Hewlett Packard Enterprise

Technical reference guide

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

HPE MSA GEN6 VIRTUAL STORAGE

CONTENTS

EXECUTIVE SUMMARY

This technical paper is an investigation into virtual storage as implemented on HPE MSA sixth-generation storage systems. It also explores the automated tiering engine, as well as new and supporting technologies of HPE MSA sixth-generation arrays. This paper can assist in the creation and implementation of ideal configurations that meet design expectations and reducing the possibility of undesired outcomes. The HPE MSA sixth-generation portfolio, which uses virtual storage exclusively, includes the HPE MSA 1060, 2060, and 2062 SAN and SAS storage arrays.

This document is not a user guide and does not list all features or explain how to configure them. For detailed information regarding the features of an HPE MSA sixth-generation array, use the links on the last page of this document to go to the core documentation.

Intended audience

This paper is for everyone involved in the design and implementation of storage solutions that include HPE MSA sixth-generation arrays. Technical sales staff tasked with designing an effective solution will benefit from an understanding of the HPE MSA architecture, and because it is a customer-installable product, the administrator that eventually configures it will also benefit. Hewlett Packard Enterprise recommends a current knowledge of basic storage concepts such as RAID, mechanical and solid-state drive (SSD) technologies, thin-provisioning, and storage networking.

SYSTEM BASICS

The design approach of an HPE MSA array features an active/active architecture that provides both flexibility and resiliency to failure. It ships in a rack-mountable 2U form factor that contains:

- Disk drive bays (either 24 x SFF[1] or 12 x LFF[1,2])
- Two hot-swappable power supplies units, each with integrated cooling fans
- Two hot-swappable controller units
- A passive midplane to which all components are connected
- Optional lockable bezel

HPE MSA arrays contain either SAN or SAS controller modules and support optional expansion disk enclosures that house additional disk drives. Expansion disk enclosures include I/O modules in place of controller modules that provide SAS connectivity between disk drives and controller units.

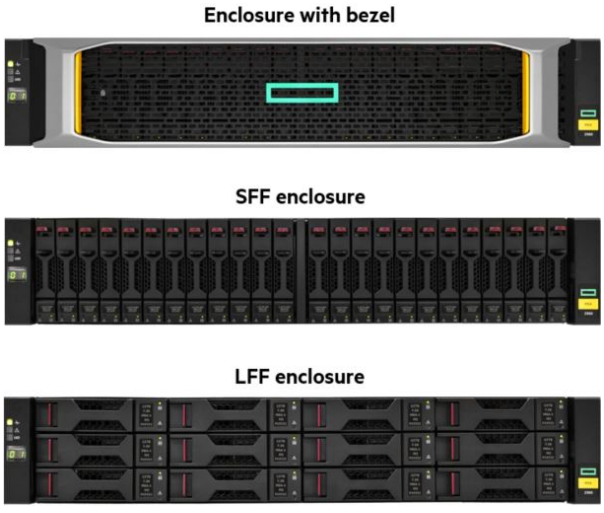

FIGURE 1. HPE MSA enclosures

[1] Small form factor 2.5"/large form factor 3.5"
[2] HPE MSA 2060 and 2062 only

HPE MSA 1060 arrays have a total of four host ports compared to the eight of the HPE MSA 2060 and 2062. However, HPE MSA 1060 SAS models support an optional fan-out cable that doubles the host port count by reducing the number of SAS lanes per port from four to two. Nevertheless, the fan-out cable provides increased scalability without compromising the performance of the array. Hewlett Packard Enterprise advises using fan-out cables even if they are not initially needed to avoid interruptions when connecting additional hosts.

FIGURE 2. Rear view of an HPE MSA 1060 array enclosure

FIGURE 3. Rear view of an HPE MSA 2060/2062 array enclosure

The HPE MSA 1060 array enclosures ship in SFF only but support a mix of up to three LFF and SFF expansion disk enclosures, which allows for a total of 96 SFF drives per array or 36 LFF drives and 24 SFF drives. HPE MSA 2060 and 2062 arrays support up to nine expansion disk enclosures and up to 120 LFF drives or 240 SFF drives.

IMPORTANT
HPE sixth-generation MSA arrays do not support disk enclosures or drives from previous generations of HPE MSA.

FIGURE 4. HPE MSA array enclosure naming convention

Controller architecture

HPE MSA arrays provide full redundancy in the event of a component failure. As shown in Figure 5, in support of availability and performance, each HPE MSA array controller contains its own set of hardware, including:

- Host ports (Fibre Channel, iSCSI, or 12Gb SAS)
- Management interfaces (Ethernet and serial over USB)
- Storage controller
- Management controller
- Memory/cache
- Internal backup power (supercapacitor)
- Nonvolatile embedded memory card (nonremovable)

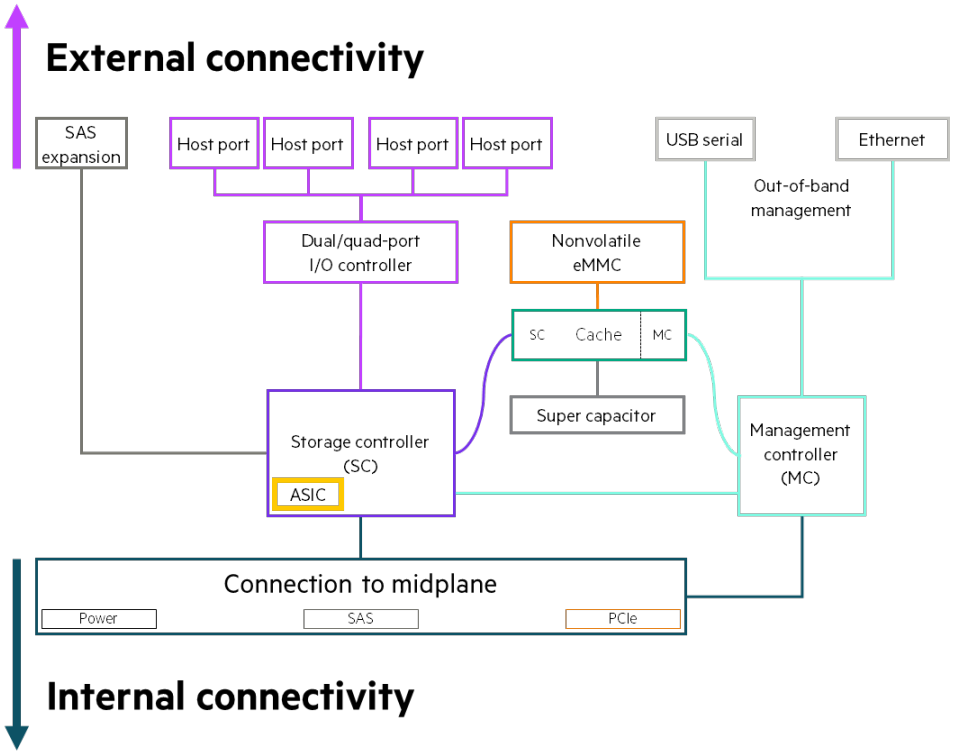

FIGURE 5. Simplified HPE MSA controller diagram

Governing each HPE MSA controller are two logical sub-controllers: the storage controller (SC) and the management controller (MC). These controllers are independent of each other and are composed of and connected to numerous subsystems that provide supporting connectivity and features.

The SC is responsible for the physical movement of data as well as maintaining data integrity. The storage controller is composed of SAS controllers and other similar low-level systems and contains an application-specific integrated circuit (ASIC) for RAID functions. The RAID ASIC supports the HPE MSA in maintaining high levels of performance, even when processing complex algorithms such as those used in RAID 6.

The management controller runs the Storage Management Utility (SMU), which is accessible by using a web browser over HTTP/HTTPS or through the command-line interface (CLI), preferably over SSH. The management controller operates out-of-band to the storage traffic. Additionally, the management controller is responsible for SNMP, SMI-S, and other system-to-system communications. To provide fault tolerance, each controller runs an instance of the SMU, and inter-controller communication occurs via a passive midplane to ensure that configuration information is kept current across both controllers.

NOTE
The HPE MSA uses stand-alone management IP addresses for each controller.

Controller cache

Each controller is an interconnected yet self-contained system; each also has an onboard cache. The cache is high-throughput, low-latency volatile memory assigned to several different system functions, as shown in Table 1.

TABLE 1. Assignment of system cache

Purpose	Per controller	System total
Total cache	12 GB	24 GB
Local read	1 GB	2 GB
Local write	1 GB	2 GB
Partner controller mirror—read	1 GB	2 GB
Partner controller mirror—write	1 GB	2 GB
Operating system overhead	8 GB	16 GB

IMPORTANT
It is a common misconception that the quantity of controller cache has a direct correlation to total system performance. Although this might be true of some architectures, it is not true of the HPE MSA, which includes an ASIC dedicated to RAID so that the general-purpose CPU is free to process other tasks such as tracking metadata. Reducing the resources required for RAID also reduces the cache needed to accommodate higher workload intensities. To accurately represent an array's sustained performance capabilities, testing by HPE engineering intentionally overwhelms controller cache to eliminate misleading and short-lived, cache-bound benefits. These results match mathematical models used for performance estimation by tools to accurately represent the array's performance in a given configuration under a defined workload.

Of the total controller read/write cache, 50% is a mirror of the contents of the partner controller, ensuring that if a controller becomes unavailable, I/O can continue through the remaining controller without data loss.

If the system loses external power, an internal supercapacitor within each controller provides the energy necessary to write the contents of cache to an embedded, nonvolatile memory card. The restoration of external power flushes the unwritten cache contents to disk.

SAS

HPE MSA array controllers communicate with all internal drives and connected disk enclosures via 12 Gb SAS. As of September 2020, the HPE MSA is the only HPE array to offer block-based shared storage over the SAS protocol. Although they share the same protocol, external host SAS traffic is entirely separate and unrelated to internal SAS traffic.

As represented in Figure 6, each HPE MSA controller has a single 12 Gb lane dedicated for each internal drive and connects to expansion disk enclosures through a four-lane 12 Gb mini-SAS connection. This architecture provides both high throughput and redundant paths to drives.

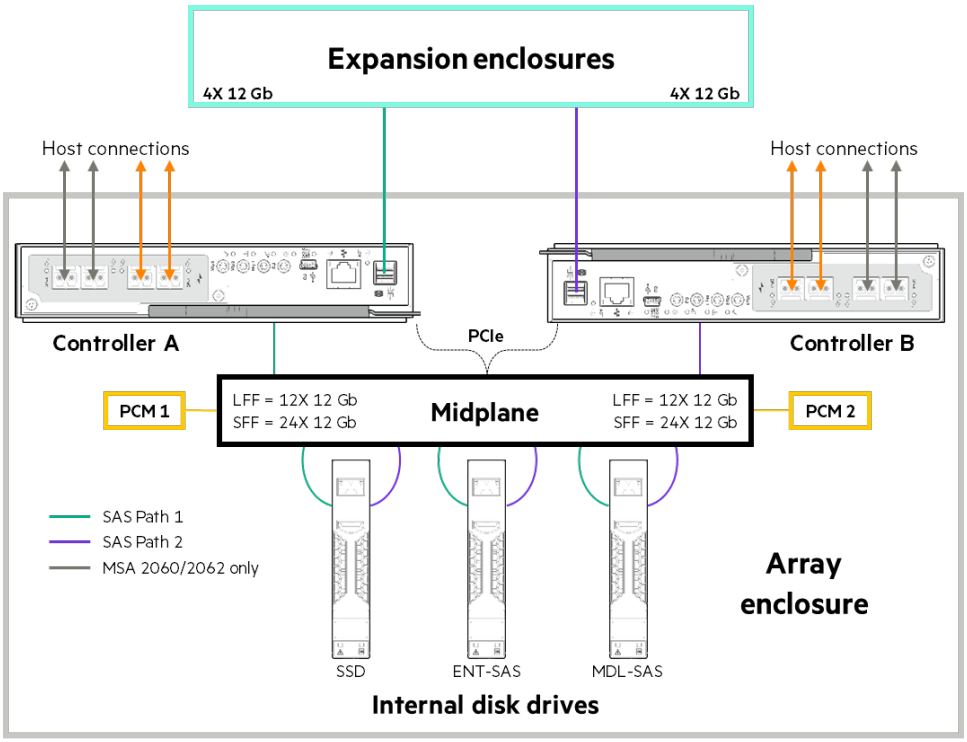

FIGURE 6. HPE MSA array architecture

IMPORTANT
For optimal performance, locate SSD drives within the array enclosure, which has a dedicated SAS lane for each drive. External enclosures connect to each array controller via two shared four-lane SAS cables.

VIRTUAL STORAGE

Virtual storage is the abstraction of underlying storage subsystems to add functionality, increase performance, and simplify the provisioning of the array. Virtual storage also makes it possible to employ software-driven data services that improve availability and overall efficiency.

In HPE MSA sixth-generation arrays, virtualization occurs at two distinct points. The first is at the disk layer in the form of hardware-accelerated RAID, and the second is at the software layer, which pools disk groups together.

RAID is a type of virtualization that aggregates multiple disk drives into a single object. HPE MSAs used RAID exclusively until firmware GL200 introduced virtual storage for fourth-generation arrays. Fundamentally, RAID increases addressable capacity, performance, and availability for all volumes located within a disk group, although the combination of benefits varies with each RAID level.

Although RAID does an excellent job of tackling these specific goals, it also results in the isolation of these same attributes. That is, volumes do not have access to the capacity and performance of other disks outside of their group. Another disadvantage of RAID-only architectures is that it is not possible to increase the capacity or the performance of a disk group without first expanding it with more disks. For small disk groups with low capacity disk drives, adding more disks might not be a cause for serious concern, but the more disks there are and the larger their capacity, the higher the risk of concurrent drive failures and subsequent data loss. Also, the expansion of a disk group initiates the restriping of existing data, thus impacting performance. Finally, a disk group can only grow to a finite size, which limits application and volume growth to a capacity less than a controller could otherwise support.

Virtual storage still uses these same RAID concepts but improves on them by not only aggregating disks but entire disk groups, thus allowing data to be wide-striped across all disks within a tier. Virtual storage enables an administrator to define the level of redundancy, capacity, and performance that provides appropriate classes of service for volumes within the pool while removing the complexities associated with traditional, RAID-only storage arrays. Additionally, virtual storage offers a mechanism for several advantageous features such as sub-LUN tiering, read cache extension, thin provisioning and rebuilds, and space-efficient redirect on write snapshots, to name a few. The benefits of virtual storage also extend to the management interface, where it is possible to carry out more tasks in fewer actions. For example, in addition to the simplification of volume creation, it is also easier to map multiple volumes to multiple hosts, as well as to visually represent array utilization more efficiently.

Controllers and pools

An HPE MSA array ships with and supports both a minimum and a maximum of two controllers supporting one pool each. A pool becomes available when the first disk group is assigned to it.

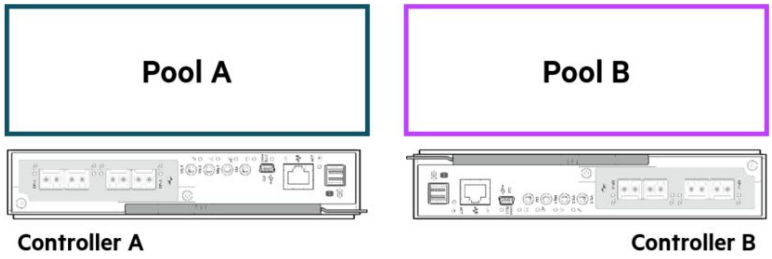

FIGURE 7. HPE MSA controllers and pools

NOTE

HPE MSA arrays are sold in dual-controller configurations only. Although an HPE MSA array can operate in a degraded state with a single controller, it introduces a single point of failure.

IMPORTANT

HPE does not sell sixth-generation controllers outside of an array enclosure except as spares for failed controllers.

As shown in Figure 8, a pool is the location of volume data. Because each controller "owns" a specific pool, a volume cannot span or otherwise use the capacity and performance of the other controller. Therefore, it is reasonable to consider each pool as entirely independent resources managed by the same interface. Nevertheless, each pool is accessible through the host ports of its partner controller through the Unified LUN Presentation (ULP) mechanism. In the event of a firmware update or controller failure, a pool remains online while the remaining controller takes temporary ownership of its disk resources.

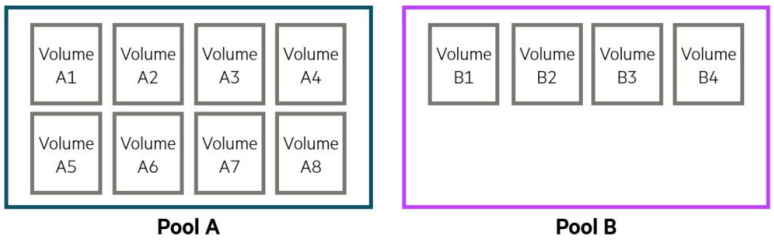

FIGURE 8. Relationship of volume and pools

It is supported and potentially advantageous to provision only one pool because this method allows a reduced impact to performance during a firmware upgrade or controller failure. For example, if the potential of a controller is 100,000 IOPS[3], then two controllers working at the same time could deliver 200,000 IOPS, which is the array's total potential. When a controller becomes unavailable, the potential of the array drops to less than 200,000 IOPS.

IMPORTANT

As of firmware I110, the array defaults for "controller-failure" and "partner-notify" are enabled although in previous firmware versions they were disabled. If a single controller becomes unavailable, these settings cause the remaining controller cache policy to switch to write-thru, which causes a degradation in write performance in return for the assurance that written data is committed to disk. Although the full performance of the remaining controller cannot be realized during single-controller operation, it is still important to consider controller headroom as the overall impact to application performance can be minimized. Refer to the HPE MSA 1060/2060/2062 Storage Arrays Best Practices document for further guidance.

[3] Example only. Real performance may be higher or lower and varies due to drive configuration and workloads.

In the example shown in Figure 9, Controller A has insufficient headroom to absorb the additional workloads temporarily relocated from Controller B. The result will be a minor but measurable degradation in overall array performance until the unavailable controller returns to an online state.

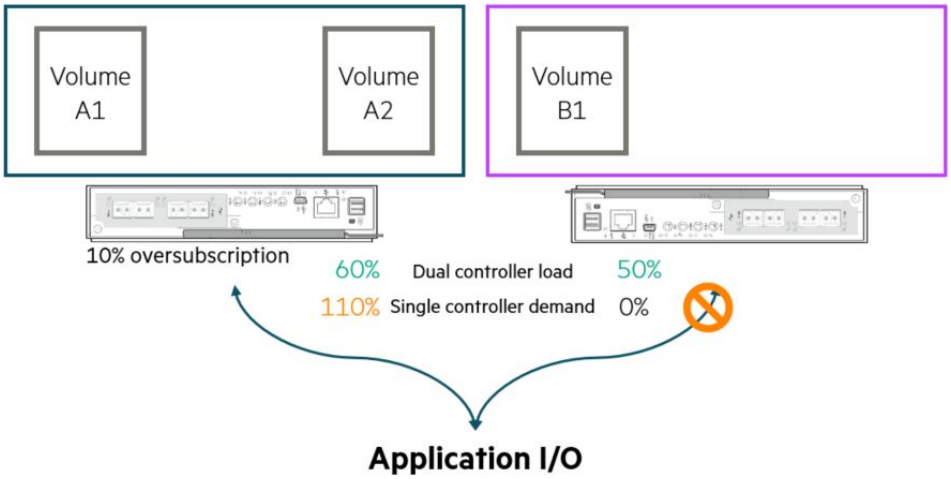

FIGURE 9. Oversubscription of Controller A during the unavailability of Controller B

As an active/active architecture, an HPE MSA supports three approaches when designing a solution regarding headroom:

- **Two pools with no headroom**—Provisioning both pools concurrently with sufficient drive resources allows workloads to consume the full potential of the array in terms of both performance and capacity. This option is sometimes chosen because controller downtime is typically very low, and planned controller firmware upgrades during quiet periods negates any performance impact. Additionally, when configured appropriately, an HPE MSA sixth-generation array performs so well that consistent consumption of more than 50% of the potential of the array is uncommon. As a result, there is often sufficient headroom, even if it is not explicitly managed.
- **One pool with 100% headroom**— Provisioning a single pool and placing all workloads in it reduces the impact to performance when a controller becomes unavailable and greatly simplifies configuration management. Because typical requirements for small and medium-sized businesses are far below the capacity and performance limits of a single pool, Hewlett Packard Enterprise recommends evaluating a single-pool strategy before considering dual-pools. System configuration limits are documented in the HPE MSA1060/2060/2062 Storage Management Guide.
- **Two pools with 100% headroom**—Similar to the second option, this method seeks to ensure that performance remains predictable during the time a controller is unavailable. This method might also be preferable when capacity beyond what a single pool can provide is required. However, keeping within this headroom after a system is in production can be challenging. For example, as a pool's capacity expands, workloads can grow and consume the additional performance and, therefore, headroom without the administrator realizing it.

IMPORTANT
When designing a storage solution, it is important to have a detailed understanding of all workloads as well as the capabilities of the array to avoid undesired outcomes.

A pool is a collection of 4 MB pages. The number of pages within a pool depends on the total capacity of all virtual disk groups associated with it. For example, if there were a single disk group with 1 TB of useable capacity, there would be a total of 250,000 4 MB pages (1,024,000,000/4,096 = 250,000). Although the number of pages in a pool can change, the size of the page never does; it is always 4 MB.

Within an HPE MSA, a volume is a collection of pages amounting to a defined capacity. The location of a volume's designated pages within the pool changes over time and is dependent on several factors, such as the use of tiering, snapshots, and thin provisioning. As shown in Figure 10, it is practical to visualize a pool as a grid filled with pages, which may or may not be in order. Each page has a contiguous range of logical block addresses (LBAs) for a volume assigned to it. A page is only associated with one volume.

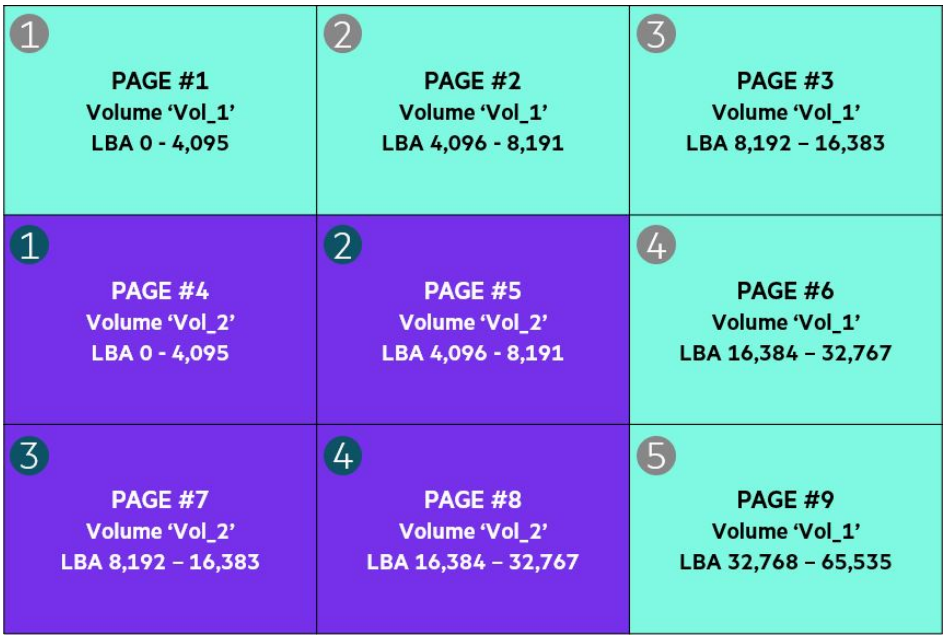

FIGURE 10. Representation of page assignment within a pool

An operating system has no awareness of data locality within the array.

As shown in Figure 11, the operating system lays a file system across what it sees as a physical disk, composed of multiple 512-byte sectors. Applications typically interact with the file system to read and write data, and the data granularity is often courser than that of the file system. In the example in Figure 11, the HPE MSA reads and writes to a single page for all sectors from #1 through to #8,192.

Host addressing	32K application block	#1								...	#128	
	4K file-system cluster	#1							...	#8	...	#1,024
	512b logical sector	#1	#2	#3	#4	#5	#6	#7	#8	...	#64	...	#8,192
	Disk LBA	#0	#1	#2	#3	#4	#5	#6	#7	...	#63	...	#8,191
Array addressing	Volume LBA	#0	#1	#2	#3	#4	#5	#6	#7	...	#63	...	#8,191
	Pool location	PAGE #1										

FIGURE 11. Addressing and data unit translation with theoretical application and file system block sizes

Automated tiering

A standout capability of virtual storage is automated sub-LUN tiering, which can increase overall system performance at a lower cost than possible with only one class of disk drive. Automated tiering can distribute the contents of a volume over multiple virtual disk groups within a tier as well as over multiple drive classes grouped to form a pool. This process exploits the performance benefits of a particular drive type, but only for the part of a volume that requires it. As shown in Table 2, an HPE MSA supports three tiers of storage; each tier relates to a specific drive class.

TABLE 2. Tiers and disk drive types

HPE MSA tier name	Disk type	Industry term
Performance	SSD	Tier 0
Standard	15K and 10K Enterprise SAS hard disk drives (HDDs)	Tier 1
Archive	7.2K MDL SAS HDDs	Tier 2

A tier is specific to a pool and includes all disk groups of a particular drive type associated with that pool. Figure 12 shows how pages of all volumes within a pool are distributed across all three tiers and all six disk groups that are associated with it.[4]

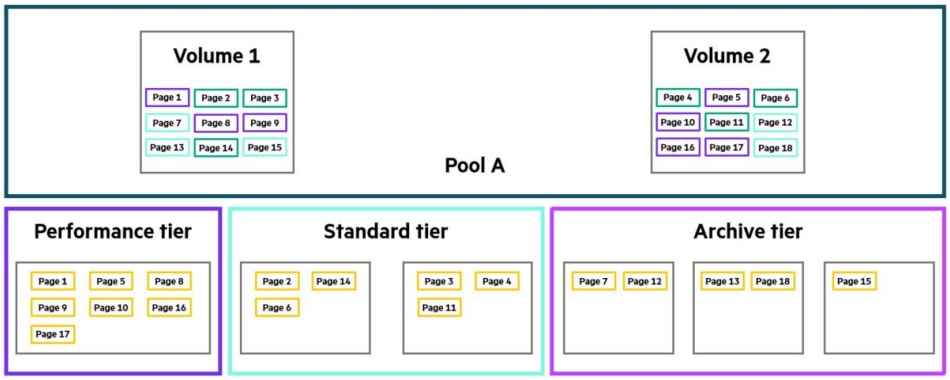

FIGURE 12. Example of volume distribution across multiple tiers within a single pool

Table 3 lists the four ways to configure a pool.

TABLE 3. Pool configurations

Configuration	Layout	Advantages/use cases
Single tier	Single drive class, wide-striping only, and no automated tiering	• Great for high capacity and sequential I/O centric configurations where mixed workloads are rare • Consistent and predictable performance for all volumes
Performance tiering	Tiering with SSDs. Can be one of two recommended layouts: • Two tiers: performance and standard tiers[5] • All three tiers	• Hybrid arrays are ideal for mixed workloads and a hands-off approach • SSDs counted towards to useable capacity
Archive tiering	Standard and archive tiers	Great for high-capacity solutions with relatively low-performance requirements
SSD read cache	SSDs used as an extension to read cache together with either a single-tier or archive tiering	Greatly accelerates random reads with a minimum investment and does not require a license

Without automated tiering, a pool would consist of only a single disk drive type and could not grow cost-effectively beyond the characteristics of that drive technology. However, a pool with a single tier can be the right choice for many solutions, particularly those requiring exceptional sequential performance and high capacities. As an entry storage array, HPE MSA arrays are typically deployed as a single solution within environments that contain multiple workloads requiring a mix of demanding random I/O and sequential I/O simultaneously.

For all and especially random workloads, flash storage is ideal because it provides an exceptional ratio of cost to performance. However, the cost-to-capacity ratio remains higher than that of mechanical drives. HPE supports configuring an HPE MSA as an all-flash array, which can be a cost-effective solution when a large capacity of low latency storage is required. However, all-flash configurations are suitable for datasets that are not strong candidates for data reduction via deduplication or compression. For such datasets, there are more suitable products within the HPE portfolio that can provide both a high level of guaranteed random I/O performance and a better ratio of cost to capacity.

Typically, hybrid storage arrays consisting of both SSDs and HDDs are the most effective route to achieving improved random and sequential performance without drastically overshooting a capacity goal or budget.

[4] Example only. Actual distribution of pages depends on several factors.
[5] Due to a substantial performance differential, performance tiering between the performance and archive tiers is supported but not considered best practice.

TABLE 4. Methods of increasing application performance

Method	Advantages	Disadvantages
Add more disk drives to a traditional RAID disk group (RAID 10, 5, 6)	• Improves performance for a disk group	• Can negatively affect the availability of the disk group and subsequently the pool. • Only improves performance for that disk group and a portion of a tier. • Consistent performance requires configuring all disk groups equally. Because it is not possible to expand a traditional RAID disk group, the disk group must be removed, resized, and reintroduced to a pool. It also requires sufficient unallocated capacity to be present within the pool to absorb the temporary loss of its capacity. • Might be disruptive to application performance. • Requires careful planning to run smoothly.
Add more disk groups to a tier	• Is a perfect solution for measurable sequential I/O performance gains • Does not impact the availability of a pool • Supports incremental growth of capacity and performance • Is the correct approach to meet capacity goals	• It might be necessary to overprovision a pool's capacity to achieve a performance goal, particularly for random I/O. • Is potentially more expensive if the goal is less than the sum of all additional drives.
Switch to a more performant drive technology	• Is a guaranteed solution to increase array performance up to the point of controller saturation or drive count limit • Is suitable for uncompromising application requirements	• Can be extremely expensive to reach both capacity and performance goals at the same time. • Temporarily invokes tiering unless data is migrated from one pool to another or restored from backup; configuring both SSD and HDD drive technologies at the same time requires a license if the array is not an HPE MSA 2062. • Might be disruptive, especially if using a backup and restore approach.
Automated tiering	• Can provide performance benefits for all volumes within a pool • Only a percentage of total capacity uses more expensive drive technology • Is simple to deploy • Can be configured online	• Requires a license for HPE MSA 1060/2060. • Can be overwhelmed by sustained intense random workloads on improperly provisioned pools.
SSD read cache	• Is very cost-effective • Requires only one SSD per pool to get started • Does not require a license • Is well-suited for random read dominant workloads	• Takes time to warm up. • Does not directly accelerate writes. • Does not accelerate sequential access. • Is limited to 4 TB per pool, which limits the pool size when staying within the minimum recommended 10% of pool capacity.

The HPE MSA automated tiering engine is responsible for ensuring the optimal location for data, including when data is first written to a volume. Some tiering solutions attempt to deliver flash-like performance by writing all inbound data to the fastest drive class within the system. However, such an approach is not as efficient as that of the HPE MSA and presents several problems:

- Migrating data from one location to another requires more back-end I/O to make space for new data.
- Scheduled page migrations require adequate "quiet time" so application performance is not impacted.
- Keeping the fastest tier full creates greater contention for hot read data.
- If writes are not separated, disk drive types that are well suited to sequential I/O cannot be taken advantage of.

An HPE MSA solution solves these problems by redirecting inbound writes to the most appropriate tier. This approach requires less back-end I/O to relocate pages later and also frees capacity within the higher performance tiers for more relevant data. As shown in Figure 13, when performance tiering is used, the engine can detect inbound data streams and will direct sequentially written application data to the fastest tier comprising traditional HDDs. Sequential I/O is well suited to mechanical disk drives, whereas random I/O is latency-sensitive and best served by SSDs. Thus, this in-line approach significantly improves overall performance and does not require any administrative intervention. In both scenarios, if the destination tier is full, then the next fastest tier with available capacity will be used instead.

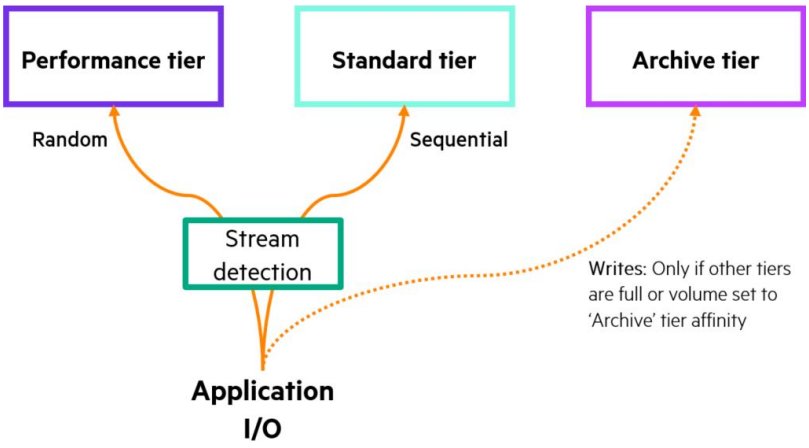

FIGURE 13. Automated redirection of inbound writes when using performance tiering

When archive tiering is used, all inbound writes land on the standard tier unless its capacity is exhausted or if a volume has the Archive tier affinity setting applied. Refer to the Tier affinity subsection of this paper for more information.

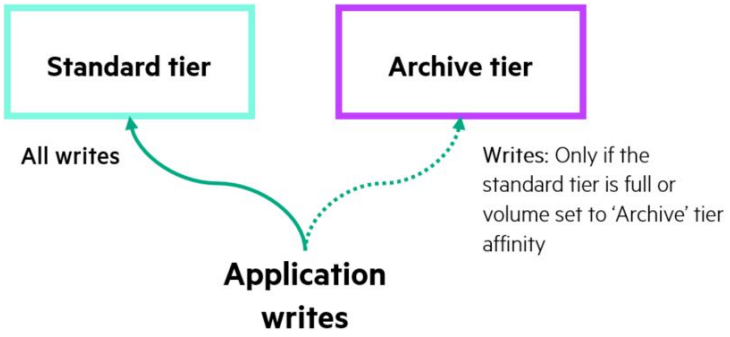

FIGURE 14. Archive tiering write mechanism

Unlike some tiering mechanisms, an HPE MSA does not schedule page migration to occur on a daily or hourly basis. It also does not require a large quantity of expensive low-latency memory or cache to buffer uncoordinated back-end writes. Instead, an HPE MSA tiering engine uses minimal resources, which allows it to relocate pages almost continuously if necessary.

With a frequency of every five seconds, the tiering engine analyzes a pool and promotes or demotes pages according to how in demand they are in a process known as **page ranking**. Page ranking seeks to keep active data on the fastest drive type within a pool by tracking what pages are accessed the most often within a given timeframe.

TABLE 5. Page ranking example

Page rank	Current page	Volume	Direction
First	Page 3	Volume A	Promoted
Second	Page 7	Volume B	Promoted
Third	Page 5	Volume A	Promoted
Fourth	Page 2	Volume A	Promoted
Fifth	Page 1	Volume A	Promoted
Sixth	Page 6	Volume A	Demoted
Seventh	Page 8	Volume B	Demoted
Eighth	Page 9	Volume A	Demoted
Ninth	Page 4	Volume B	Demoted

Another way to think of ranking is as a heat map where frequently accessed pages are hot and infrequently accessed pages are cold. As data cools, it becomes eligible for eviction as needed, to lower-performing and less-costly tiers.

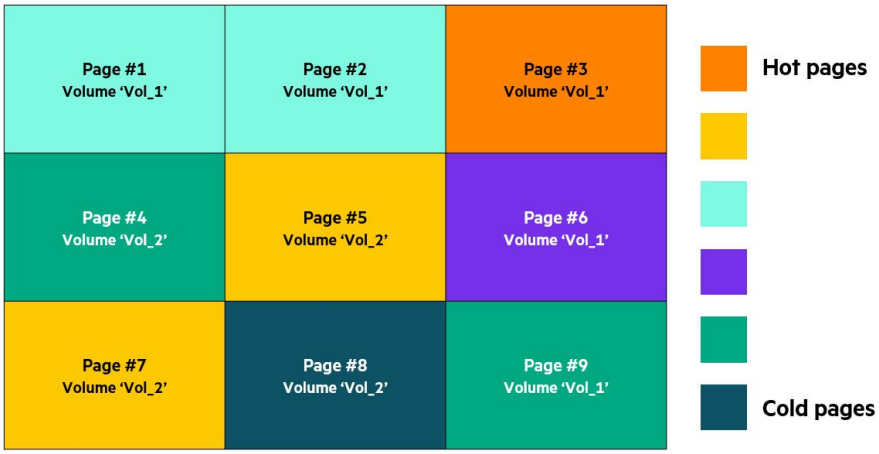

FIGURE 15. Heat map example of Table 5

IMPORTANT
Tiering is not a function that can be enabled or disabled. The tiering engine is always "on," and page ranking takes place even in a single-tiered system. The volume-level tier affinity setting is the only influence available to the user to manipulate the tiering engine.

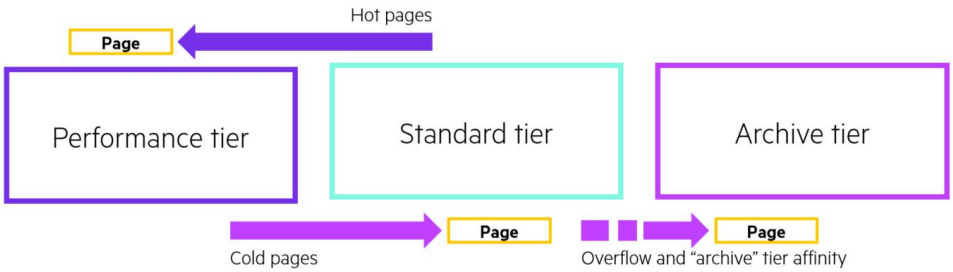

FIGURE 16. Performance tiering behavior

The migration of a page requires that it be read from one disk group and then written to another, which is an action that consumes performance. Under regular operation, several mechanisms are in place to ensure that reallocations do not harm performance. The only condition for circumventing these mechanisms is during the removal of a virtual disk group from a pool. If there is sufficient capacity remaining in the pool, removing a disk group drains pages allocated to it to the remaining disk groups.

General page migration rules:

- A page will not migrate more often than every 15 minutes.
- Page movement is throttled to move not more than a fixed number of times within a five-second interval.
- Unless a volume's tier affinity is set to **archive**, cold pages are not demoted from a higher tier unless to make room for hot pages.

IMPORTANT
Ideally, the performance tier should account for 80% or more of typical I/O within 24 hours. After it is configured and workloads are established, evaluate daily capacity usage using the in-array I/O Workload graph from within the Capacity menu. A safe quantity of SSDs to use when sizing a solution before an array enters production is between 10% and 20% of the standard and performance tiers capacity combined. Although some workloads may still benefit, Hewlett Packard Enterprise recommends that this ratio not drop lower than 10%. If less than 10%, performance would likely become less predictable over time as the pool fills with data. Conversely, increasing the performance tier beyond 20% may devalue the cost savings made possible by the tiering engine.

Tier affinity

The HPE MSA automated tiering engine adapts to changing workloads across all volumes within a pool and efficiently locates data to the right tier automatically. However, users accustomed to alternative tiering mechanisms might want to pin a volume to a specific tier—for example, the performance tier—to ensure that it receives SSD performance 100% of the time. However, due to the effectiveness of the HPE MSA tiering implementation, this is unnecessary. Pinning a volume to the performance tier reduces the SSD capacity available to other volumes, thus introducing wastage and possibly causing a pool-wide degradation in performance.

As an example, database administrators are often concerned about the performance of the TempDB volume, which can impede the performance of a system if not kept in an appropriate location. However, this type of volume is consistently hot and naturally finds itself promoted to the fastest tier within a pool[6].

Nevertheless, in some scenarios, the ability to locate most of a volume within a specific can prove desirable, and the tier affinity setting provides a mechanism that attempts to do so.

Tier affinity is a per-volume setting that can be applied during or after the creation of a volume. Its function is to modify the default page ranking algorithm so that pages for a volume are more likely to be favorably located in either the performance or archive tier. There are three settings, as shown in the following table.

TABLE 6. Tier affinity settings

Setting	Effect
Performance	Increases the likelihood of being promoted to the highest tier in the pool.
No affinity	Default setting and standard tiering engine rules: • Hot pages—Performance or highest tier • Cold pages—Standard or next highest tier from current location • Overflow[7]—Archive tier
Archive	The lowest tier in the pool absorbs newly written data, and there is an increased likelihood of page demotion when other volumes require capacity in the higher tiers.

NOTE

Changes to the tier affinity setting do not produce measurable results immediately. Tier affinity only acts on an existing page when it is either ranked for promotion (performance affinity) or the space it occupies within a tier is needed for higher priority data (archive affinity).

The No Affinity setting is the system default and is recommended for most volumes. The No Affinity setting attempts to keep the hottest pages in the highest tier and all other pages in the next highest tier that has free capacity. When a pool has three tiers, the archive tier remains unused until no capacity remains on the tiers above it to avoid the needless reduction in performance that using it could cause. The archive tier is first accessed for page overflow and is used to store only the coldest pages or new data when all other tiers are full.

The Performance tier affinity setting might sound like a good choice for any volume but should be applied sparingly. The typical use of the Performance tier affinity setting is for application data frequently in need of flash performance but that has extended periods between heightened read activity. In a correctly proportioned pool, it should not be possible to migrate the contents of all volumes to the performance tier by using the Performance tier affinity setting because there would be insufficient capacity, which limits its effectiveness. If all volumes must be in the performance tier, it would be more appropriate to have a dedicated all-flash pool or select a dedicated all-flash array from elsewhere in the HPE Storage portfolio.

More often, the most useful tier affinity setting other than the default is the Archive affinity setting, which allows the use of the archive tier even when capacity remains in the other tiers. Because it is the lowest-performing tier in a pool, the archive tier is well suited to data that will not be accessed after it is initially written to disk, or when high performance is not required. Examples include disk images, test data, or CCTV video streams, which are all good candidates for this setting. There are two distinct benefits to employing this strategy:

- It frees up capacity in higher-performing tiers for use by more performance-sensitive volumes.
- It reduces the back-end I/O required to migrate pages from one tier to another later on.

[6] Provided the proportion of SSD within the pool is appropriately sized

[7] Overflows are pages that are either demoted to the lowest tier to make room for hotter data, or that will not fit into the higher tiers during ingest.

SSD read cache

As a rule, performance tiering is the best way to increase overall system performance. However, in environments where workloads are mostly random read-heavy, SSD read cache offers excellent benefits at an even lower cost. SSD read cache does not count toward array capacity and therefore does not require disk-level redundancy or a license. Additionally, although designed to accelerate random reads, the offloading of reads to SSD frees HDDs from processing I/O, potentially further improving performance.

SSD read cache supports pools containing either a single-tier or archive tiering, and functions within the following set of rules:

- The minimum is one drive per pool (NRAID)[8].
- The maximum is two drives per pool (RAID 0)[9].
- A maximum addressable read cache size is 4 TB per pool.
- SSD read cache cannot be used with performance tiering in the same pool.
- One pool configured with SSD read cache and the other pool with performance tiering.

IMPORTANT

Ideally, SSD read cache should account for 80% or more of typical I/O within 24 hours. After configuring SSD read cache and establishing workloads, evaluate daily capacity usage using the in-array I/O Workload graph from within the Capacity menu. A safe quantity of SSD is between 10% and 20% of the standard tier when sizing a solution before an array enters production. Although some workloads may still benefit, Hewlett Packard Enterprise recommends that this ratio stay greater than 10%. If less than 10%, performance could become less predictable as the pool fills with data. Conversely, increasing read cache beyond 20% may devalue the cost savings made possible by the SSD read cache mechanism.

To use the maximum addressable read cache size of 4 TB per pool efficiently, Hewlett Packard Enterprise recommends a 3.84 TB read-intensive (RI) SSD.

Figure 17 illustrates how data is fetched when using SSD read cache. However, except for Step 4a, the process is the same regardless of the pool's configuration. The array always completes a read I/O using the fastest storage medium that holds a copy of the data. If performance tiering instead of SSD read cache is used, SSD is a possible location for requested data when that data is not in controller cache (Step 4b).

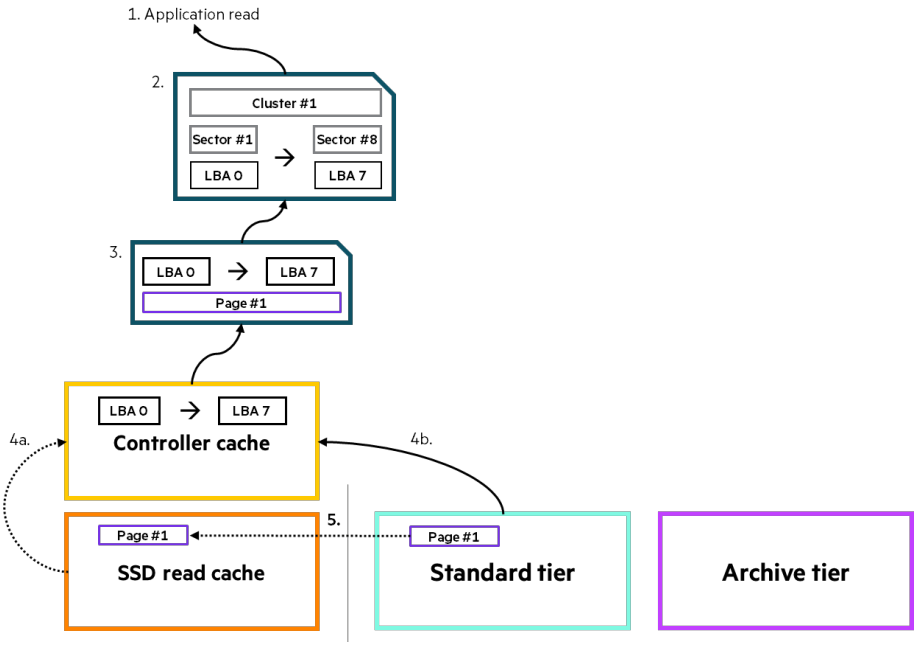

FIGURE 17. Read mechanism

[8] NRAID (No RAID) is only applicable to read cache disk groups.
[9] RAID 0 is only applicable to read cache disk groups.

The steps illustrated in Figure 17 are:

1. An application requests data from the file system.
2. The file system reads the appropriate disk sectors.
3. HPE MSA virtual storage matches the resulting LBAs to the relevant pages within the pool.
4. Unless caching is disabled for a volume, recently read, written, or prefetched data might already be in controller cache, in which case the I/O will immediately complete. Otherwise:
   a. SSD read cache, which is the next fastest medium, might contain the page as a result of a page copy during Step 5. If so, the relevant LBAs are read from the page and copied to controller cache for I/O completion.
   b. If SSD read cache does not contain a copy of the page, then the LBAs will be read from the mechanical disk group that holds the requested data.
5. Virtual storage uses hints from controller cache regarding whether to copy the page into the SSD read cache. For example, a page containing randomly accessed blocks is likely to be nominated, but a page read sequentially is not.

Virtual disk groups and wide-striping

Virtual disk groups are the fundamental building block of a pool and defined as a group of individual disks using RAID to provide a logical unit of capacity and performance. A pool can consist of one to 16 disk groups. There are no enforced rules regarding the geometry of these disk groups, although there are best practices to ensure consistent performance. A goal of this paper is to explain HPE MSA technology well enough to understand why certain best practices are recommended, but not to define them all. The HPE MSA 1060/2060/2062 Storage Arrays Best Practices paper contains the specific best practices that should be applied.

Virtual storage uses a controller's general-purpose CPU to track page metadata and location. In contrast, RAID is an ASIC accelerated disk group level function responsible for the distribution of a page across individual disks. Although an HPE MSA array can support NRAID and RAID 0 for read cache disk groups, neither can be used in capacity disk groups because neither scheme protects from drive failure. Table 7 summarizes the RAID levels supported for use in virtual disk groups to provide capacity to a pool.

TABLE 7. Supported RAID levels for virtual disk groups

RAID level	Protection	Notes	Recommended tier[10]
RAID 1	Mirroring	• Excellent performance for random workloads • More practical than RAID 10 when using MSA-DP+ for lower tiers	Performance
RAID 10	Mirroring and striping	• Same performance benefits as RAID 1 but allows for up to 256 drives per pool compared to 32 using RAID 1.	Performance
RAID 5	Distributed parity (D+P)[11]	• Good performance • Cost optimized	Performance
RAID 6	Advanced Data Guarding (D+P+Q)[11]	• Great availability • Recommended when using less than 12 HDDs including spares	Standard Archive
MSA-DP+	Distributed erasure coding (D+P+Q+nS)[11]	• Excellent availability • Great performance • Excellent rebuild times • Expandable • Recommended when using 12 HDDs or more	Standard Archive

NOTE

MSA-DP+ is the default and recommend protection choice for both the standard and archive tiers (HDDs). Although supported, MSA-DP+ is not necessary within the performance tier because drive failure and scalability are not common challenges for solid-state drives.

Figure 18 illustrates the relationships among physical disks, disk groups, tiers, and pools. Note that each disk group uses a given RAID type only as an example.

[10] Based on drive capacity and potential rebuild times
[11] D=data, P=parity chunk #1, Q=parity chunk #2, nS=multiples of spare chunks

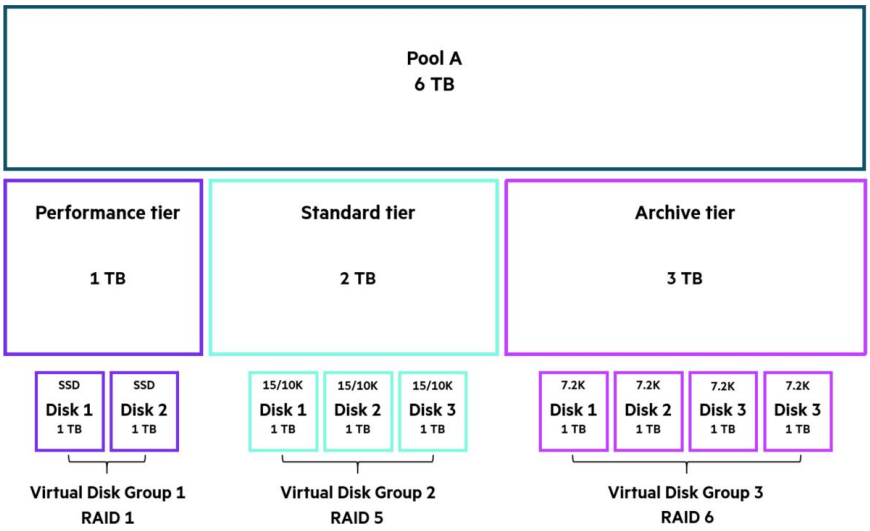

FIGURE 18. Representation of a multi-tiered pool layout

Virtual storage distributes pages one at a time across all disk groups within the tier as part of a process known as **wide striping**. For tiers composed of mechanical drives, wide striping offers significant performance benefits by both reducing the effects of latency and by multiplying data transfer rates.

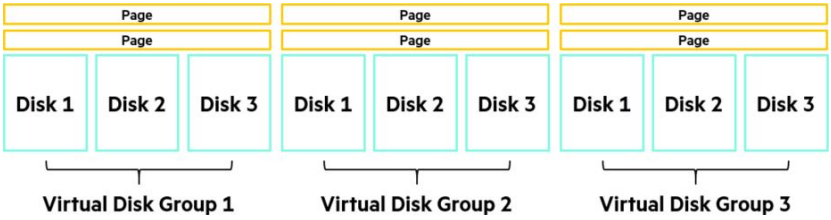

FIGURE 19. Wide-striped page distribution

It is helpful to understand the physical constraints of mechanical disk drives to understand how wide-striping improves performance. Within a mechanical disk drive, a head connected to an actuator arm can move one-dimensionally between the outer and inner extremes of the disk to access data. It is the rotation of the platter that brings a second dimension, thereby allowing heads to access the entire disk surface.

However, the longer the platter takes to revolve, the longer I/O must wait to complete. The time it takes to move a head to the correct track, rotate the platter so that it aligns with a given sector, and begin data transfer is known as the **access time**. Access times translate into latency, and the higher the latency, the slower an application will perform. Random I/O can easily overwhelm a mechanical disk drive due to constant consecutive random sector access and results in poor performance. However, sequential I/O is likely to result in fewer mechanical movements to seek a sector and is well suited to this workload type. Additionally, if they are not disabled, controller cache prefetching algorithms increase sequential read performance by copying the next range of data into memory ahead of time from the next disk.

Figure 20 uses a dotted arc to depict the location of data. The figure illustrates how sequential I/O is well suited to spinning media because it does not require dramatic physical movements of the read/write head.

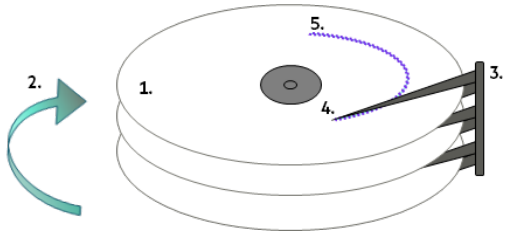

FIGURE 20. Internal view of a disk drive processing a sequential read/write operation

Components of a spinning disk drive illustrated in Figure 20 are:

1. Disk platter
2. Rotational direction of the disk
3. Read/write heads and actuator arm assembly
4. First sector where data is to be accessed
5. Last sector where data is to be accessed

In contrast, data accessed randomly requires a lot of physical movement of read/write heads and the platter, which leads to longer access times.

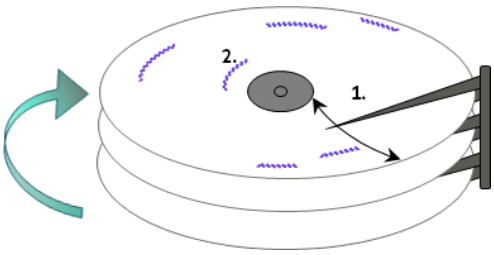

FIGURE 21. Internal view of a disk drive processing a series of random read/write operations

Figure 21 shows:

1. Movement of actuator arm across the disk surface
2. Data to be accessed concurrently but at various physical locations

MSA-DP+ disk groups

MSA-DP+ is a RAID type that offers superior performance, availability, flexibility, and rebuild times when compared to other disk groups, especially those that employ parity-based RAID schemes. MSA-DP+ uses distributed RAID to allow disk groups to grow from 12 to 128 drives incrementally and to feature integrated sparing.

A problem that keeps traditional disk groups limited to a relatively low number of drives is that the greater the drive count, the greater the probability of disk failure within that group. RAID serves to protect against such failures, but as drive capacities increase, the time to rebuild lost data increases with it linearly. It is the combination of both drive capacity and their number within a disk group that introduces higher risk as either variable increases.

Although the HPE MSA intelligently rebuilds only allocated capacity, it still requires that there be compatible drives available or global spares assigned to the role. Additionally, the management of spare drives is often a growing challenge that can lead to unplanned downtime, typically because they are unknowingly consumed by degraded disk groups. MSA-DP+ solves all of these challenges and more.

Unlike other disk group types where every drive participates in a stripe, MSA-DP+ disk groups contain multiple stripe zones that are of a fixed size. The more drives that participate in the MSA-DP+ disk group, the more stripe zones there are. A stripe zone is composed of 2,048 stripes laid out in a contiguous manner across the LBAs of each participating drive. Each of the 2,048 stripes within a stripe zone stores a single 4 MB page, totaling 8 GB of data across the zone. To demonstrate the rules of stripe zone distribution within a disk group, Figure 22 shows a small subsection of a much larger disk group composed of the 12-drive minimum for an MSA-DP+ disk group.

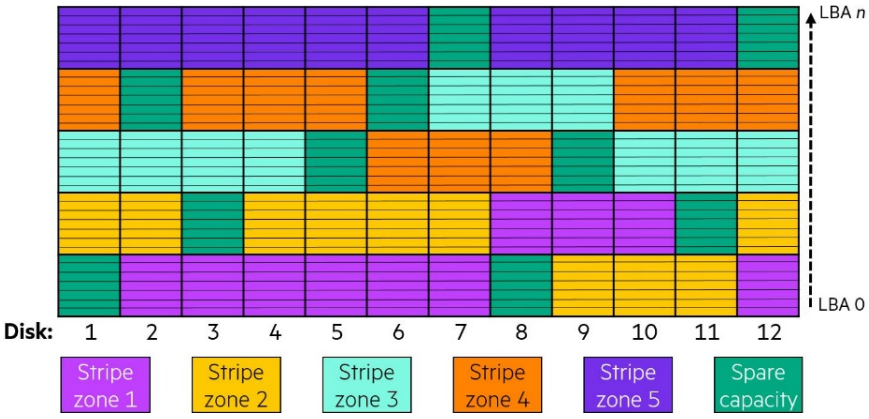

FIGURE 22. Example of multiple stripe zone and spare capacity within an MSA-DP+ disk group

Stripe zones obey fixed rules across the disk group:

- A stripe zone always spans exactly ten drives, but the drives do not need to be adjacent physically or logically.
- Each drive can only contain one section of a stripe zone (2,048 chunks[12]).

As illustrated in Figure 23, stripe zones are protected internally by RAID 6 in an 8+2 layout, which is eight chunks of data and two of parity (P&Q). Any additional drives in the disk group will not participate in the stripe zone, which for this example are Drives 1 and 11. As with any other RAID 6 implementation, parity rotates across all participating drives with each stripe. Each drive in a stripe holds 512 KB of data or parity information.

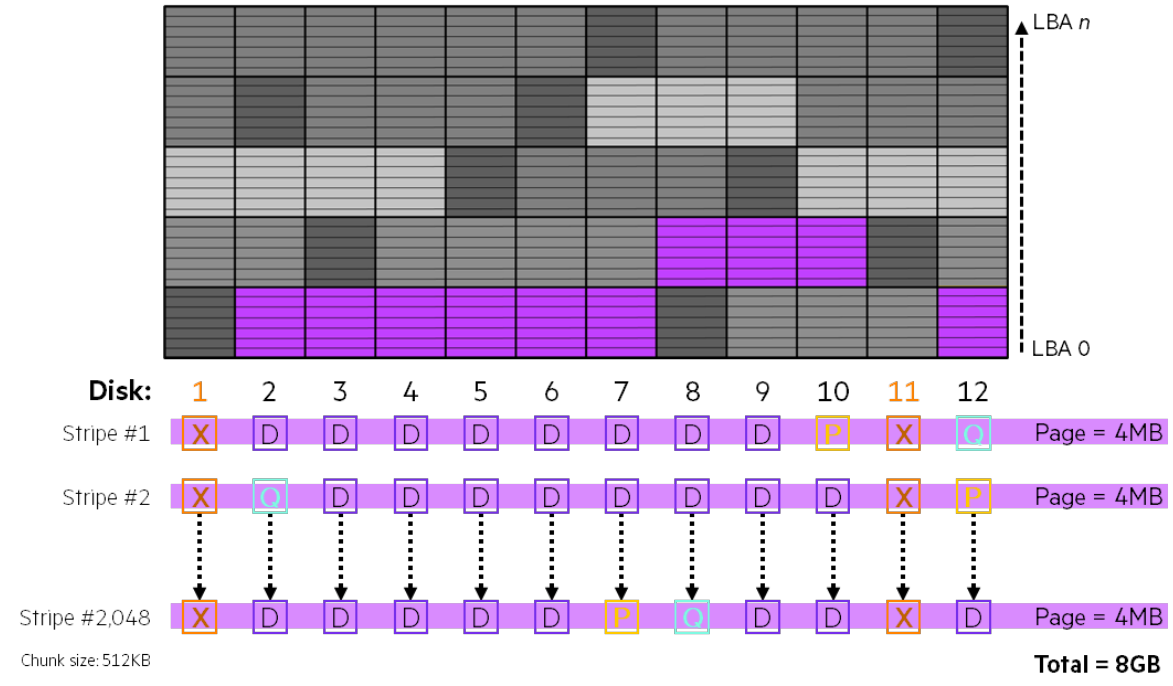

FIGURE 23. Internal protection of an MSA-DP+ stripe zone using RAID 6

As drive technology evolves and their capacities increase, so do the size of the disk groups, which often fuels growing concerns regarding availability. However, MSA-DP+ employs several techniques to reduce the time necessary to recover from drive failure, which leads to exponentially decreasing rebuild times the more drives there are within a disk group.

[12] A chunk is the amount of contiguous data written per disk in a stripe.

TABLE 8. Comparing recovery times from critical to degraded

Metric	RAID 6 (16 drives)	MSA-DP+ (16 drives)	MSA-DP+ (53 drives)
Random read IOPS performance impact (no rebuild I/O)			
1 drive failed	-44%	-30%	-12%
2 drives failed	-64%	-47%	-21%
Time to rebuild data from one failed drive[13]			
2 drives failed	55 hours	21 hours	2 hours
Time to fully fault-tolerant			
1 drive failed	55 hours	37 hours	11 hours
2 drives failed	55 hours	71 hours	20 hours

In addition to stripe zones, MSA-DP+ disk groups contain distributed spare capacity, which by default and at minimum equals the capacity of two of the largest in the disk group. For example, if every drive is 10 TB and there are 12 drives within the disk group, then the group has 20 TB of spare capacity.

Distributing spare capacity across all drives eliminates the two most significant disadvantages of dedicated spare drives:

- A spare drive will ordinarily sit idle and not contribute towards the performance of a disk group.
- Disk group rebuild times are dependent on the performance of a single drive.

In the simplified representation of an MSA-DP+ disk group shown in Figure 24, each drive accounts for five groups of 2,048 chunks. Therefore, the capacity of two drives to be used for sparing equates to ten groups of 2,048 chunks distributed across ten drives within the disk group.

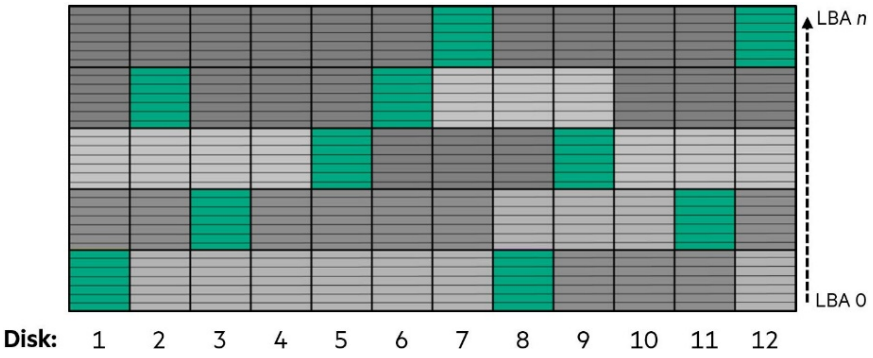

FIGURE 24. Example of spare capacity distribution within an MSA-DP+ disk group

To provide additional resiliency and as an equivalent of defining additional spare drives for traditional disk groups, it is possible to define a target spare capacity larger than the sum of two drive's capacity. For example, if a disk group contained thirteen 1.2 TB drives, it would be possible to set a target capacity of 3,600 GB.

IMPORTANT
After setting a target capacity higher than what is currently available, it is necessary to expand the MSA-DP+ disk group. However, this action can only occur after the disk group has finished initializing.

[13] Based on 10 TB drives and 50 MB/s per drive rebuild rate

NOTE
Unlike other disk group sparing mechanisms, MSA-DP+ disk groups observe slot affinity for failed drives. If a drive fails and a new drive is installed in its place in the same location, the new drive will automatically join the failed disk group, and the rebuilding of missing data will commence.

If a drive fails, missing sections of affected stripe zones rebuild into available spare capacity within the disk group. However, data cannot rebuild to a drive that already holds data for the affected zone as it would otherwise reduce the availability of the disk group and pool. Because a disk contains multiple zones, all drives that hold data for those stripes work in unison to rebuild data in a many-to-many relationship rather than many-to-one when using dedicated spare drives.

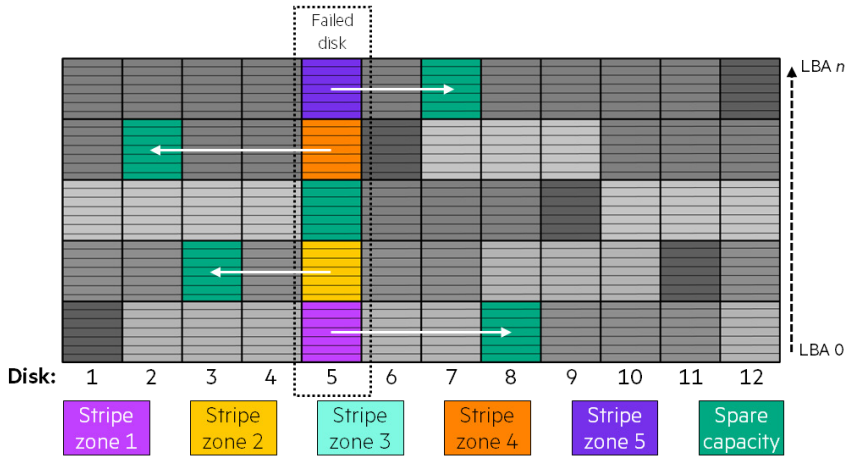

FIGURE 25. Example of rebuilding data located on a failed drive to spare capacity

When a single drive fails within an MSA-DP+ disk group, there will likely be a mixture of unaffected (fault-tolerant) stripes and some that are degraded, meaning that one of the ten drives that hold data for that zone is unavailable. When a second drive fails, there will likely be a mixture of fault-tolerant, degraded, and critical stripes that have lost two of their ten drives. To minimize the impact on availability, MSA-DP+ disk groups recover allocated capacity in two stages:

- Stage 1: Rebuild stripes that have sustained losses from both drives (Figure 26)
- Stage 2: Rebuild all remaining stripes (Figure 27)

Stage 1 of a rebuild assigns priority to returning critical zones to a degraded state as it protects the zone from data loss should a third drive holding zone data fail. After Stage 1 is complete, Stage 2 rebuilds remaining degraded zones until they are fully fault-tolerant.

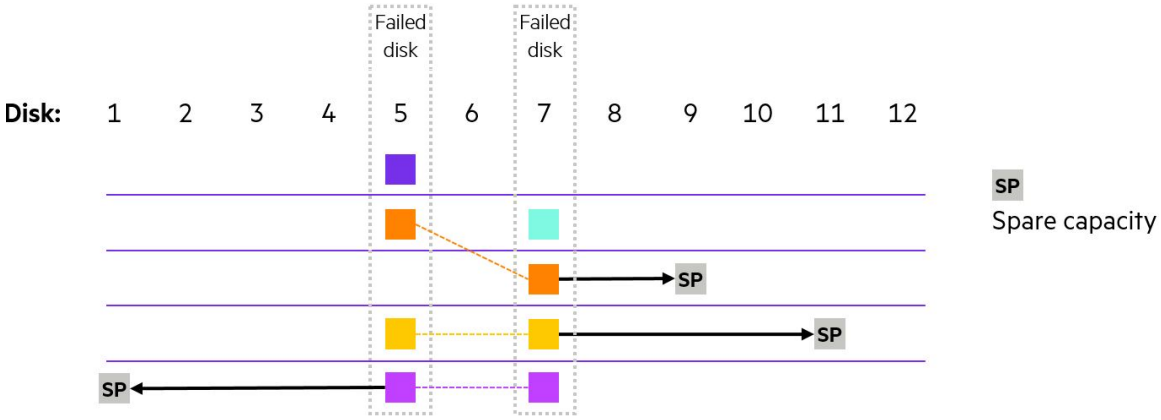

FIGURE 26. MSA-DP+ double disk failure rebuild Stage 1 is to protect critical stripe zones

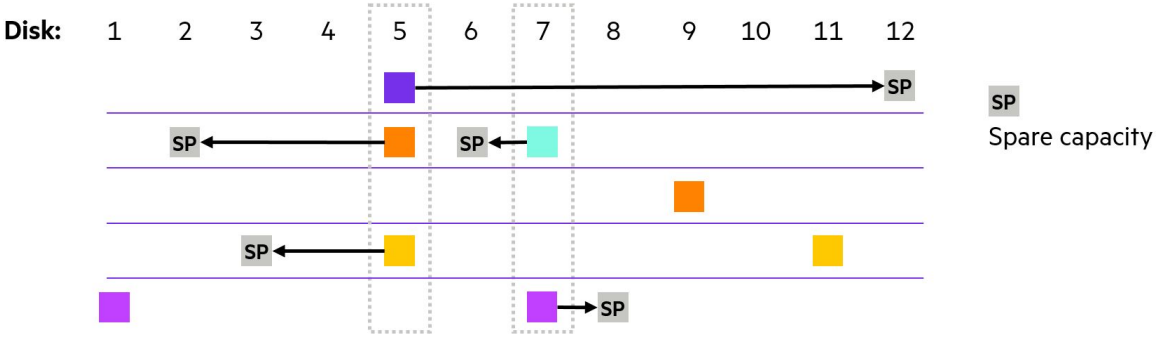

FIGURE 27. MSA-DP+ double disk failure rebuild Stage 2 is to return the disk group to a fault-tolerant state

If no spare capacity remains and two more drive failures occur, some stripe zones, the disk group, and the pool will all be in a critical state. Depending on how many drives are in the disk group, some or perhaps many stripe zones may still be fault-tolerant. To mitigate the danger of data loss, Rebalance Fault-Tolerant Stripes (REFT) attempts to reallocate chunks from a fault-tolerant stripe zone to another that is critical. However, the same rules for zone distribution still apply, which means a drive cannot donate capacity to a zone that it already participates in.

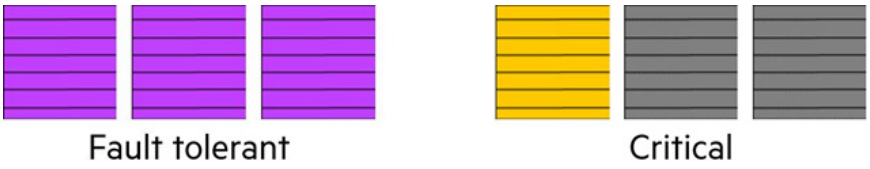

FIGURE 28. Critical stripe zone, disk group, and pool before REFT

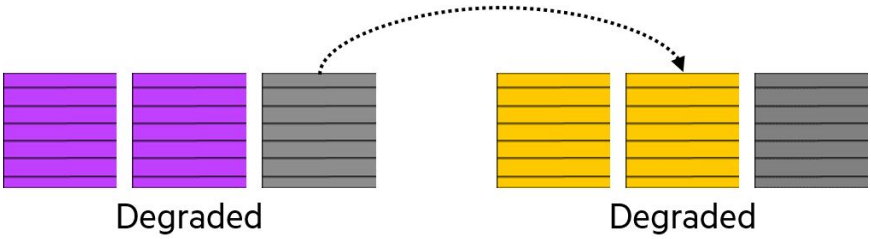

FIGURE 29. Degraded zones, disk group, and pool after REFT

IMPORTANT
MSA-DP+ disk groups will only use REFT if a system administrator has failed to replace failed drives. As per best practices, ensure that notifications are configured and working correctly.

Sequential write optimization

The Power of 2 rule is a highly recommended best practice to ensure that sequential write performance does not degrade as a result of a partial page write when configuring virtual disk groups with either RAID 5 or 6. This rule dictates that the number of disk drives within a disk group holding data chunks rather than parity be a power of 2. Because the maximum number of disk drives supported within a disk group including parity is 16, this means either 2, 4, or 8 data chunks.

TABLE 9. Power of 2 chunk distribution

RAID level	# drives in a disk group	# data chunks in a stripe	# parity chunks in a stripe
RAID 5	3	2	1
RAID 5	5	4	1
RAID 5	9	8	1
RAID 6	4	2	2
RAID 6	6	4	2
RAID 6	10	8	2

RAID distributes data across all disk drives within a disk group through a process known as **striping**. As shown in Figure 30, when a parity RAID scheme such as RAID 5 is used, a stripe contains one parity chunk and as many data chunks as there are remaining disk drives within the disk group. For optimal performance, the array automatically chooses a disk group chunk size of 512 KB when following the Power of 2 rule.

NOTE
Because of the high performance of SSDs, it is not necessary to follow the Power of 2 rule for the performance tier or all-flash pools.

NOTE
An MSA-DP+ disk group incorporates the Power of 2 rule automatically regardless of drive count, and no action is required to maintain conformity with this best practice.

In the example illustrated in Figure 30, each stripe uses two data chunks to store 1 MB of data. A parity chunk consumes the same amount of disk space as a data chunk but does not count toward capacity. This example requires four stripes to write a page. However, the number of stripes needed depends on how many disks there are within a virtual disk group. For example, a virtual disk group with nine disk drives requires one stripe to write a full page.

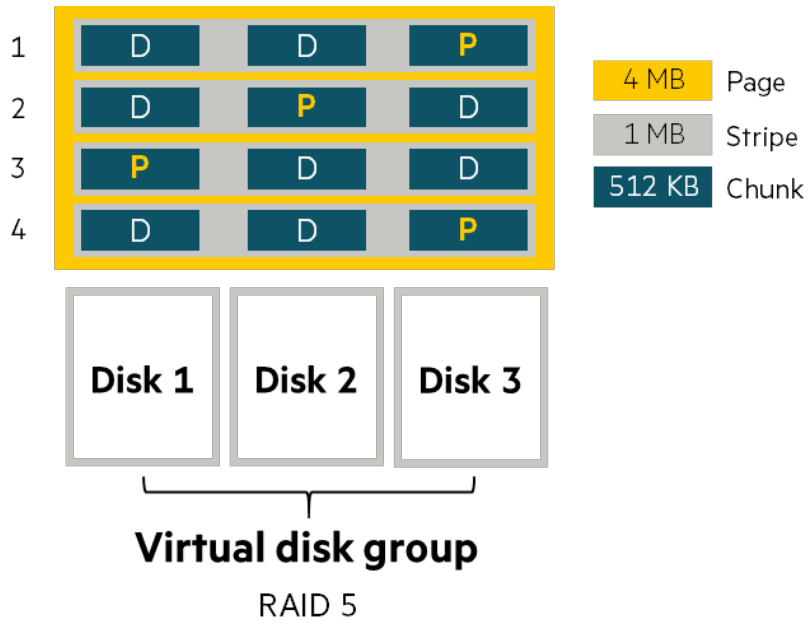

FIGURE 30. Example of how a page is distributed across physical disks when the Power of 2 rule is followed

Disk groups that do not follow the Power of 2 rule incur a performance penalty when writing sequential data, caused by two pages sharing the same parity chunk. As shown in Table 10, each write to a disk group requires additional I/Os to recalculate and write parity. For example, when a single new chunk is written to one disk drive using RAID 5, four I/Os are required:

- Read the original data chunk
- Read the original parity chunk
- Write the new data chunk
- Write the new parity chunk

TABLE 10. RAID write penalties

RAID level	Disk group write I/O	I/O penalty
RAID 1	1	2
RAID 5	1	4
RAID 6	1	6
MSA-DP+	1	6

Page overlap leads to double the number of parity recalculations for shared stripes. The recalculation of parity is not a burden for the HPE MSA RAID ASIC. However, the physical limitations of mechanical hard drives persist, and additional I/Os means more work for the disk drives.

When the Power of 2 rule is not applied, the array tries to reduce the negative performance impact by decreasing the chunk size to 64 KB. This increases the number of stripes required to write a page, but also decreases the amount of data that needs to be read and written when recalculating parity for a share stripe.

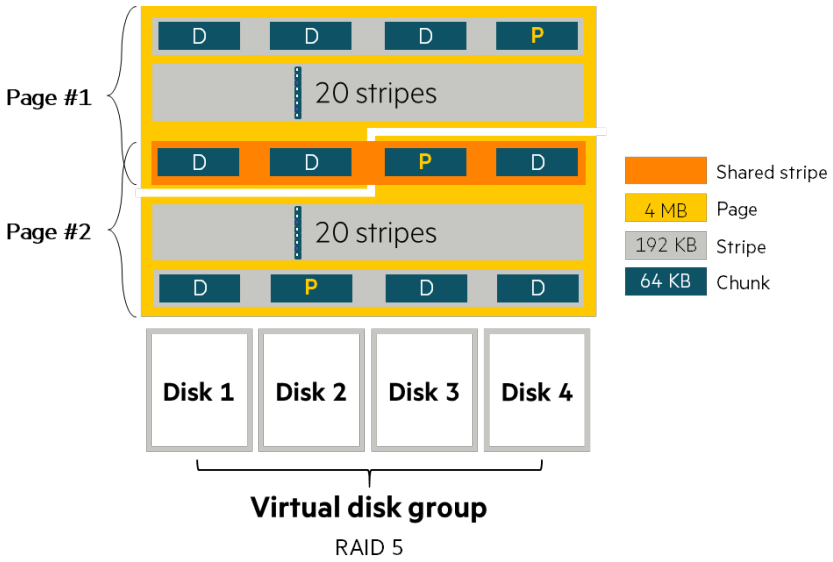

FIGURE 31. Example of how a page is laid out across physical disks when the Power of 2 rule is not followed

Capacity and performance expansion

The capacity of a pool is expanded online by either adding more disk groups or by adding additional drives to MSA-DP+ disk groups. It is not possible to expand any other disk group except for those configured using MSA-DP+. An important consideration is to ensure that when the size of either the standard or archive tier is increased, the proportion of flash storage does not fall below the recommended minimums. Additionally, the more drives there are within a system, the higher the probability of drive failure. It is therefore critical to configure spare capacity for disk groups protected by MSA-DP+ or global spares for all others as per current recommendations found in the HPE MSA 1060/2060/2062 Storage Arrays Best Practices paper.

Figures 32 and 33 illustrate before and after an expansion of a pool by adding additional disk groups. Figures 34 and 35 add additional drives to MSA-DP+ protected disk groups.

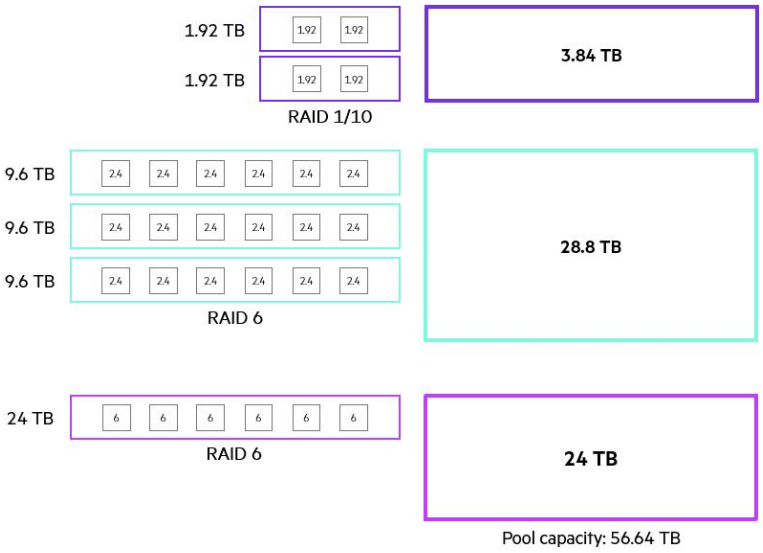

FIGURE 32. Example of a pool layout using traditional RAID disk groups before capacity expansion

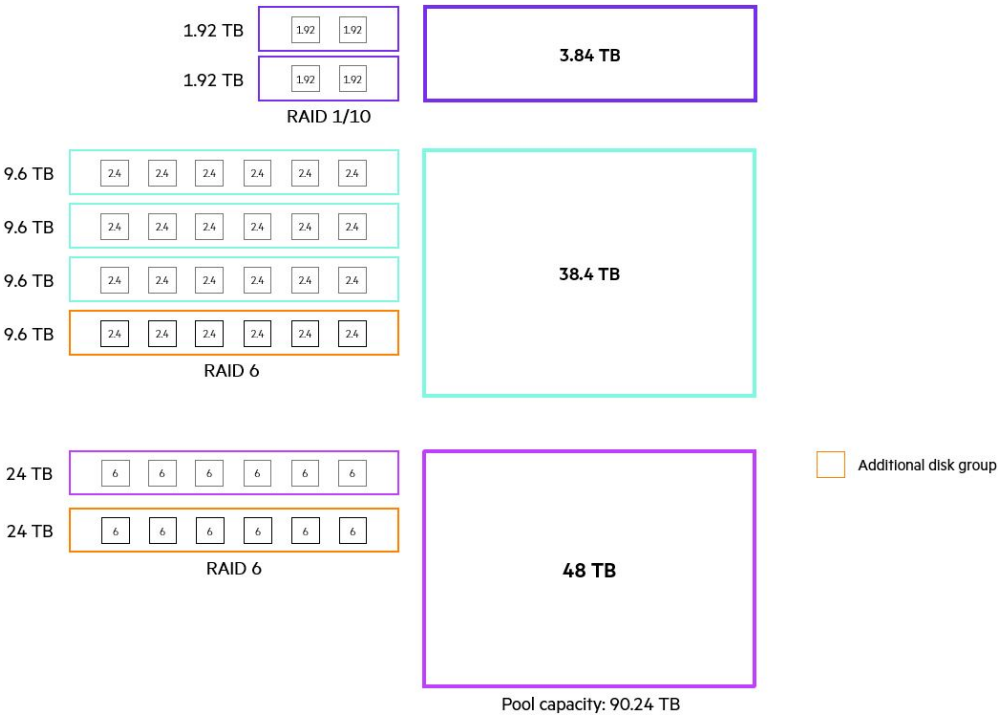

FIGURE 33. Example of pool layout after capacity expansion by adding additional disk groups

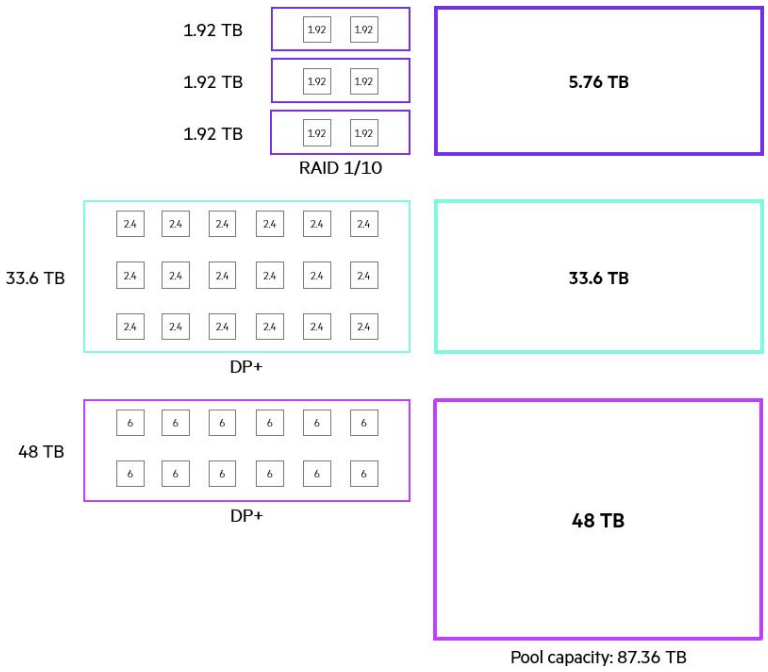

FIGURE 34. Example of pool layout using MSA-DP+ HDD disk groups before capacity expansion

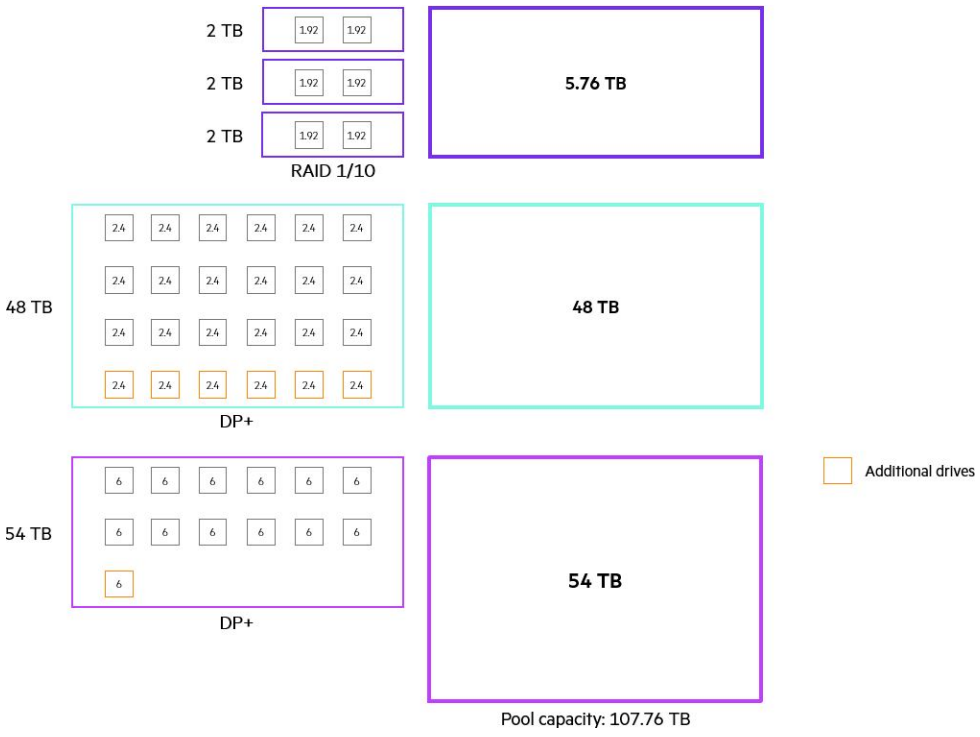

FIGURE 35. Example of pool layout using MSA-DP+ HDD disk groups after capacity expansion

Adding disk groups to a pool both expands available capacity and initiates a rebalance of allocated pages across the tier to which the disk group belongs.

When an MSA-DP+ disk group is expanded, new drives first replenish spare capacity that has been consumed due to a failed drive or to satisfy a new spare capacity target, or both. The remaining capacity is then added to the disk group and allocated stripe zones rebalance across disk group.

Figure 36 demonstrates the expansion of capacity in addition to the increase of spare capacity by adding ten more drives to the disk group.

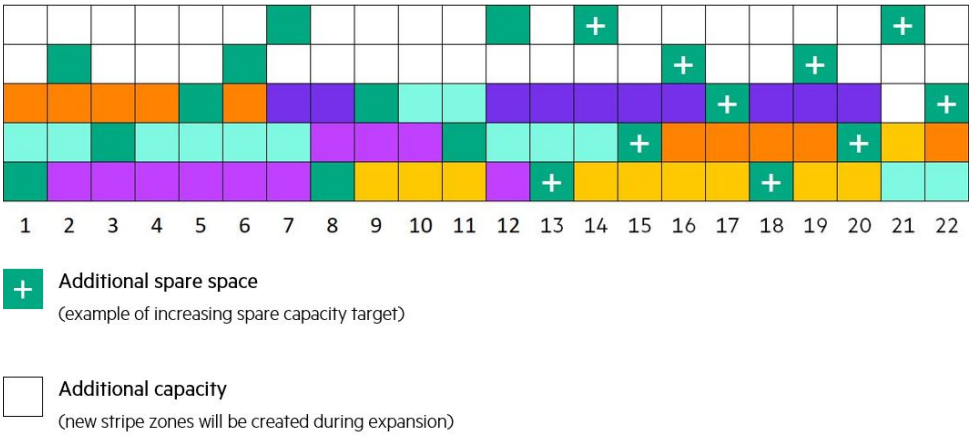

FIGURE 36. Example of an MSA-DP+ disk group layout after expansion and allocation of additional spare capacity

NOTE
It is supported to expand an MSA-DP+ disk group with larger drives, but only recommended if they are not larger than by a factor of two than the dominant drive in the group. For example, expanding a disk group containing 1 TB drives by adding 2 TB drives is supported, but adding 4 TB drives is not recommended.

SSD drive endurance

Another benefit of virtual storage, automated tiering, and SSD read cache is the ability to reduce the wear of SSDs, which have cells that eventually terminally degrade after a certain number of write cycles. Several techniques are used within an SSD to prevent accelerated wear, and some storage arrays take this further by using sophisticated leveling strategies. However, much of the concern around the wear of SSDs stems from an era when memory cell technology was more expensive to produce, and capacities were far smaller. Comparatively, SSDs today typically hold between five and twenty times the amount of data, which means that the likelihood of reaching a drive's physical write limits has declined dramatically. Coupled with the effectiveness of the HPE MSA page ranking algorithms and with wide-striping, it is possible to exclusively use lower-cost RI drives under all workloads in either hybrid or all-flash configuration without introducing a risk of failure.

TABLE 11. Example of SSD drive wear with exaggerated random-write workloads in a hybrid configuration

1 TB total data per day	Performance tier	Standard tier
Share of workload	90% (random)	10% (sequential)
Reads/writes %	10 / 90	50 / 50
Tier configuration	• Four 1.92 TB RI SSDs • One RAID 10 virtual disk group	• 12 x 2.4 TB 10K enterprise SAS • One 12-drive MSA-DP+ virtual disk group
Pool capacity	23.04 TB	
Tier capacity	3.84 TB	19.20 TB
Proportion of pool	17%	83%
Daily writes to tier	810 GB[14]	190 GB
Daily drive writes	405 GB[15]	N/A
Years to drive wear-out	23.7[16]	N/A

[14] 90% of 90%
[15] 21% of DWPD
[16] 20.7 years after warranty expiration

Table 11 demonstrates how both the automated tiering engine and wide striping decouples front-end I/O from the back-end I/O, thus reducing total drive writes. In the example, 1 TB of data is written to and read from the pool in 24 hours. Of the share of workload, 90% of the I/O is random, and 90% of that are writes, which means a single drive absorbs 405 GB of data during that period. Although 405 GB might appear to be a lot, it accounts for only 21% of the 1.92 TB of data that can be written to each drive every day for five years before it would wear out. At this rate, each SSD would have 2.76 PB of unused wear after five years.

If an HPE MSA array were in an all-flash configuration, each drive would absorb more writes compared to a hybrid configuration. However, as shown in the example in Table 12, even with a similar pool capacity and no incremental capacity growth, an unrealistically high amount of daily writes would have to occur before a drive would wear out.

TABLE 12. Example of time until drive wear-out in an all-flash configuration and typical capacity target

Condition	All-flash
Share of workload	100% (mixed)
Pool configuration	• Eight 3.84 TB RI SSDs • One RAID 5 virtual disk group
Pool capacity	26.88 TB
Total SSD lifetime writes	49 PB
Writes to pool (per day) until drive wear-out	• Three years (warranty): 44 TB • Five years (lifetime): 26.88 TB • Ten years: 13.44 TB • Twenty years: 6.72 TB • Thirty years: 4.48 TB

Thin provisioning

Thin provisioning is a well-established technology designed to limit initial monetary investment into physical storage capacity. Thin provisioning is achieved by reporting a requested volume size and required geometry to an operating system. It does not physically allocate pages within the pool. A page is allocated, and physical capacity is consumed only when data is written to a volume. In addition to thin provisioning, virtual storage also provides two mechanisms for returning allocated capacity that is no longer in use back to the pool to be used by other volumes and snapshots. These mechanisms enable a pool to grow and contract as required and result in a conservative approach to storage provisioning:

- The SCSI UNMAP command can be issued by modern operating systems to free LBAs after deleting files from the file system. When a contiguous range of sectors amounting to 4 MB has the UNMAP command applied, the page is released.
- For older operating systems that do not support UNMAP, software tools or full-formatting can be used to write zeroes across the entire volume and, therefore, pages. Pages containing zeroes are detected and released as part of the disk scrubbing maintenance task.

Thin provisioning requires monitoring and adequate planning to ensure that physical capacity is not exhausted. Upon reaching defined or system default utilization thresholds, an HPE MSA storage system will send alerts to defined users. It is possible to disable overcommitment at the pool level, which can assist with managing capacity allocation for base volumes. However, disabling overcommitment also requires careful management because snapshots also require the allocation of capacity. Because snapshot quantities tend to be high and less predictable, the inability to overprovision the pool can easily lead to unexpected capacity consumption.

Thin rebuilds

Because unallocated pages do not consume physical disk space, recovery from a degraded disk group is faster. Except for SSD read cache, if a disk fails or goes offline, the system will begin rebuilding the disk group starting first with allocated pages, thereby minimizing the time needed to bring the system back to a fault-tolerant status.

LICENSING

The HPE Advanced Data Services (ADS) Suite is a single-license SKU for HPE MSA sixth-generation systems. It includes:

- Performance tiering
- 512 snapshots and volume copy
- Remote snapshot replication

NOTE
The ADS Suite is optional for HPE MSA 1060 and 2060 arrays. The HPE MSA 2062 systems include the license along with two 1.92 TB SSDs.

IMPORTANT
If both SSDs and HDDs are used in tiers in the system, even if in different pools, then a license is needed. A license is not needed when SSDs are used as read cache.

Table 13 makes clear when a license is required based on the combination of drives and the model of the HPE MSA array.

TABLE 13. License requirements

Array	Single drive type (HDD or SSD)	Mixing any HDD[17] types within the same system[18]	Coexistence of both a performance tier and any other tier within the same system	SSD read cache
HPE MSA 1060	No	No	Yes	No
HPE MSA 2060	No	No	Yes	No
HPE MSA 2062	No	No	Included	No

DATA PROTECTION

An HPE MSA array provides several technologies in addition to RAID to provide increased resiliency and recovery from failure. This paper does not address the array-based asynchronous replication capabilities of the Remote Snap Replication (RSR) feature. If near-synchronous replication is required, consider Zerto through the HPE Complete Program.

Snapshots

The HPE MSA uses redirect-on-write (RoW) snapshot technology to protect volumes. RoW snapshots are superior to the most common alternative mechanism, copy-on-write (CoW), because it is flexible and eliminates the performance penalty associated with copying data elsewhere in the system.

In taking a snapshot, the system transparently creates a duplicate volume, which becomes the location of all future writes. However, rather than copying data, lookup tables are used by both the base volume and the snapshot to reference shared pages. The system tracks how many volumes reference the page and locates new data accordingly. When there are no snapshots for a volume, the reference count is 1, and data written as usual. However, one or more snapshots of that volume will increase the reference count to two or more, which results in the allocation of new pages from the pool, thus preserving the original data.

[17] 15K, 10K, or 7.2K hard disk drives
[18] The term *system* refers to the HPE MSA array and includes either pool.

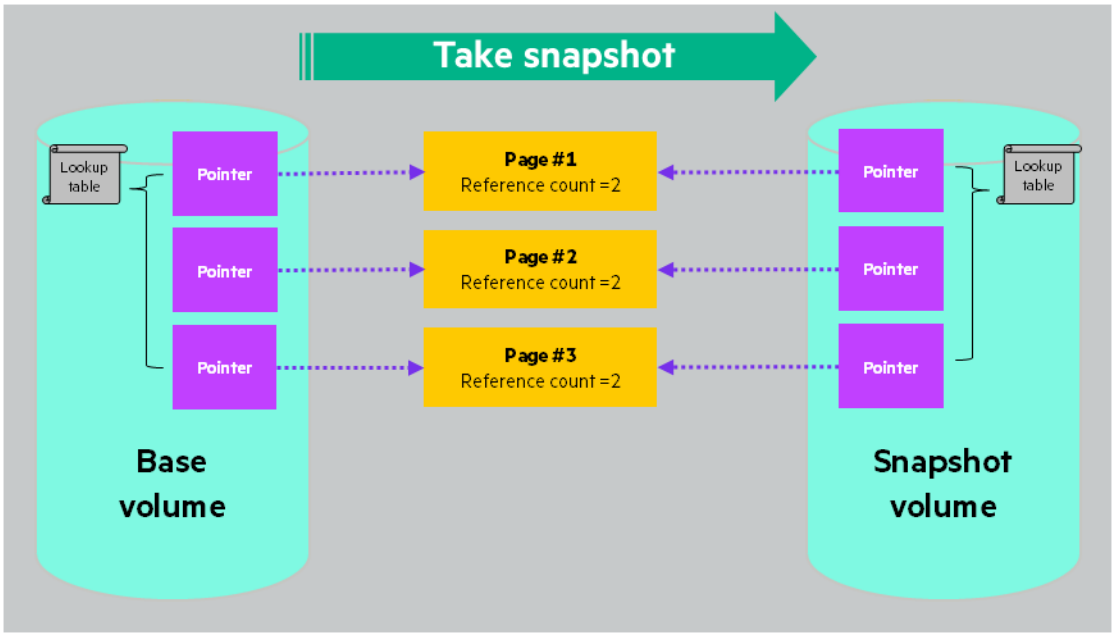

FIGURE 37. RoW snapshot mechanism

CONCLUSION

Virtual storage provides the foundation for a multitude of essential features and capabilities that tackle the challenges faced by small and medium-sized customers. Virtual storage, as found in sixth-generation MSA, alleviates management overhead, and provides advanced technology to reduce both initial and ongoing costs, all while improving availability and performance far beyond all previous generations.

Resources, contacts, or additional links

HPE MSA 1060 Storage QuickSpecs
hpe.com/support/MSA1060QuickSpecs

HPE MSA 2060 Storage QuickSpecs
hpe.com/support/MSA2060QuickSpecs

HPE MSA 2062 Storage QuickSpecs
hpe.com/support/MSA2062QuickSpecs

HPE MSA 1060/2060/2062 Storage Best Practices
https://www.hpe.com/h20195/v2/Getdocument.aspx?docname=a00105260enw

HPE MSA Health Check
www.hpe.com/storage/MSAHealthCheck

Sign up for HPE updates
h41360.www4.hpe.com/alerts-signup.php

LEARN MORE AT

hpe.com/storage/msa

a00103247ENW, July 2021, Rev. 1

Hewlett Packard Enterprise