setting an overprovisioning warning or a limit.

Setting the overprovisioning ratio warning

- To set or update the optional overprovisioning ratio warning, issue the `setsys OverprovRatioWarning <value>` command.

The <value> of the OverprovRatioWarning parameter is a ratio. A ratio of 0 (default) means no warning alert is generated. To set a warning threshold, the value must be 1.5 or above. A ratio of 3 means that there is three times the size virtually available than what is physically available.

Setting the overprovisioning ratio limit

- To set or update the optional overprovisioning ratio limit, issue the `setsys OverprovRatioLimit <value>` command.

The <value> of the OverprovRatioLimit parameter is a ratio. A ratio of 0 (default) means no limit is enforced. To set a limit threshold, the value must be 1.5 or above. A ratio of 3 means that there is three times the size virtually available than what is physically available. If a limit is set, the `createvv` and `createsv` commands will return an error when the per-CPG, per-disk-type, or system wide overprovisioning ratio reaches the limit.



NOTE

In systems with many CPGs, setting the overprovisioning limit and warning can slow the performance of CLI commands or steps that use the system manager (sysmgr). I/O is not affected.

TRACKING VOLUME SPACE USAGE

Thin volumes consume user space, admin space, and possibly snapshot space on the disk array. The following sections provide the CLI commands needed to determine how much space of every type is being consumed within the array. The output of these CLI commands shows the reserved and the raw reserved space. The reserved space is what is offered by the array as usable space to the host. This value is also shown in the HPE StoreServ Management Console in the Reserved User Size column for a TPVV and in the pie chart for the logical option in the Summary tab for the virtual volume details screen. The raw reserved space is calculated from the reserved space by multiplying the latter by its RAID overhead factor. For example, this factor has a value of 8/6 for RAID 6 with a set size equal to 8. The HPE StoreServ Management Console shows the raw reserved space in the pie chart for the Raw option in the Summary tab for the virtual volume details. In chargeback models, most IT departments bill their customers on the amount of raw reserved space consumed.

Volumes

Use the `showvv -s` command to see how much user and snapshot space is used by each thin volume. The reserved totals show how much space has been allocated, whereas the used totals show how much of the space is currently in use by the VV. A significant difference between the space in use and the reserved space would indicate that space reclaim has been initiated on the VV, and the reserved space will decrease over time as the space is reclaimed in the background. This is an example of the `showvv -s` output:

```
cli% showvv -s -p -comp No
-----Snp-----Usr-----Total-----
--(MiB)-- -(% VSize)-- --(MiB)---- -(% VSize)-- -----(MiB)----- --Efficiency--
Id Name Prov Compr Dedup Type Rsvd Used Used Wn Lim Rsvd Used Used Wn Lim Rsvd Used HostWr VSize Compact Compress
902 vv2 tppv No No base 0 0 0.0 -- -- 482176 465856 44.4 0 0 482944 465856 -- 1048576 2.25 --
901 vv4 tdvv No Yes base 0 0 0.0 -- -- 107520 97115 9.3 0 0 108544 97115 465871 1048576 10.80 --
-----
2 total 0 0 589696 562971 591488 562971 465871 2097152
```

CPGs

Space in use on the array can be tracked per CPG. The `showcpg -r` command shows the user, snapshot, and free (unallocated) space in Used and Raw Used amounts.

```
cli% showcpg -r CPG?
----Volumes---- -Usage- -----(MiB)-----
Id Name Warn% VVs TPVs TDVs Usr Snp Base RBase Snp RSnp Free RFree Total RTotal
5 CPG1 - 1 1 0 1 1 900608 1080727 135040 162047 8064 9678 1043712 1252453
6 CPG2 - 2 0 1 2 1 352000 422399 2560 3071 54784 65741 409344 491211
-----
2 total 3 2 1252608 1503126 137600 165118 62848 75419 1453056 1743664
```

In addition to showing the CPG usage, the `showspace -cpg` command will also show how much LD space may still be created, given the amount of free chunklets in the system and the CPG parameters (for example, RAID set size, HA level, device types, and so on).

```
cli% showspace -cpg CPG?
------(MiB)-----
CPG -----EstFree----- -----Efficiency-----
Name RawFree LDFree OPFree Base Snp Free Total Compact Dedup Compress DataReduce Overprov
CPG1 10282284 8568576 - 900608 135040 8064 1043712 2.05 - - - 0.33
CPG2 10282284 8568576 - 352000 2560 54784 409344 3.12 1.71 1.95 2.70 0.02
```



System space

Usage reporting and trend analysis

The CLI command `showcpg -hist <CPG>` gives a daily account of CPG usage split into user, snapshot, and free space.

```
cli% showcpg -hist SSD_r6
CPG SSD_r6
```

----Volumes----				-Usage-		-----[MiB]-----				
Time	Warn%	VVs	TPVVs	TDVVs	Usr	Snp	Base	Snp	Free	Total
Jul 18 16:10:12	-	20	0	19	20	15	2315776	7680	11264	2334720
Jul 18 03:37:01	-	20	0	19	20	15	2315776	7680	576512	2899968
Jul 17 03:37:01	-	20	0	19	20	15	2315776	7680	576512	2899968
Jul 16 03:37:01	-	20	0	19	20	15	2315776	7680	576512	2899968
Jul 15 03:37:00	-	20	0	19	20	15	2315776	7680	576512	2899968
Jul 14 03:37:01	-	20	0	19	20	15	2315776	7680	576512	2899968
Jul 13 03:37:01	-	20	0	19	20	15	2315776	7680	576512	2899968
Jul 12 03:37:01	-	20	0	19	20	15	2316160	7680	576128	2899968
Jul 11 03:37:01	-	20	0	19	20	15	2316160	7680	576128	2899968
Jul 10 03:37:00	-	20	1	19	20	15	2760064	7680	9344	2777088
Jul 09 03:37:01	-	20	1	19	20	15	2422144	7680	162944	2592768
Jul 08 03:37:00	-	20	1	19	20	15	2422144	7680	162944	2592768

The command `showspace -cpg <CPG> -hist` also shows this information along with the capacity efficiency ratios.

cli% showspace -cpg SSD_r6 -hist

-----[MiB]-----															
CPG SSD_r6		-----EstFree-----			-ReduceRate/hr-			-----Efficiency-----							
Time	HrsAgo	RawFree	LDFree	OPFree	RawFree	LDFree	Base	Snp	Free	Total	Compact	Dedup	Compress	DataReduce	Overprov
Jul 18 16:13:25	0	10321920	7741440	-	-	-	2315776	7680	11264	2334720	1.65	1.23	1.26	1.45	0.13
Jul 18 03:37:01	12	11018240	8263680	-	55234	41425	2315776	7680	576512	2899968	1.65	1.23	1.26	1.45	0.0
Jul 17 03:37:01	36	11362304	8521728	-	14336	10752	2315776	7680	576512	2899968	1.65	1.23	1.26	1.45	0.0
Jul 16 03:37:01	60	11362304	8521728	-	0	0	2315776	7680	576512	2899968	1.65	1.23	1.26	1.45	0.0
Jul 15 03:37:00	84	11362304	8521728	-	0	0	2315776	7680	576512	2899968	1.65	1.23	1.26	1.45	0.0
Jul 14 03:37:01	108	11362304	8521728	-	0	0	2315776	7680	576512	2899968	1.65	1.23	1.26	1.45	0.0
Jul 13 03:37:01	132	11362304	8521728	-	0	0	2315776	7680	576512	2899968	1.65	1.23	1.26	1.45	0.0
Jul 12 03:37:01	156	11911168	8933376	-	22869	17152	2316160	7680	576128	2899968	1.65	1.23	1.26	1.45	0.0
Jul 11 03:37:01	180	12255232	9191424	-	14336	10752	2316160	7680	576128	2899968	1.65	1.23	1.26	1.45	0.0
Jul 10 03:37:00	204	13017088	9762816	-	31743	23807	2760064	7680	9344	2777088	1.79	1.22	1.27	1.44	0.0
Jul 09 03:37:01	228	13262848	9947136	-	10240	7680	2422144	7680	162944	2592768	2.06	1.22	1.27	1.44	0.0
Jul 08 03:37:00	252	13262848	9947136	-	0	0	2422144	7680	162944	2592768	2.06	1.22	1.27	1.44	0.0



You can also use the `sncpgspace` and `snvvspace` commands to query the internal HPE System Reporter database. Additionally, the optional HPE Primera System Reporter software has the ability to track the CPG and VV usage for comprehensive usage and trend analysis.

The following example shows the output of a `sncpgspace` command for a CPG.

```
cli% sncpgspace -hourly -btsecs -1h CPG2
```

		-----[MB]-----				-[KB/s]-		-----Efficiency-----					
Time	Secs	PrivateBase	PrivateSnap	Shared	Free	Total	UsableFree	Dedup_GC	Compact	Dedup	Compress	DataReduce	OverProv
2019-07-18 16:00:00 BST	1563462000	212224	2560	139776	54784	409344	8568570	0.0	3.12	1.71	1.95	2.7	0.02

MONITORING AND ALERTING

Allocation warnings

Allocation warnings provide a mechanism for informing storage administrators when a specific capacity threshold is reached. An allocation warning can be specified independently for each VV and each CPG. It is recommended that allocation warnings be used, at least on the CPG level, and acted upon when they are triggered.

The relevant CLI commands for setting allocation and growth warnings are:

`setvv -usr_aw <percent> <VV>`: Sets the allocation warning for the user space of the VV as a percentage of the VV size

`setvv -snap_aw <percent> <VV>`: Sets the allocation warning for the snapshot space of the VV as a percentage of the VV size

`setcpg -sdgw <num> <CPG>`: Sets the growth warning for the CPG in MiB (append to the value num g or G for GiB or t or T for TiB)

These warnings can be changed at any time and are effective immediately. The CLI commands `showvv -alert` and `showcpg -alert` lists the allocation warnings that were set per VV and CPG.

The VV allocation limits and warnings can be set with the HPE StoreServ Management Console by selecting **Advanced options** checkbox when creating or editing a VV as shown in Figure 5.

Create Virtual Volume General ▾ ?

General ☒ Advanced options

Name:

System:

Domain:

Policy:

Provisioning:

Dedup and Compression:

CPG:

RAID 6 SSD

Size:

App Volume set:

Allocation warning: ☒ Enabled %

Allocation limit:

FIGURE 5. VV allocation limit and warning options



The CPG growth limits and warnings can be set with the HPE StoreServ Management Console by selecting **Advanced options** checkbox when creating or editing a CPG. This will display the growth options as shown in Figure 6.

Growth

Growth increment

8

GiB

Growth limit

Disabled

Growth warning

Enabled

5

TiB

FIGURE 6. CPG growth limit and warning options

It is important to note that the growth limit for a CPG is a hard limit and the CPG will not grow beyond it. Once the CPG hard limit has been reached, any VVs that require more space will not be able to grow. This will result in write errors to host systems until the CPG allocation limit is raised. Therefore, it is recommended that VV, CPG and free space warnings and limits are set to sensible levels and managed when they are triggered. As an example, the CPG warning limit should be set sufficiently below the CPG allocation limit so that it alerts the storage administrator with ample time to react before the CPG allocation limit is reached.

Remaining physical capacity alerts

As available, physical capacity across the HPE Primera storage gets consumed by VVs, preconfigured alerts are generated at 50%, 75%, 85%, and 95% of physical capacity in use per drive type (FC, NL, or SSD). Furthermore, the storage administrator can use the CLI command `setsys RawSpaceAlertSSD <value>` where value is the remaining capacity on SSD disks in GiB

These serve as array-wide, advance warnings to the storage administrator to plan for and add necessary physical capacity. The alerts generated should be monitored and promptly acted upon to prevent all free space of a particular drive type from being consumed.

The system limits and warnings can be set with the HPE StoreServ Management Console by selecting the **Edit** action when viewing a storage system as shown in Figure 7.

System Parameters

SSD raw space alert

Enabled

10000

GB

Overprovisioning limit

Disabled

Overprovisioning warning

Disabled

Max volume retention

Enabled

14

Days

FIGURE 7. System limit and warning options



Remaining CPG free space alerts

HPE Primera OS samples the space available to CPGs once per day. The history of used and free CPG space is stored in an internal table and can be displayed using the `-hist` option in the `showspace` and `showcpg` commands. An alert is automatically generated if the available free space for a CPG falls below the CPG warning limit or the CPG allocation limit.

View logged warning and limit alerts

All warning and limit alerts previously mentioned can be viewed in several ways:

- The CLI commands `showalert` and `showeventlog` list the alerts in various formats and with various options.
- The HPE StoreServ Management Console shows the alerts in the Events section.
- Storage Management Initiative Specification (SMI-S) software integrated in HPE Primera OS provides asynchronous notification of events for changes in the elements managed by the Common Information Model (CIM) server. A CIM client can subscribe to selected CIM indications to receive event notifications from the CIM server.
- The SNMP agent within HPE Primera OS allows for retrieving the alerts by remote SNMP clients.

Alerts can be forwarded (`setsys RemoteSyslogHost`) to a log host for viewing them in an enterprise management application.

User recommendations

The monitoring of alerts for available capacity by storage administrators and internal business processes are a critical component of a successful HPE Primera Thin Provisioning management and administration strategy. You should nominate a primary and if possible a backup storage administrator for each site with HPE Primera equipment. The storage administrator's roles include:

- Proactively monitor free space availability per VV and CPG
- Proactively monitor consumption rates for VVs and CPGs
- Proactively monitor consumed VV capacity and compare to licensed thin-provisioning capacity
- Proactively monitor physical capacity thresholds for each disk type and for the entire array
- Ensure adequate purchasing and installation of additional physical disk capacity buffer upgrades in a timely manner
- Nominate an escalation contact who has proper authority to drive the customer responsibilities outlined in this document if the nominated storage administrators fail to carry out their responsibilities

If you have a network connection with HPE InfoSight via the HPE Primera UI, the health of the HPE Primera can be proactively monitored for CPG growth problems. You can request to receive thin provisioning and other alerts by mail or via phone. You retain responsibility for managing the thin-provisioning capacity and CPGs; Hewlett Packard Enterprise is not responsible for any failure when thresholds are met or exceeded.

MIGRATE BETWEEN VOLUME TYPES ON THE SAME ARRAY

All HPE Primera systems have the ability to make a conversion between volume types without requiring an offline transition. This allows the storage administrator to select which data reduction technologies are enabled on a per-volume basis for complete flexibility. They can choose the optimal volume type to match the application characteristics or performance requirements.

The SSMC or the `tunevv` command can be used to perform the conversion between the volume types. Table 3 shows the command line options for tuning to the different volume types.

The conversion process does not support virtual volumes within Remote Copy groups. A virtual volume using snapshots can only be converted if the `-keepvv` option is used, but the snapshots will be associated with the virtual volume specified by the `-keepvv` option. The conversion will automatically roll back on a failure.

TABLE 3. Command line options for tuning to the different volume types

Volume type	Command line
Thin provisioned	<code>tunevv usr_cpg <CPG> -tpvv <VV Name></code>
Data reduction	<code>tunevv usr_cpg <CPG> -reduce <VV Name></code>



HPE PRIMERA THIN PERSISTENCE SOFTWARE

Traditionally, when data is deleted on a host, the OS will report that space has been freed, but the storage will not be informed that the data is no longer in use. With thinly allocated volumes, the unused space will remain allocated on the array causing the volumes to grow over time. This creates a hidden utilization penalty that can significantly reduce the space savings of data reduction. On systems that overprovision the virtual to physical capacity, if this unused space is not claimed, there is risk of running out of space on the system.

HPE Primera Thin Persistence software is able to maintain the benefits of HPE Primera Data Reduction by enabling HPE Primera Thin Reclamation of allocated but unused capacity so the thin volumes are able to stay as lean and efficient as possible.

It leverages the HPE Primera OS support for the WRITE_SAME and UNMAP commands of the T10 SCSI Primary Commands - 3 (SPC-3) standard and the unique zero-detection capabilities of the HPE Primera ASIC. These give HPE Primera storage the power to reclaim unused space associated with deleted data simply, quickly, and nondisruptively.

UNMAP is a SCSI command that a host can issue to tell the storage that blocks are no longer needed to be allocated. This is particularly important in thin-provisioned environments, as it allows the storage array to recognize that these blocks are not used and to return them to the free capacity pool for reuse by other volumes.

The HPE Primera ASIC features an efficient, silicon-based zero-detection mechanism. This unique hardware capability gives HPE Primera Storage the ability to remove allocated but unused space as small as 16 KiB on the fly without impacting performance.

Also, the benefits of HPE Primera Thin Persistence are available to read/write snapshots of VVs. The mechanism for initiating reclamation is the same as for the parent TPVV: writing zeros to the allocated but unused space in a read/write snapshot will trigger the ASIC to initiate reclamation of the deleted space. To benefit from thin reclamation, the zero-detect policy needs to be enabled on each read/write snapshot.

With data reduction VVs, there are internal differences in the way zero detection operates, as the zero blocks will be reduced to a single zero block by the deduplication engine.

HPE Primera Thin Persistence reclamation

HPE Primera Thin Persistence reclamation occurs at several levels. Initially, all freed 16 KiB pages are returned to the VV. This means that on a file system that supports automatic reclaim, the space freed by an UNMAP after a file deletion is immediately available for reuse by a file creation or file extension operation on the same file system.

To make space available for reuse by other volumes, there is a reclaim thread that returns freed 128 MiB regions allocated to a VV back to the CPG. This thread scans volumes every five minutes for space that potentially can be reclaimed. If a VV has free 128 MiB regions or there is enough free space to warrant a defragmentation of the VV, then space will be reclaimed back to the CPG. Defragmentation occurs if there is more than 1 GiB of available space to reclaim and results in free 128 MiB regions. Up to 16 volumes at a time can be queued for reclaim processing.

How quickly the space is reclaimed depends on a number of factors. If there is a large amount of freed space on a volume, then this may not be processed within a single reclaim period. Once the reclaim process runs on a VV, the reclaim process will not run again on that VV again for at least 90 minutes. Therefore, large space reclaims can take several hours to complete.

Also, the reclamation on a VV can be deferred for various reasons. For example, if the space of a VV is grown, then reclaims on the volume will be deferred for 60 minutes. Also, if reclaim is defragmenting a VV and the defragmentation does not complete during the reclaim interval, reclaim will be deferred further for four hours.

HPE Thin Persistence reclamation may not reclaim all the free space on a volume. There is a 4 GiB per node threshold below which the VV will not be inspected for available 128 MiB regions that can be reclaimed back to the CPG. The free space will still be available for reuse by the VV.

Those new to thin provisioning often like to verify HPE Thin Persistence reclamation by creating a test scenario of filling a file system then deleting the files and running a space reclamation tool. It is important to understand that the space will not be returned to the CPG immediately. The `showvv -s` command will show how much space has been allocated to the VV and the difference between the space in use and the reserved space shows the amount of space reclaimed for use within the VV. The amount of reserved space will decrease over time as the space is reclaimed back to the CPG in the background by the reclaim thread.



HPE Primera Thin Persistence methods

The most optimal HPE Primera Thin Persistence method is for the host OS to issue SCSI UNMAP commands for unwanted data blocks. Typically, this would be done by the file system when files are deleted. However, if the OS offers a suitable UNMAP application programming interface (API), an application can directly communicate to the HPE Primera system that some data is no longer needed. This method provides continual reclaiming of space to allow the storage volumes to stay thin. The disadvantage is it requires significant changes to the storage stack and only the most modern OSs have implemented native UNMAP support.

Another method is to have an OS utility that will examine the blocks of a volume and issue UNMAPs for those that are not being used. This type of utility also requires host OS UNMAP support but to a lesser extent the continual method, and they are specific to a file system or volume manager type. Most of these utilities can be run when the data is online but as they generate the UNMAP requests for all the unused blocks in one go, they are:

1. Generally run manually during an outage window
2. Scheduled to run during a quiet period so the reclaims do not adversely impact other workloads on the storage

The final method is to reclaim space using a zerofile utility that writes zeros to all allocated but unused space on a file system. On HPE Primera systems, zero detection intercepts the blocks of zeros being written and automatically triggers the reclamation of the space. The advantage of this method is that it does not require any special OS support, and the utilities to generate zerofiles are often supplied with the base OS distribution. To achieve good reclaim rates, the utilities need to fill the majority of the available free space so they are:

1. Generally run manually during an outage window
2. Scheduled to run during a quiet period to avoid applications failing due to a lack of file system space

For the zerofile utilities to work, the zero-detect policy needs to be set for each VV. Blocks of 16 KiB of contiguous zeros are freed and returned for reuse by the VV. If 128 MiB of space is freed, it is returned to the CPG for use by other volumes.

For more information on HPE Primera Thin Persistence methods for various OSs, see Appendix A.

APPENDIX A—HPE PRIMERA THIN PERSISTENCE METHODS

HPE Primera Thin Reclamation for Microsoft Windows Server® 2012

Windows Server 2012 integrates very well with thin-provisioned volumes on HPE Primera. It identifies thin-provisioned volumes on HPE Primera systems, writes entries in the Windows® Event Log file when storage thresholds are reached on the CPG and the TPVV level, and supports active reclaim. This is done by issuing UNMAPs upon file deletion or file shrinking on thin-provisioned volumes on NTFS formatted volumes. The standard defragmentation scheduled task also automatically reclaims storage.

Also, Windows Server 2012 extends UNMAP support to the virtual layer. Hyper-V VHDX disks report themselves to the guest OSs as being **thin-provision capable**. This means that if the guest OS is UNMAP-capable, it can send UNMAPs to the VHDX file, which will then be used to help ensure that block allocations within the VHDX file are freed up for subsequent allocations as well as forwarding the UNMAP requests to the physical storage.

There is also a File TRIM API, which is mapped to the TRIM command for ATA devices and the UNMAP command for SCSI devices. TRIM hints allow the application to notify the storage that blocks that, previously were allocated, are no longer needed and can be reclaimed.

In summary, the following operations trigger storage space reclamation in Windows Server 2012:

- Deletion of a file from a file system on a thin-provisioned volume
- Running storage optimization,¹ a new feature of Windows Server 2012 disk management
 - You can use manual or automatic scheduling of the Optimize operation utility.
 - The standard Defrag scheduled task automatically runs Optimize.
- UNMAP requests from a Hyper-V guest OS
 - Deleting a file from the file system of an UNMAP-capable guest OS sends UNMAP requests to the driver stack of the Hyper-V host
- UNMAP requests from applications using the TRIM API

¹ Optimization is only available on NTFS volumes with allocation units of 16 KiB or less.



The default behavior of issuing UNMAPs on file deletion can be disabled on a Windows Server 2012 by setting the `DisableDeleteNotify` parameter of the `fsutil` command. This will prevent reclaim operations from being issued against all volumes on the server.

To disable reclaim operations, run the following PowerShell command:

```
fsutil behavior set DisableDeleteNotify 1
```

HPE Primera Thin Reclamation for Microsoft Windows Server 2003 and 2008

Windows Server versions prior to 2012 do not implement UNMAP support and therefore to reclaim thin-provisioned storage, you must leverage the zero-detection capabilities of HPE Primera.

Microsoft provides a Sysinternals advanced system utilities suite that includes the Secure Delete (SDelete) application that can be used to overwrite deleted files on disk data to make disk data unrecoverable.

As well as overwriting a single file data space, SDelete can indirectly overwrite free space by allocating the largest file it can. Then, it performs a secure overwrite to ensure that all the disk space that was previously free becomes securely cleansed. You can utilize this feature of SDelete to perform HPE Primera Thin Reclamation on zero-detect enabled HPE Primera volumes by specifying the `-z` flag when running SDelete to write zeros to the free space.

One disadvantage of SDelete is that it does not support mount points so a volume must have a drive letter associated with it before the utility can be run. Using the `subst` command, one can temporarily attach a drive letter to a mount point before running SDelete.

It is recommended that applications are shut down before running SDelete, as it can cause a `file_system_full` condition due to consuming all free space. An alternative solution is to create a PowerShell script that uses `fsutil` to create a balloon file that is limited to a certain percentage of the free space.

This is an example of using `fsutil` to zero 100 MiB of space:

```
fsutil file createnew zerotemp.txt 104857600
fsutil file seVValiddata zerotemp.txt 104857600
fsutil file setzerodata offset=0 length=104857600 zerotemp.txt del zerotemp.txt
```

HPE Primera Thin Reclamation for VMware vSphere

Raw device mapping

Raw device mapping (RDM) is used to map a storage LUN directly to a VM bypassing VMFS layer. This LUN can be used by the VM without the need to format it using VMFS and placing VMDK file on top of it. All SCSI commands from the VM are passed to RDM LUN directly except the REPORT LUNS command, which is virtualized by VMkernel. The LUN will be seen by the guest OS as a SCSI version 6 (SPC-4) device as demonstrated in the following example from a Linux® VM:

```
# sg_inq /dev/sdb -d
standard INQUIRY:
PQual=0 Device_type=0 RMB=0 LU_CONG=0 version=0x06 [SPC-4]
[AERC=0] [TrmTsk=0] NormACA=0 HiSUP=1 Resp_data_format=2
SCCS=0 ACC=0 TPGS=1 3PC=1 Protect=0 [BQue=0]
EncServ=0 MultiP=1 [VS=0] [MChngr=0] [ACKREQQ=0] Addr16=0
[RelAdr=0] WBus16=1 Sync=1 [Linked=0] [TranDis=0] CmdQue=1
[SPI: Clocking=0x0 QAS=0 IUS=0]
length=172 [0xac] Peripheral device type: disk
Vendor identification: Primeradata
Product identification: VV
Product revision level: 3310
Unit serial number: 4UW0001439
```



UNMAPs can be issued to the RDM LUN if the guest OS supports them. See the information in the HPE Primera Thin Persistence method section for the guest OS type and version. The following example from a Linux VM using the `sg_vpd` command shows the UNMAP support flags for a compressed volume presented as an RDM:

```
# sg_vpd /dev/sdb -p lbpv
Logical block provisioning VPD page [SBC]:
  Unmap command supported (LBPU): 1
  Write same (16) with unmap bit supported (LBWS): 1
  Write same (10) with unmap bit supported (LBWS10): 1
  Logical block provisioning read zeros (LBPRZ): 1
  Anchored LBAs supported (ANC_SUP): 0
  Threshold exponent: 19
  Descriptor present (DP): 0
  Provisioning type: 2
```

vSphere 6.0

In vSphere 6.0, support was introduced for passing guest OS UNMAPs through to the storage for VVol and VMFS backed VMs. For VVols, no extra configuration is required, but for VMFS datastores, the `EnableBlockDelete` parameter must be set and the VMDKs must be thin-provisioned. You can view the value of `EnableBlockDelete` using the following `esxcli` command:

```
# esxcli system settings advanced list --option /VMFS3/EnableBlockDelete
Path: /VMFS3/EnableBlockDelete
Type: integer
Int Value: 0
Default Int Value: 0
Min Value: 0
Max Value: 1
String Value:
Default String Value:
Valid Characters:
Description: Enable VMFS block delete when UNMAP is issued from guest OS
```

To enable guest UNMAPs, use the following `esxcli` command:

```
# esxcli system settings advanced set --int-value 1 --option /VMFS3/EnableBlockDelete
```

However, support for guest UNMAPs depends on the type of the guest OS. The LUN will be seen by the guest OS as a SCSI version 2 (SCSI-2) device and SCSI-2 does not have support for the UNMAP command. This means that the only guest OSs that support UNMAP are Windows 2008 and above. For example, Windows Server 2012 VMs that run on vSphere 6.0 can generate UNMAP commands. Other guest OSs, such as Linux, will not generate UNMAP commands because they expect a version 5 (SPC-3) or higher standard.

To support UNMAP automatically, the VMs must be of hardware version 11 or above.

vSphere 6.5

In vSphere 6.5 and above, automatic space reclaim for files deleted from VMFS datastores is available. The system will send UNMAP commands asynchronously in the background and the speed of reclaim is based on priority levels set on individual VMFS datastores.

However, the automatic reclaim is only available on VMFS 6 datastores. VMFS 5 datastores must still use the manual method using the `esxcli` command to reclaim space with the balloon file method. When you create a VMFS, you will have the option to select between VMFS 5 and VMFS 6 as shown in Figure 8.



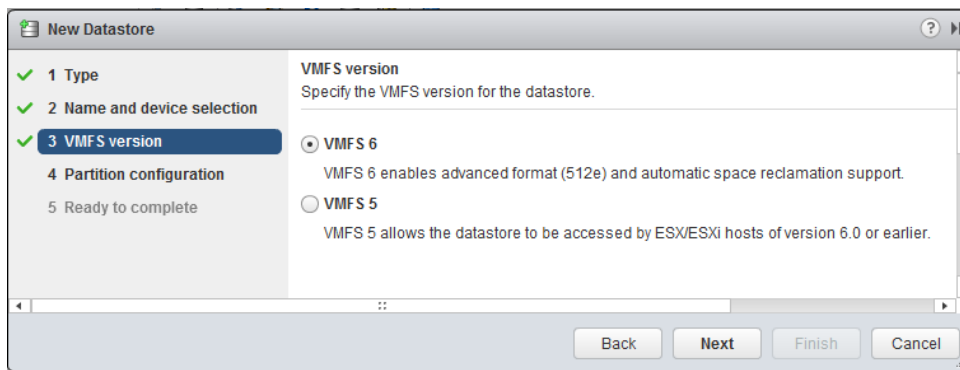


FIGURE 8. VMFS version selection in vSphere 6.5

The reclaim priority on VMFS 6 datastore will default to Low. In the VMware vSphere® Web Client, the priority can be set to either None or Low with None disabling UNMAP completely. However, using the `esxcli` command (`esxcli storage vmfs reclaim config`) you can also change this setting to Medium or High that increases the frequency in which UNMAP commands are sent by 2X (Medium) and 3X (High) over the Low setting.

Note that a higher reclaim priority may generate heavier loads on the storage array, which could negatively impact the normal VM workloads. It is therefore recommended to leave the priority as the default Low value.

vSphere 6.5 also introduced a virtual disk format 2.0 with SCSI version 6 (SPC-4) support for both VMDK and VVol-based disks. This means that all guest OSs that support UNMAP can issue them to the devices and they will be passed through to the storage.

The LUN is seen by the guest OS as a SCSI version 6 (SPC-4) device as demonstrated in the following example from a Linux VM:

```
# sg_inq /dev/sda -d
standard INQUIRY:
    PQual=0 Device_type=0 RMB=0 LU_CONG=0 version=0x06 [SPC-4]
    [AERC=0] [TrmTsk=0] NormACA=0 HiSUP=0 Resp_data_format=2
    SCCS=0 ACC=0 TPGS=0 3PC=0 Protect=0 [BQue=0]
    EncServ=0 MultiP=0 [MChngr=0] [ACKREQQ=0] Addr16=0
    [RelAdr=0] WBus16=1 Sync=1 [Linked=0] [TranDis=0] CmdQue=1
    length=36 [0x24] Peripheral device type: disk
Vendor identification: VMware
Product identification: Virtual disk
Product revision level: 2.0

No version descriptors available
```

The following example from a Linux VM using the `sg_vpd` command shows the UNMAP support flags for a VMFS VMDK volume:

```
# sg_vpd /dev/sda -p lbpv
Logical block provisioning VPD page [SBC]:
    Unmap command supported [LBPU]: 1
    Write same [16] with unmap bit supported [LBWS]: 0
    Write same [10] with unmap bit supported [LBWS10]: 0
    Logical block provisioning read zeros [LBPRZ]: 1
    Anchored LBAs supported [ANC_SUP]: 0
    Threshold exponent: 1
    Descriptor present [DP]: 0
    Provisioning type: 2
```



To support UNMAP automatically from virtual disk 2.0 devices, the VMs must be of hardware version 13 or above, the datastores must be VMFS 6 format or VMFS 5 format with EnableBlockDelete set, and the VMDKs must be thin provisioned.

HPE Primera Thin Reclamation for Linux

Linux uses the term discard for the act of informing a storage device that blocks are no longer in use. This is because it uses the same mechanism to support both TRIM commands on ATA SSDs and UNMAP commands on SCSI devices, so discard was chosen as a protocol neutral name. Red Hat® Enterprise Linux (RHEL) 6 was the first major distribution to support discards and offers both real-time and batched support.

Real-time support is offered by an ext4 or XFS file system when mounted with the `-o discard` option; the default is not to issue discards. When data or files are deleted on a discard enabled ext4 file system, UNMAPs are generated to free up space on the thin-provisioned virtual volume on the storage. The LVM and the device mapper (DM) targets also support discards so space reclaim will also work on file systems created on LVM and/or DM volumes.

For example, to mount the DM device `tpvv_lun` on `/mnt` with discards enabled, run:

```
# mount -t ext4 -o discard /dev/mapper/tpvv_lun /mnt
```

This will cause the RHEL 6 to issue the UNMAP command, which in turn causes space to be released back to the array from the TPVV volumes for any deletions in that ext4 file system. This is not applicable for full-provisioned virtual volumes.

Also, the `mke2fs`, `e2fsck`, and `resize2fs` utilities also support discards to help ensure the TPVV volumes are enhanced when administration tasks are performed.

There is also batched discard support available using the `fstrim` command. This can be used on a mounted file system to discard blocks, which are not in use by the file system. It supports ext3 and ext4 file systems and can also rethin a file system not mounted with the discard option.

For example, to initiate storage reclaim on the `/mnt` file system, run:

```
# fstrim -v /mnt
```

```
/mnt: 21567070208 bytes were trimmed
```

The `fstrim` can be run when the file system is online. However, as it generates UNMAP requests for all the unused blocks in one go, consideration should be given to running it during a quiet period so the reclaims do not adversely impact other workloads on the storage.

It is also possible to configure the Logical Volume Manager (LVM) to automatically issue discards to a logical volume's underlying physical volumes when the logical volume is no longer using the physical volumes' space (for example, by using the `lvremove` or `lvreduce` commands). This can be done by enabling the option `issue_discards` in the LVM configuration file:

```
$ cat /etc/lvm/lvm.conf
```

```
devices {  
    issue_discards = 1  
}
```

Then, when removing a volume, you would see the following message:

```
# lvremove vg1/lvol0
```

```
Do you really want to remove and DISCARD logical volume lvol0? [y/n]:
```

The Linux swap code will also automatically issue discard commands for unused blocks on discard-enabled devices and there is no option to control this behavior.



HPE Primera Thin Reclamation for HP-UX

HP-UX systems using Veritas Storage Foundation 5.1 or higher can reclaim space associated with file deletions or file shrinking (see the [HPE Primera Thin Reclamation for Symantec Storage Foundation](#) section for more details). However, non-VxFS file systems or VxFS file systems on LVM volumes will need to reclaim space using a **zerofile** script that writes zeros to all allocated but unused space on a file system. The zero-detection capability of the HPE Primera ASIC will intercept the blocks of zeros being written and automatically trigger the reclamation of the space.

HPE Primera Thin Reclamation for UNIX®

On UNIX systems or Linux distributions that do not support discard, you will need to reclaim space using a **zerofile** script that writes zeros to all allocated but unused space on a file system. The zero-detection capability of the HPE Primera ASIC will intercept the blocks of zeros being written and automatically trigger the reclamation of the space.

The script would use the `dd` command to copy zero data blocks from the `/dev/zero` device to a file in the file system. However, it is recommended that the size of the space to zero is not more than 70% of the free space as a very large zerofile could cause the creation of new files to fail while the zerofile script is running.

HPE Primera Thin Reclamation for Symantec Storage Foundation

Since the introduction of Symantec Storage Foundation 5.1, hosts with VxFS file systems managed by the Veritas Volume Manager (VxVM) software can also reclaim space associated with file deletions or file shrinking on HPE Primera storage.

The space reclamation is not automatic. The VxFS file system informs the VxVM about deleted blocks and a VxVM command has to be manually run that sends WRITE SAME SCSI commands with the UNMAP bit turned on to the HPE Primera. No tool to write zeros to the deleted space in the file systems is required for reclaiming the space.

The list of disks, whose allocated but unused space can be reclaimed, is given by the command executed on the host:

```
# vxdisk -o thin,fssize -u m list
```

This will display the VV allocated space and the file system usage. The space reclamation is initiated by the VxVM `vxdisk` command:

```
# vxdisk reclaim [<disk>|<dg>|<encl>]
```

By default, the reclamation does not affect unmarked space, which is the unused space between subdisks. If a LUN has a lot of physical space that was previously allocated, the space between the subdisks might be substantial. Use the `-o full` option to reclaim the unmarked space.

```
# /opt/VRTS/bin/fsadm -V vxfs -R /<VxFS_mount_point>
```

To monitor the reclamation status, run the following command:

```
# vxtask list
```

```
TASKID PTID TYPE/STATE PCT    PROGRESS
171    RECLAIM/R      00.00% 0/41875931136/0 RECLAIM vol100 dg100
```

The `vxrelocd` daemon tracks the disks that require reclamation. The schedule for reclamation can be controlled using the `vxdefault` command. The `reclaim_on_delete_wait_period` parameter specifies the number of days after a volume or plex is deleted when VxVM reclaims the storage space. The default value is 1, which means the volume is deleted the next day. A value of -1 indicates that the storage is reclaimed immediately and a value of 367 indicates that the storage space can only be reclaimed manually using the `vxdisk reclaim` command. The `reclaim_on_delete_start_time` parameter specifies the time of day when VxVM starts the reclamation for deleted volumes and this defaults to 22:10.

To completely disable thin-reclaim operations, add the parameter `reclaim=off` to the `/etc/vxdefault/vxdisk` file.



HPE Primera Thin Reclamation for Oracle Databases

During an Oracle Automatic Storage Management (ASM) database lifecycle, the utilization of allocated storage capacity in a thin-provisioned volume can decrease, as changes are made to the database through common operations such as:

- Dropping of a tablespace or database upon deletion of transient data
- Resizing of an Oracle datafile upon shrinking a tablespace
- Addition of new disks to an ASM disk group to accommodate growth or load balance performance

These changes result in the creation of unused ASM disk space that can build up over time and although this space is available for reuse within ASM, it remains allocated on the storage array. The net result is that the storage utilization on the array eventually falls below desirable levels.

To solve this problem, HPE and Oracle have partnered to improve storage efficiency for Oracle Database 10g and 11g environments by reclaiming unused (but allocated) ASM disk space in thin-provisioned environments. The Oracle ASRU is a stand-alone utility that works with HPE Primera Thin Persistence software to reclaim storage in an ASM disk group that was previously allocated but is no longer in use. Oracle ASRU compacts the ASM disks, writes blocks of zeroes to the free space, and resizes the ASM disks to original size with a single command, online and nondisruptively. The HPE Primera, using the zero-detect capability of the HPE Primera ASIC, will detect these zero blocks and reclaim any corresponding physical storage.

You can issue a SQL query to verify that ASM has free space available that can be reclaimed:

```
SQL> select name, state, type, total_mb, free_mb from vasm_diskgroup where name = 'LDATA';
```

NAME	STATE	TYPE	TOTAL_MB	FREE_MB
-----	-----	-----	-----	-----
LDATA	MOUNTED	EXTERN	1023984	610948

Run the Oracle ASRU utility as the Oracle user with the name of the disk group for which space should be reclaimed:

```
# bash ASRU LDATA
```

```
Checking the system ...done
```

```
Calculating the new sizes of the disks ...done Writing the data to a file ...done
```

```
Resizing the disks...done
```

```
/u03/app/oracle/product/11.2.0/grid/perl/bin/perl -I /u03/app/oracle/product/
```

```
11.2.0/grid/perl/lib/5.10.0 /home/ora/zerofill 5 /dev/oracleasm/disks/LDATA2
```

```
129081 255996 /dev/oracleasm/disks/LDATA3 129070 255996 /dev/oracleasm/disks/
```

```
LDATA4 129081 255996 /dev/oracleasm/disks/LDATA1 129068 255996
```

```
126928+0 records in
```

```
126928+0 records out
```

```
133093654528 bytes (133 GB) copied, 2436.45 seconds, 54.6 MB/s
```

```
126915+0 records in
```

```
126915+0 records out
```

```
133080023040 bytes (133 GB) copied, 2511.25 seconds, 53.0 MB/s
```

```
126926+0 records in
```

```
126926+0 records out
```

```
133091557376 bytes (133 GB) copied, 2514.57 seconds, 52.9 MB/s
```



```
126915+0 records in
126915+0 records out
133080023040 bytes [133 GB] copied, 2524.14 seconds, 52.7 MB/s
Calculating the new sizes of the disks ...done
Resizing the disks...done
Dropping the file ...done
```

APPENDIX B—FILE SYSTEMS AND DEDUPLICATION

File systems with lots of common files such as home directory shares are ideal candidates for deduplication. However, the space savings achieved can vary greatly depending on the file system in use. HPE Primera Deduplication uses a granularity, also called chunk size, of 16 KiB and therefore, optimum deduplication is achieved when the file system aligns files with this granularity.

There are many different file system choices available across the supported OSs and they work in very different ways with deduplication. Some file systems work perfectly, some need online tuning, some need reformatting with different options, and some cannot be changed. In general, older file system technologies designed to work with discrete drives do not perform as well as modern log structured file systems.

It is often thought that deduplication ratio is dependent on the file system block size alone, but it actually depends on block size, extent size, and alignment, all of which vary from file system to file system. Block size is the basic unit of storage for the file system and it is usually, but not always, page size of the processor architecture. Extent size is the number of contiguous blocks that can be written. Alignment is the starting offset of a file; most file systems align to block size boundaries.

For normal volumes, changing the block size or alignment can increase the space usage for very small files (although it may improve performance for larger files). However, by aligning the files, the increase in deduplication should more than offset any space increase for small files, so the overall space usage will reduce.

Note that with virtualized systems, the file systems inside the VM, which may be under the control of a tenant, should be aligned as well as the file system of the hypervisor.

Microsoft Windows

New Technology File System

New Technology File System (NTFS) is the default file system for Microsoft Windows and by default has an allocation unit of 4 KiB. For optimal deduplication, consider setting the allocation unit to 16 KiB or a multiple of 16 KiB when the file system is created. Do this by selecting the desired allocation unit in the format dialog box when creating the file system.²

Using a 16 KiB allocation unit size not only ensures that all files start on a 16 KiB boundary, relative to the offset of the partition, but also any UNMAP operations will also be aligned to 16 KiB boundaries.

To determine the allocation unit size of an existing NTFS file system, run the `fsutil fsinfo ntfsinfo` command and check the Bytes Per Cluster value. The following example shows a file system with a 16 KiB allocation unit.

```
C:\>fsutil fsinfo ntfsinfo f:

NTFS Volume Serial Number :      0x004eab4c4eab38f4

Version :                        3.1

Number Sectors :                  0x0000000006f997ff

Total Clusters :                  0x00000000001be65f

Free Clusters :                   0x00000000000317a5

Total Reserved :                  0x0000000000000000

Bytes Per Sector :                512
```

² For Microsoft Windows PowerShell Format-Volume cmdlet syntax, see technet.microsoft.com/en-us/itpro/powershell/windows/storage/format-volume.



```

Bytes Per Physical Sector :      512
Bytes Per Cluster :            16384
Bytes Per FileRecord Segment : 1024
Clusters Per FileRecord Segment : 0
Mft Valid Data Length :        0x00000000000040000
Mft Start Lcn :                0x00000000000018000
Mft2 Start Lcn :               0x00000000000000001
Mft Zone Start :               0x00000000000018000
Mft Zone End :                 0x00000000000019920
RM Identifier:                  D252D40C-592D-11E4-803D-A0E201E4931F

```

The following example shows the effect of allocation unit on dedup ratios. Four deduplication volume were created and then formatted with NTFS file systems using allocation units of 4 KiB–32 KiB. Four copies of a compressed archive were put into each file system.

```
cli% showcpg -space NTFS*
```

```

      -Private[MiB]-  -----[MiB]-----  Efficiency  -----
Id Name   Warn%   Base   Snp Shared  Free Total Compact Dedup Compress DataReduce Overprov
 7 NTFS4K   -    4608    0   4608 11008 20224   17.54  1.29    1.26    1.32    0.00
 8 NTFS8K   -    4608    0   4608 11008 20224   17.54  1.99    1.35    2.02    0.00
 9 NTFS16K  -    4608    0   4608 11008 20224   17.54  3.94    1.26    3.94    0.00
10 NTFS32K  -    4608    0   4608 11008 20224   17.52  3.96    2.26    3.98    0.00
-----
 4 total      18432    0  18432 44032 80896   17.54  3.29    2.08    3.32    0.00

```

It is clear that the file system with allocation units of 16 KiB and above have significantly higher dedup ratios.

Setting boot disk allocation units

While changing the NTFS allocation unit on a file system requires a reformat, it is relatively easy to do for nonsystem disks. The Windows OS unfortunately does not give an option to set the allocation unit during installation, so the system disk will have 4 KiB by default. Although there is little dedupable data within a single Windows OS instance, having a 4 KiB allocation unit reduces the deduplication achievable between multiple Windows Servers or VMs. Therefore, it is still desirable to set 16 KiB allocation units for boot disks.

The Windows installation does not give an option to change the allocation unit on Windows boot disks and Microsoft does not have a documented procedure. The challenge is that Windows boots from a hidden System Partition, which must have a 4 KiB allocation unit. A solution is at the start of the installation process is to create a small (at least 500 MB) System Partition formatted with a 4 KiB allocation unit and then the main Windows partition formatted with a 16 KiB allocation unit.

The following is an example of the procedure:

1. Start Windows installation
2. Press Shift + F10 to open CMD
 - a. diskpart
 - b. select disk 0
 - c. create partition primary size=1000
 - d. format quick
 - e. create partition primary
 - f. format quick unit=16K



3. Close CMD
4. Install Windows

This method has been tested on Windows Server 2012.

If the Windows System Partition is not formatted with a 4 KiB allocation unit, the message shown in Figure 9 will be seen during installation.

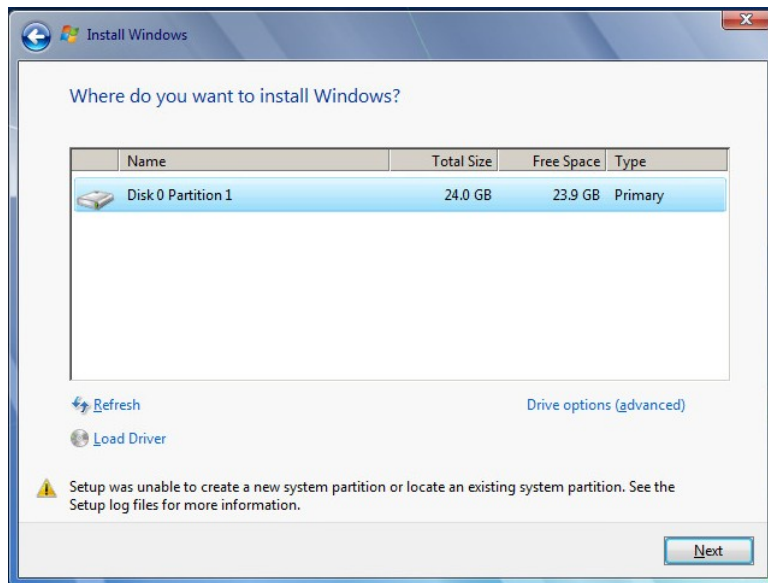


FIGURE 9. System Partition error

Resilient File System

Resilient File System (ReFS) is a new file system introduced with Windows Server 2012. It was designed to have higher data integrity and scalability than NTFS. Internally, ReFS allocates data on disk in clusters of 64 KiB and metadata in clusters of 16 KiB. Therefore, no tuning is necessary for use with HPE Primera Deduplication.

Formatting a volume with ReFS can be done from the UI by choosing ReFS as the file system in the drop-down of the Format dialog box or from the command line using the following syntax:

Command line

```
Format /fs:ReFS /q J:
```

PowerShell equivalent

```
Format-Volume -DriveLetter J -FileSystem ReFS
```

Microsoft Hyper-V

Windows Server 2012 brought many new virtualization improvements, including the introduction of the VHDX file format. In Windows Server 2008, Hyper-V used a virtual hard disk (VHD) file format that had a 2 TiB limit. The new VHDX file format has a maximum capacity 64 TiB. The advantages of VHDX aren't limited to improved capacity, it has a 4 KiB logical sector size that improves performance and can increase deduplication compared with VHD files. It is therefore recommended to convert VHD files to VHDX format when using thin volumes.

Use one of the following methods to convert between formats:

1. Use the Hyper-V UI in Windows Server 2012, select Edit the VHD file, and choose to convert to VHDX
2. Use the new Convert-VHD PowerShell cmdlet³

Note that the VHD conversion must be done with the VM shutdown.

Furthermore, the VMs using the VHDs may have file systems that have alignments that are different from the host file system. It is therefore recommended to follow the OS guidelines from this section for each VM.

³ For Microsoft Windows PowerShell Convert-VHD cmdlet syntax, see technet.microsoft.com/en-us/itpro/powershell/windows/hyper-v/convert-vhd.



VMware vSphere

The VMFS file systems use a 1 MiB block size; therefore, no tuning is necessary for deduplication with VMware®.

However, the VMs using the VMDK files within the VMFS may have file systems that have alignments that are not optimal for deduplication. It is therefore recommended to follow the OS guidelines from this section for each VM.

The VMware vCenter Converter Standalone can automate this process for Windows VMs, as it provides an option to change the Volume Cluster Size (NTFS allocation unit) during the conversion. On the Options page of the Conversion wizard, click **Data to copy** in the options list and follow this procedure:

- 1. From the **Data copy type** drop-down menu, choose **Select volumes to copy**.
- 2. Click **Advanced** and select the **Destination layout** tab.
- 3. Select the volume for which you want to change the cluster size.
- 4. In the **Cluster size** column, select the new cluster size of the destination volume.
- 5. In the **Copy type** column, you can verify that the cloning operation type is set to file level.

Figure 10 shows the Cluster size selector in the Conversion wizard.

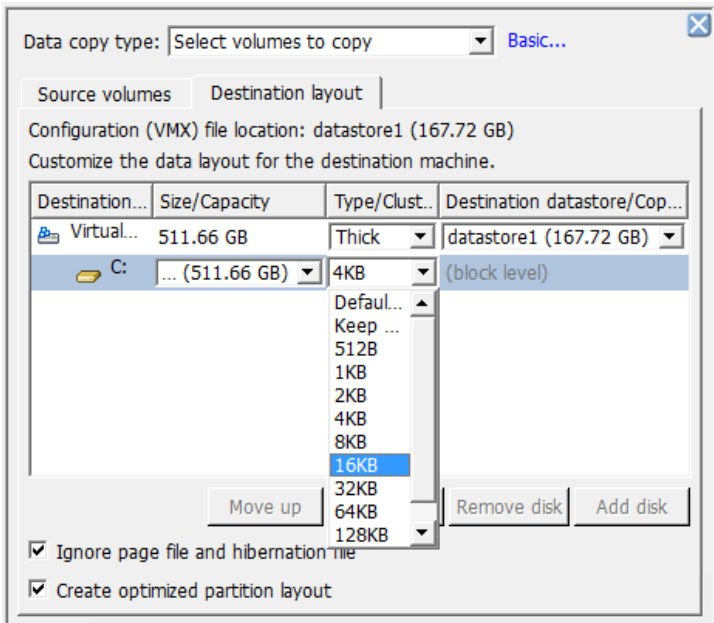


FIGURE 10. Selecting the new cluster size

Oracle Solaris

ZFS

ZFS is a combined file system and logical volume manager for Oracle Solaris systems. It uses variable-sized blocks, with 128 KiB as the default size. While this is good for deduplication within a file, it aligns files on physical block boundaries, which are 512 bytes by default and this significantly reduces deduplication between files containing like data. The interfile deduplication rate can be increased by configuring Solaris to use 16 KiB physical block sizes for HPE Primera volumes.

On SPARC-based Solaris platforms, the SAN SCSI Disk driver configuration file /kernel/drv/ssd.conf should be edited to include the following line:

```
ssd-config-list="PrimeradataVV","physical-block-size:16384";
```

On Intel®-based Solaris platforms, the SCSI Disk driver configuration file /kernel/drv/sd.conf should be edited to include the following line:

```
sd-config-list="PrimeradataVV","physical-block-size:16384";
```



After the system is rebooted, any new zpools created should have a 16 KiB block size. You can verify this by running the command `zdb -L`. The zpools created with 16 KiB block sizes will have an ashift value of 14 and the zpools created with the default block size will have an ashift value of 9 as shown in the following example.

```
# zdb -L | grep ashift ashift: 14
ashift: 9
```

UNIX file system

UNIX file system (UFS) is the original Solaris file system and is based on the BSD Fast File System. The maximum file system size is 1 TiB on Intel Solaris (x86) and 16 TiB on SPARC Solaris. The default logical block size to which files are aligned to is 8 KiB and this is also the maximum block size. Therefore, to obtain the best deduplication rate, it may be necessary to migrate to ZFS with the aforementioned block size tuning.

Linux

There are many different file systems available on Linux, but the most widely used are ext3, ext4, and XFS.

ext3 is an extension of the ext2 file system with journal support. It does not support extents and therefore files will be fragmented into block-sized chunks, which have a default, and maximum, size of 4 KiB. ext3 was introduced to the Linux kernel in 2001 and for a long time was considered to be the standard Linux file system.

ext4 superseded ext3 and replaced the traditional block-mapping scheme by extents. An extent is a range of contiguous physical blocks, improving large file performance and reducing fragmentation. ext4 uses a delayed allocation mechanism that buffers data and allows it to allocate blocks in groups. ext3 file systems can be migrated offline to ext4 by using the `tune2fs` command; however, existing files will retain their ext3 layout. Therefore, backup and restore operations are required to take full advantage of the ext4 improvements. ext4 was made the default file system in RHEL 6.

Extend file system (XFS) is a highly scalable, high-performance file system with metadata journaling and online defragmentation. Compared to ext4, it excels at multithreaded, parallel I/O workloads and scales higher with support for file systems up to 16 EiB and files of 8 EiB. There is no direct conversion available from ext3/ext4 to XFS, so migration involves backup and restore operations. XFS was the default file system in SUSE Linux Enterprise Server from SLES 8–SLES 11 and is the default file system in RHEL 7.

B-tree file system (Btrfs) is a copy-on-write logging-style file system. Rather than journaling changes before writing them in place, it writes them to a new location, then links it in. With SLES 12, the default file system for the OS is Btrfs, but XFS is the default for all other use cases.

The following example shows four CPGs with ext3, ext4, XFS, and Btrfs file systems created on a data reduction volume within them. Four copies of a Linux system backup were restored to each file system.

```
cli% showcp -s EXT3 EXT4 XFS BTRFS
      -Private[MiB]- -----[MiB]----- ----- Efficiency -----
```

Id	Name	Warn%	Base	Snp	Shared	Free	Total	Compact	Dedup	Compress	DataReduce	Overprov
14	BTRFS	-	8704	0	10752	21248	40704	4.29	1.38	1.55	1.57	0.00
11	EXT3	-	8704	0	6656	25344	40704	7.43	2.12	2.04	2.86	0.00
12	EXT4	-	8704	0	6656	25344	40704	7.45	2.14	1.83	2.73	0.00
13	XFS	-	4608	0	6656	8960	20224	10.96	2.92	2.27	3.79	0.00

It is clear from the example that Btrfs has the worst performance with data reduction and the two ext versions offer similar performance. However, the more sophisticated XFS provides the greatest space savings and therefore consideration should be given to migrating existing file systems to XFS when using data reduction volumes.



Symantec Storage Foundation

The Veritas File System (VxFS) of Symantec Storage Foundation is available on many platforms including AIX, HP-UX, Linux, Solaris, and Windows. It is an extent based file system that allocates storage in groups of extents rather than a block at a time.

VxFS file systems are formatted with a fixed block size ranging from 1 KiB to 8 KiB, which represents the smallest amount of disk space allocated to a file. The block size largely used to define the maximum size of the file system—a block size of 1 KiB allows a maximum file system size of up to 32 TiB and a block size of 8 KiB allows for a file system up to 256 TiB.

The extent-based nature of VxFS means that no tuning is necessary for use with HPE Primera Deduplication, as the alignment is not based on the block size.

HP-UX

Journaled File System

Journaled File System (JFS) is an implementation of the VxFS and due to the extent-based nature of VxFS, it means that no tuning is necessary for use with HPE Primera Deduplication.

Hierarchical File System

Hierarchical File System (HFS) is based on the BSD Fast File System. HFS file systems are not commonly used as their maximum size is limited to 128 GiB. However, HFS does support a wider range of block sizes (4 KiB–64 KiB) than other FFS implementations and therefore is suitable for use with HPE Primera Deduplication.

The default primary block size of HFS is 8192 bytes and can be checked by looking at the `bsize` field of the `dumpfs` command output.

```
$ dumpfs /dev/rdisk/disk63
**/dev/rdisk/disk63:
magic    5231994 time    Fri Mar 20 09:30:54 2015
sblkno   16      cblkno  24      iblkno   32      dblkno   192
sbsize   2048    cgsz    2048    cgoffset  32      cgmask   0xffffffff0
ncg      12800   size    104857600    blocks  102604584
bsize    8192   shift   13      mask    0xfffffe000
fsize    1024   shift   10      mask    0xffffffc00
frag     8      shift   3       fsbtodb  0
minfree  10%    maxbpg  256
maxcontig 1     rotdelay 0ms    rps      60
csaddr   192    cssize   204800  shift    9      mask    0xfffffe000
ntrak    16     nsect    32      spc       512    ncyl     204800
cpg      16     bpg      1024    fpg       8192   ipg      1280
nindir   2048   inopb    64      nspf      1
```

To create a new HFS file system suitable for use with HPE Primera Deduplication, use the `-b` flag of the `newfs` command to set a 16 KiB block size.

```
$ newfs -F hfs -b 16384 -f 4096 /dev/rdisk/disk63
mkfs [hfs]: Warning - 608 sector[s] in the last cylinder are not allocated.
mkfs [hfs]: /dev/rdisk/disk63 - 104857600 sectors in 168042 cylinders of 16 tracks, 39 sectors
107374.2Mb in 10503 cyl groups [16 c/g, 10.22Mb/g, 1536 i/g]
```



IBM AIX

JFS is a 64-bit extent-based file system. The initial version of JFS (also called JFS1) can support a file system of up to 1 TiB and a file size of 64 GiB. JFS2 (the Enhanced JFS) is similar to JFS but supports a file system of up to 32 TiB and a file size of 16 TiB.

Both JFS and JFS2 use a block size of 4 KiB and use extents of up to 64 GiB to allocate blocks to files. JFS2 also divides the space into allocation groups that are ranging from 8 MiB to 64 MiB, which allow the resource allocation policies to cluster disk blocks for related data. Due to the extent-based nature of JFS, no tuning is necessary for use with HPE Primera Data Reduction.

LEARN MORE AT

hpe.com/storage/hpeprimera

Check if the document is available
in the language of your choice.



Make the right purchase decision.
Contact our presales specialists.



Chat



Email



Call



Share now



Get updates

© Copyright 2019 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Intel is a trademark of Intel Corporation in the U.S. and other countries. Microsoft, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Oracle is a registered trademark of Oracle and/or its affiliates. Red Hat is a registered trademark of Red Hat, Inc. in the United States and other countries. SAP HANA is a trademark or registered trademark of SAP SE in Germany and in several other countries. UNIX is a registered trademark of The Open Group. Linux is the registered trademark of Linus Torvalds in the U.S. and other countries. VMware, VMware vSphere, and VMware vSphere Web Client are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions. All third-party marks are property of their respective owners.