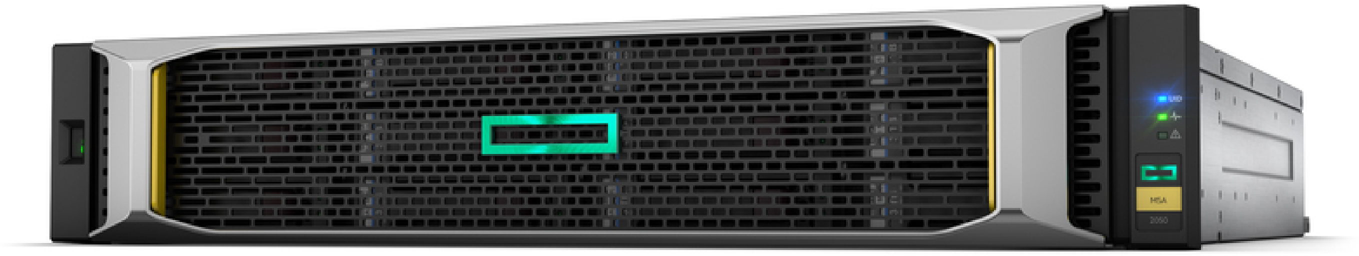
Hewlett Packard Enterprise

Technical white paper

HPE MSA Gen5 virtual storage

Technical Reference Guide



Contents

Executive summary

This white paper is an investigation into virtual storage as implemented on HPE MSA fifth-generation storage systems. It also explores the automated tiering engine and supporting technologies of HPE MSA fifth-generation arrays. This white 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 fifth-generation portfolio, which uses virtual storage exclusively, includes the HPE MSA 1050, 2050, and 2052 SAN and SAS storage arrays.

Note
HPE MSA third-generation arrays used linear storage; fourth-generation arrays used linear and virtual storage. HPE MSA fifth-generation arrays are the first generation in the portfolio to use only virtual storage technology.

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 fifth-generation array, use the links on the last page of this document to go to the core documentation.

Intended audience

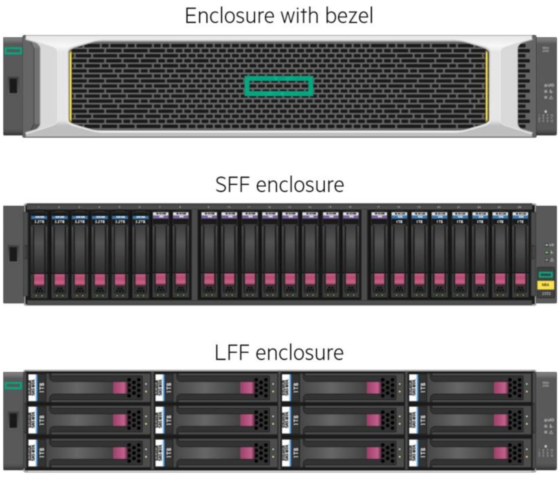
This white paper is for everyone involved in the design and implementation of storage solutions that include HPE MSA fifth-generation arrays. Technical sales staff tasked with designing an effective solution will benefit from an understanding of the HPE MSA architecture, and as a customer-installable product, so will the administrator that eventually configures it. HPE recommends a current understanding 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])
- 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 bezel

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



Enclosure with bezel

SFF enclosure

LFF enclosure

Figure 1. HPE MSA enclosures

[1] Small form factor 2.5"/large form factor 3.5"

HPE MSA 1050 arrays have a total of four host ports compared to the eight of the HPE MSA 2050 and 2052. However, HPE MSA 1050 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. HPE advises using fan-out cables even if they are not initially needed to avoid interruptions when connecting additional hosts.

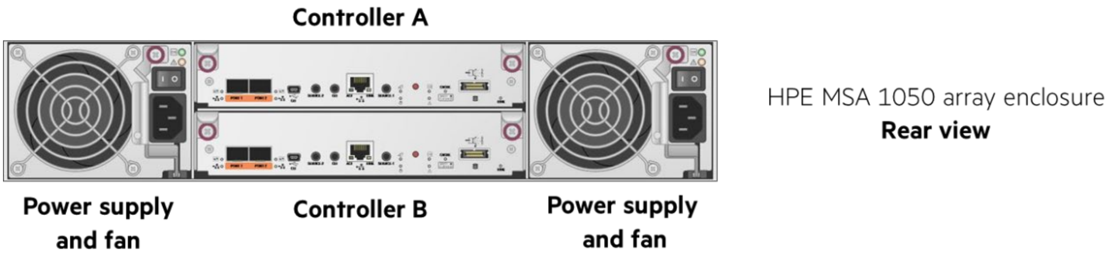


Figure 2. Rear view of an HPE MSA 1050 array enclosure

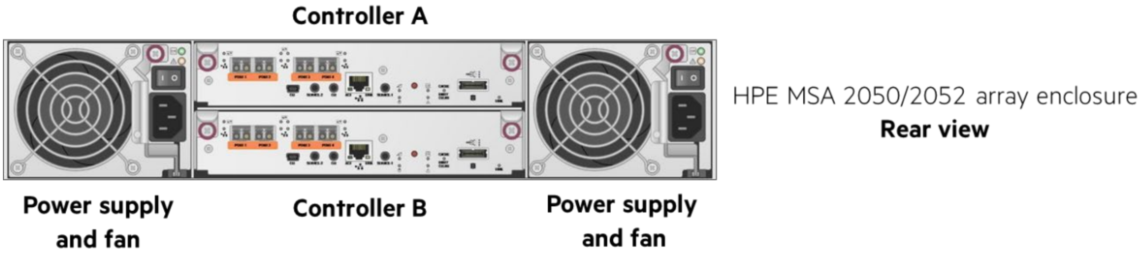


Figure 3. Rear view of an HPE MSA 2050/2052 array enclosure

The HPE MSA 1050 supports up to three expansion disk enclosures totaling either 48 LFF drives or 96 SFF drives. HPE MSA 2050 and 2052 arrays support up to seven expansion disk enclosures and up to 96 LFF drives or 192 SFF drives. Both arrays support the HPE MSA 2040 LFF enclosure and D2700 SFF enclosure in upgrade scenarios. Refer to the Upgrading to HPE MSA 1050/2050/2052 white paper for more information.

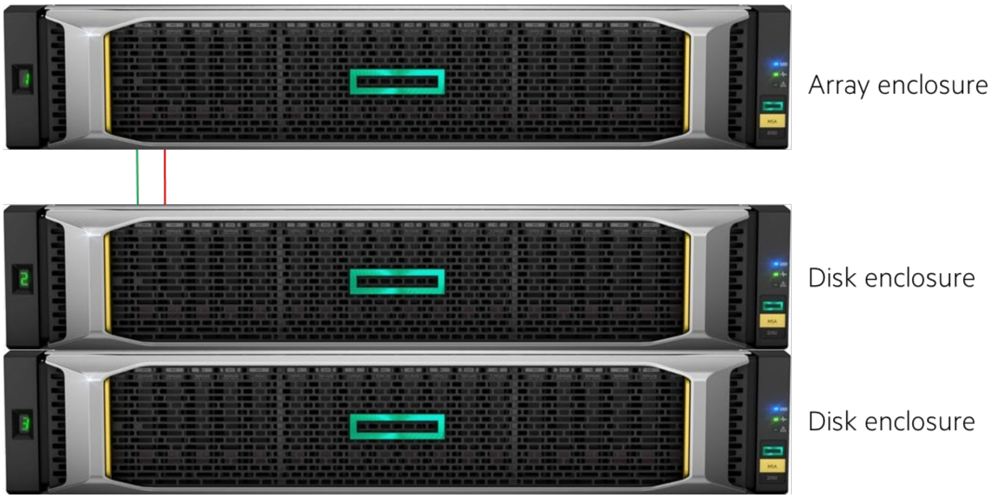


Figure 4. HPE MSA array enclosure naming

Controller architecture

HPE MSA arrays are designed to 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, Serial over USB)
- Storage controller
- Management controller
- Memory/cache
- Internal backup power (supercapacitor)
- Removable nonvolatile memory card (CompactFlash [CF] Express)

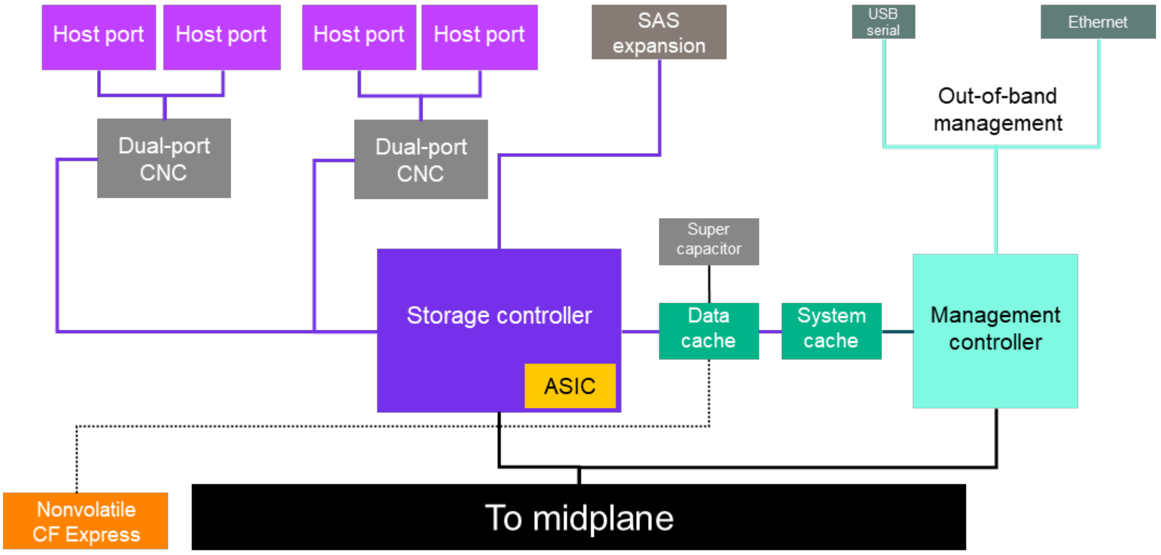


Figure 5. Simplified HPE MSA controller diagram

Governing each HPE MSA controller are two logical components, the storage controller (SC) and the management controller (MC). The controllers are composed of numerous subsystems that provide supporting connectivity and features independent of each other.

The storage controller is the fundamental function and 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 SMU runs over Ethernet with out-of-band management to 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 the array 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 and does not offer a single, highly available IP address.

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, and is assigned to several different system functions. As shown in Table 1, an HPE MSA 1050 array has a total of 12 GB, and an HPE MSA 2050/2052 has 16 GB of cache allocated to various tasks.

Table 1. Assignment of system cache.

Purpose	HPE MSA 1050 per controller	System total	HPE MSA 2050/2052 per controller	System total
Total cache	6 GB	12 GB	8 GB	16 GB
Local read	1 GB	2 GB	1 GB	2 GB
Local write	1 GB	2 GB	1 GB	2 GB
Partner controller mirror-read	1 GB	2 GB	1 GB	2 GB
Partner controller mirror-write	1 GB	2 GB	1 GB	2 GB
Operating system overhead	2 GB	4 GB	4 GB	8 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 a dedicated ASIC for RAID to offload the general-purpose CPU, freeing it to process other tasks such as tracking metadata. By reducing the resources required for RAID, less cache is needed to counterbalance a higher load. To accurately represent an array's sustained performance capabilities, HPE testing intentionally overwhelms controller cache to eliminate misleading and short-lived, cache-bound benefits. These results match mathematical models used for performance estimation by tools such as the HPE Storage Sizing Tool 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 in the event of controller failure I/O can continue through the remaining controller and without data loss.

Should external power be removed from a controller, an internal supercapacitor provides enough energy to write the contents of controller cache to a removable nonvolatile CF Express memory card, allowing for the long-term retention of data not yet committed to disk. The restoration of external power then flushes the unwritten cache contents to disk.

SAS

An HPE MSA array communicates with all drives and expansion enclosures through 6 Gb SAS. SATA drives are not supported. However, an HPE MSA supports up to 12 Gb SAS connectivity to hosts, and as of August 2019 an HPE MSA is the only HPE array to offer block-based shared storage over the SAS protocol. External host traffic is entirely separate from internal SAS traffic and although they share the same protocol, they are otherwise unrelated.

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

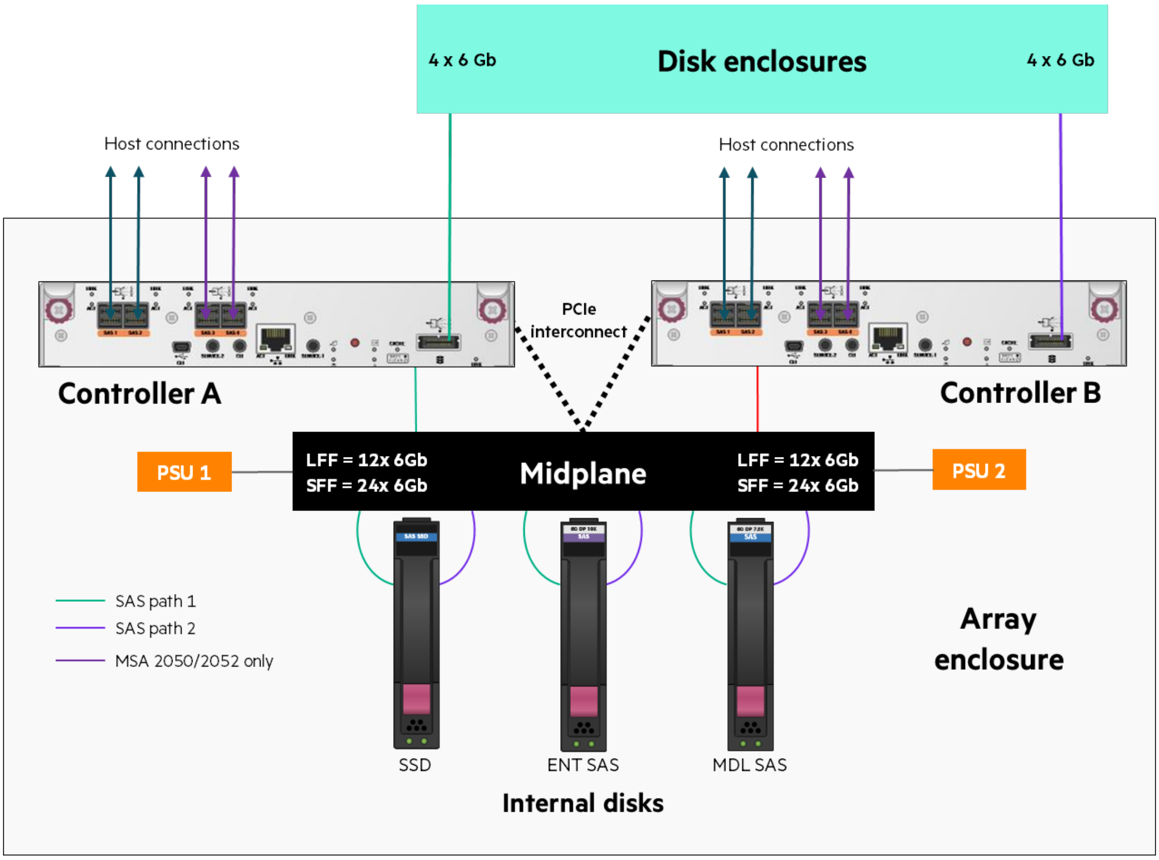


Figure 6. HPE MSA array architecture

Important
For optimal performance, locate SSD drives within the array enclosure that has a dedicated SAS lane for each drive, compared to external enclosures that connect to an array controller through a four-lane SAS connector. Although additional lanes do not result in multiples of performance, they do reduce overall link latency, thus improving overall performance.

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 fifth-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 as a “pool.”

RAID is a type of virtualization that aggregates multiple disk drives into a single object. HPE MSAs used this linear concept 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 this 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 this architecture is that it is not possible to increase the capacity nor 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 there are and the larger their capacity, the greater the risk of concurrent disk 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 something 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 also by aggregating entire disk groups, thus allowing data to be wide-striped across all disk groups 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 provides 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 improvements of virtual storage also stretch 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, each supporting one pool. A pool becomes available when a disk group is assigned to it.

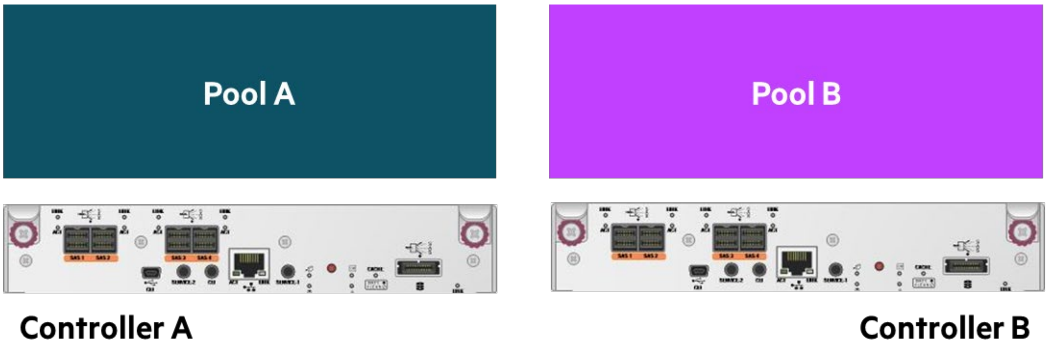


Figure 7. HPE MSA controllers and pools

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. A second controller should be reintroduced as soon as possible.

Important

HPE does not sell fifth-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 nor otherwise utilize 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 ULP mechanism. In the event of firmware update or controller failure, a pool will remain online while the remaining controller takes temporary ownership of it.

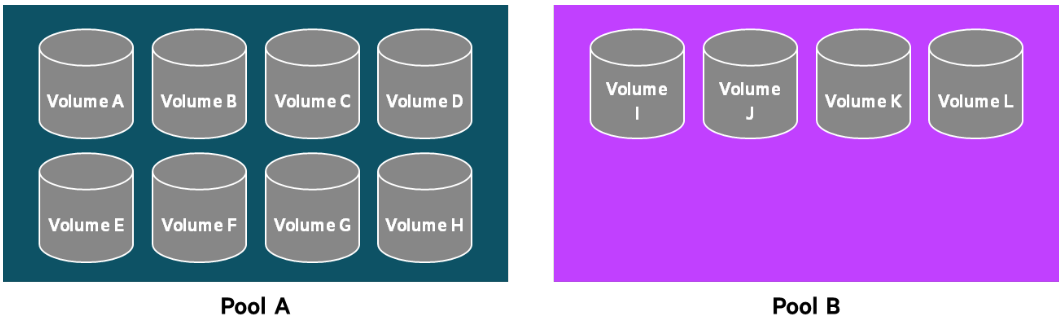


Figure 8. Relationship of volume and pools

It is supported and potentially advantageous only to provision one pool because this allows for consistent performance during a firmware upgrade or controller failure[2]. 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 100,000 IOPS[3], which is that of one controller. 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.

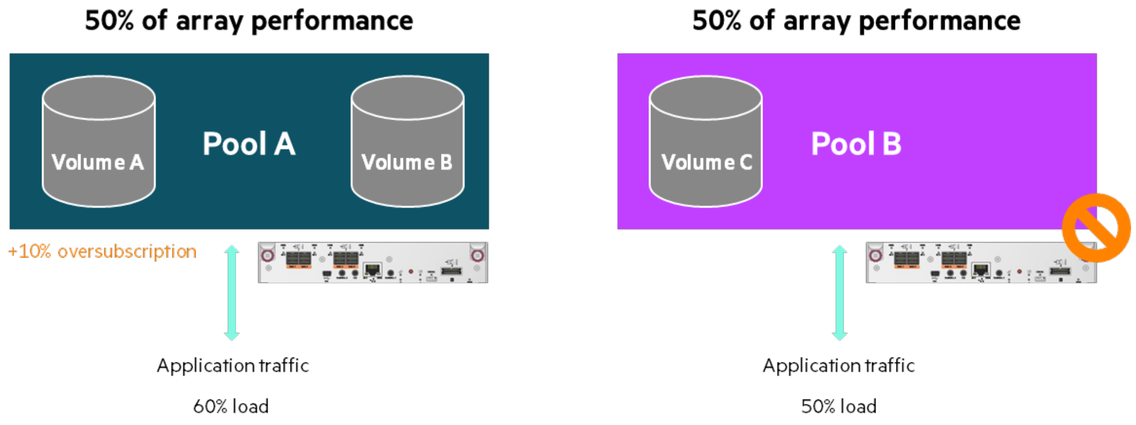


Figure 9. Oversubscription of Controller A during the unavailability of Controller B

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

1. **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 the most commonly 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 fifth-generation array performs so well that it is uncommon to consistently consume more than 50% of the potential of the array. As a result, there will almost always be headroom, even if it is not explicitly managed.
2. **One pool, with 100% headroom.** Provisioning a single pool and placing all workloads in it ensures that performance will be consistent even if a controller becomes unavailable. This approach relies on the storage solution being sized appropriately to the requirements, which can be difficult using today's sizing tools[4] because they do not contain performance modeling for an asymmetrical pool layout. Nevertheless, simple mathematical division when comparing array performance maximums within the QuickSpecs and total disk group count within the HPE storage sizing tool makes this possible, although not precise.

   When there is a fixed capacity goal less than 128 TiB[5], a single pool also allows a host to access the total capacity as a single volume if the operating system file system supports it. This approach might be crucial for a small configuration where there might only be a single data store in use by a hypervisor.
3. **Two pools, with 100% headroom.** Similar to the second option, this methodology 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.

   Two pools are simpler to size for performance because data exists for a symmetrical pool layout. However, over time, keeping within this headroom can be challenging. For example, as a pool's capacity expands, workloads can grow and consume the additional performance without the administrator realizing the problem. When this happens, headroom will decline.

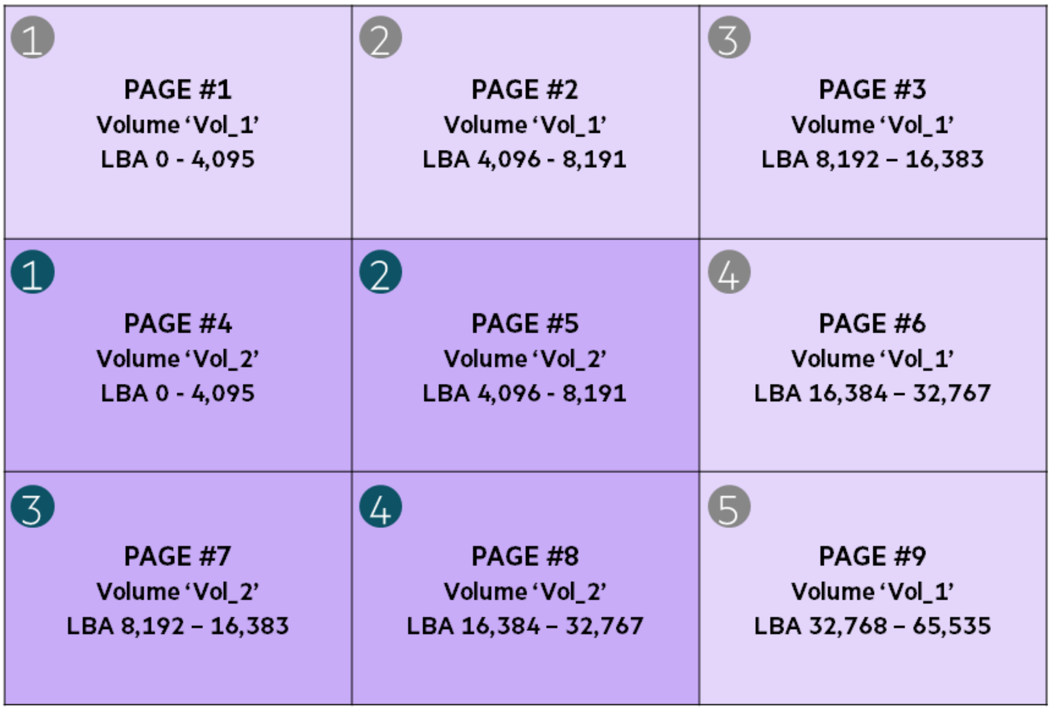
Important

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.

[2] As of August 2019, the best practices white paper states that pools should be provisioned symmetrically. That white paper will be updated to agree with this document.
[3] Example only. Real performance may be higher or lower and varies due to drive configuration and workloads.
[4] HPE Storage Sizing Tool
[5] Maximum virtual volume size as of August 2019with firmware VE270 (MSA 1050) or VL270 (MSA 2050/2052).

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.

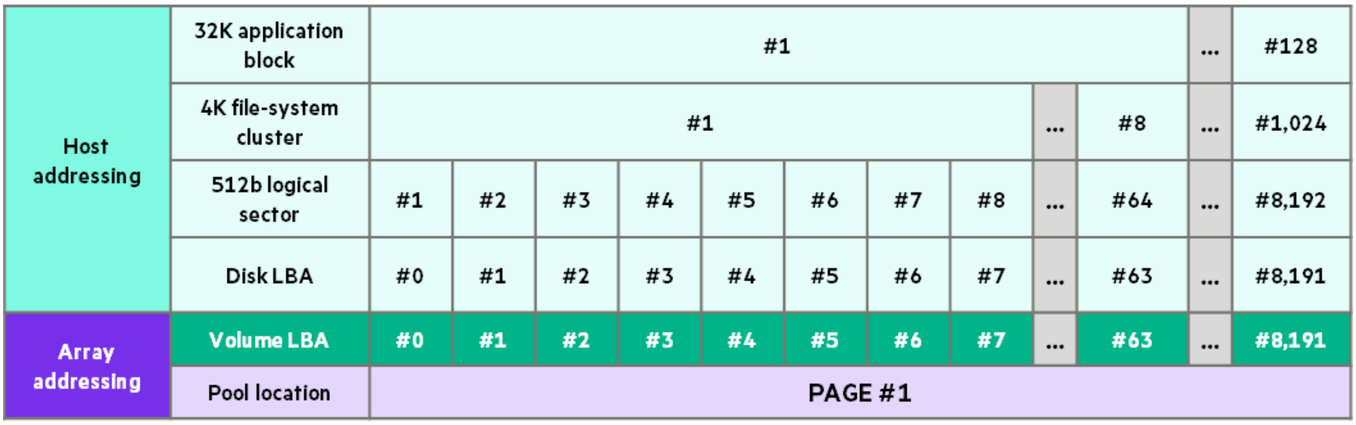
Within an HPE MSA, a volume is a collection of pages amounting to a defined capacity. The location of a volume's designated pages within the pool changes over time and is dependent on several factors such as the use of tiering, snapshots, and thin provisioning. As shown in Figure 10, it is practical 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.



1 PAGE #1 Volume 'Vol_1' LBA 0 - 4,095	2 PAGE #2 Volume 'Vol_1' LBA 4,096 - 8,191	3 PAGE #3 Volume 'Vol_1' LBA 8,192 – 16,383
1 PAGE #4 Volume 'Vol_2' LBA 0 - 4,095	2 PAGE #5 Volume 'Vol_2' LBA 4,096 - 8,191	4 PAGE #6 Volume 'Vol_1' LBA 16,384 – 32,767
3 PAGE #7 Volume 'Vol_2' LBA 8,192 – 16,383	4 PAGE #8 Volume 'Vol_2' LBA 16,384 – 32,767	5 PAGE #9 Volume 'Vol_1' LBA 32,768 – 65,535

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, an operating system will 'see' a physical disk containing a quantity of 512-byte sectors on which a file system is laid out. Applications will typically interact with the file system to read and write data, and their data granularity is often courser than that of the file system. In this example, the HPE MSA will read and write 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 sample 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 distributes the contents of a volume over multiple virtual disk groups within a tier as well as over multiple drive classes that are 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

Important

Disk drives are assigned to an appropriate tier automatically when assigned to a disk group, and it is not possible to change this assignment. For example, it is not possible to allocate 15K disk drives to the performance tier.

A tier is specific to a pool and includes all disk groups of a particular drive type associated with that pool. In Figure 12, you can see how pages of all volumes within a pool are distributed across all three tiers and all six disk groups that are associated with it.[6]

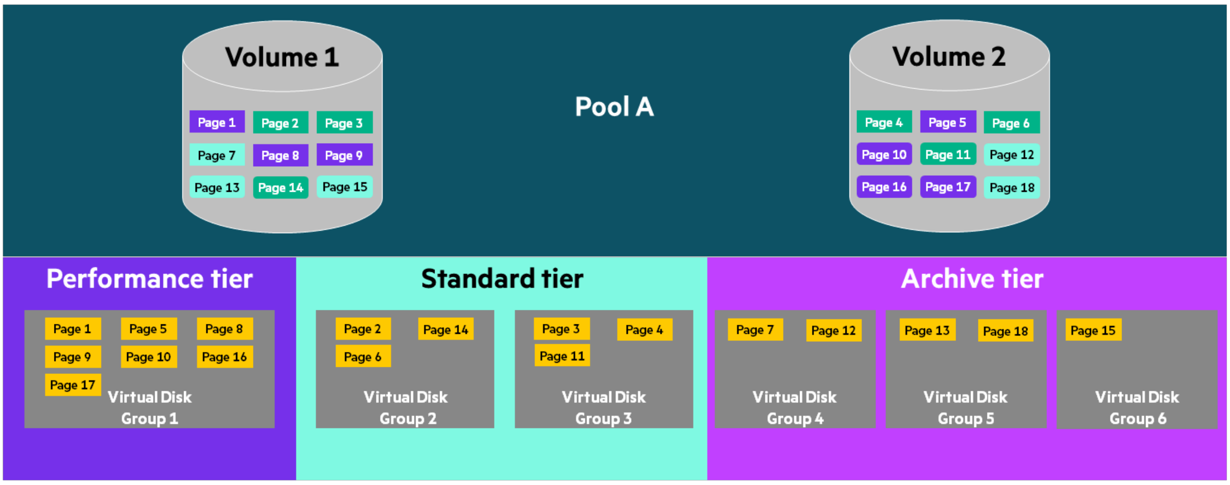


Figure 12. Example of volume distribution across multiple tiers within a single pool

Table 3 lists the four ways to configure a pool.

Table 3. Pool configurations

Configuration	Layout	Advantage
Single tier	Single drive class, wide-striping only, and no automated tiering	• Great for high capacity and sequential I/O centric configurations • Consistent performance for all volumes
Performance tiering	Tiering to SSD; can be one of two recommended layouts: • Two tiers: Performance and standard tiers[7] • All three tiers	• Hybrid array, ideal for mixed workloads and a hands-off approach • Adds to useable capacity
Archive tiering	Standard and archive tiers	Great for high capacity solutions with low-performance requirements
SSD read cache	SSDs used as read cache extension together with either a single tier or archive tiering	Greatly accelerates random reads with a minimum investment, and does not require a license

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

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

For mixed workloads, flash storage is ideal because it provides exceptional performance, although flash is more expensive than traditional HDDs. HPE supports configuring an HPE MSA as an all-flash storage array, but there are more suitable products within the HPE portfolio for scenarios requiring such a high level of guaranteed random I/O performance. Instead, hybrid storage arrays consisting of both SSDs and HDDs are a far more effective route to achieving improved random and sequential performance, without drastically overshooting a capacity goal or budget.

Table 4. Methods of increasing application performance

Method	Advantages	Disadvantages
Add more disk drives to a disk group	• Improves performance for a disk group	• Negatively affects availability for the disk group and subsequently the pool • Only improves performance for that disk group and a portion of a tier • Consistent performance requires that the action be carried out for every disk group within a tier. Because it is not possible to expand a virtual disk group, the disk group must be removed, resized, and reintroduced to a pool. It also requires sufficient unallocated capacity within a pool to absorb the temporary loss of its capacity. • Not cost-effective • Disruptive
Add more disk groups to a tier	• Is a perfect solution for 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 may be necessary to over-provision capacity to achieve a performance goal, particularly for random I/O. • Potentially expensive
Switch to a more performant drive technology	• A guaranteed solution to increase performance up to the point of controller saturation or drive count limit • Suitable for uncompromising application requirements	• Can be extremely expensive to reach both capacity and performance goals at the same time • Tiering is used temporarily 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 2052 • Disruptive, especially if using a backup and restore approach
Automated tiering	• Provides 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 1050/2050 • Can be overwhelmed by sustained workloads on improperly provisioned pools
SSD read cache	• Is very cost-effective • Requires only 1 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 recommended 10% of pool capacity

The HPE MSA automated tiering engine is responsible for ensuring the optimal location for data, including when 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:

1. Results in more back-end I/O being required to make space for new data
2. Leaves less room for “hot” read data
3. Does not take advantage of disk drive types better suited to sequential I/O

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 on, and also frees capacity within the higher performance tiers for more relevant data. As shown in Figure 13, when using performance tiering, the engine detects inbound data written as a stream and actively directs those writes to the fastest tier comprising traditional hard disk drives. Sequentially written data is well suited to mechanical disk drives, whereas random writes are 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.

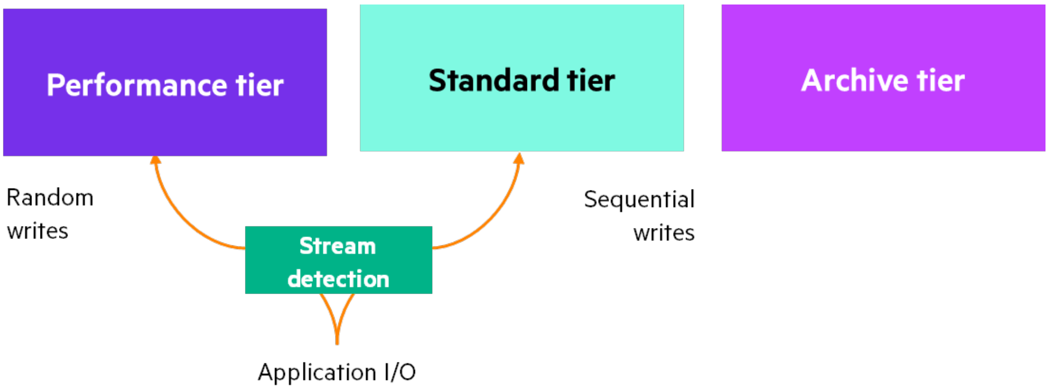


Figure 13. Automated redirection of inbound writes when using performance tiering

When using archive tiering, all inbound writes will 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 white paper for more information.

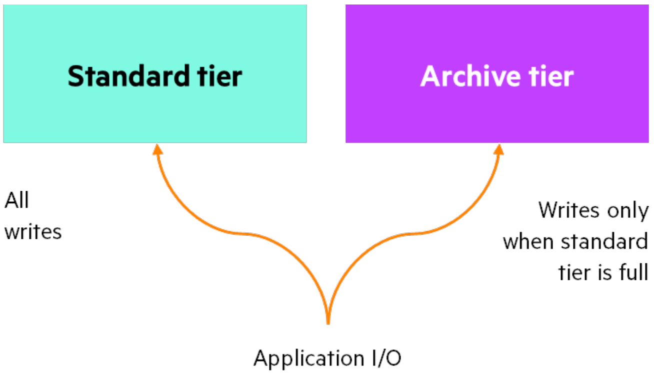


Figure 14. Archive tiering write mechanism

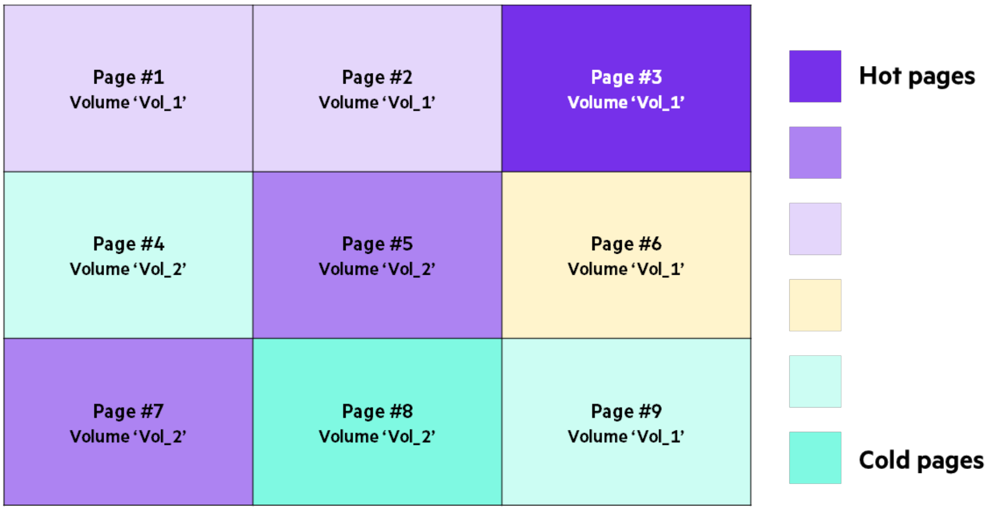
Unlike some tiering mechanisms, an HPE MSA does not schedule page migration to occur on a daily or hourly basis. It also does not require large quantities of expensive low-latency memory or cache to buffer uncoordinated back-end writes. Instead, an HPE MSA tiering engine uses minimal resources, which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

With a frequency of every five seconds, the tiering engine will analyze the pool and promote or demote pages according to how in demand they are in a process known as page ranking. Page ranking is a bid to keep the most active "hot" pages on the fastest tier within a pool by maintaining a list of those which are most accessed. In the same process, pages which have become "cold" due to infrequent access, can be demoted to a less costly and lower performance tier so to make room.

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 the hottest data is the most frequently accessed pages. As a page becomes colder because more active pages are moving up the ranking table, the pages will eventually be evicted from the ranking table entirely until they are accessed again.



Page #1 **Volume 'Vol_1'**	**Page #2** **Volume 'Vol_1'**	**Page #3** **Volume 'Vol_1'**
Page #4 **Volume 'Vol_2'**	**Page #5** **Volume 'Vol_2'**	**Page #6** **Volume 'Vol_1'**
Page #7 **Volume 'Vol_2'**	**Page #8** **Volume 'Vol_2'**	**Page #9** **Volume 'Vol_1'**

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 outside of what disk drives are used.

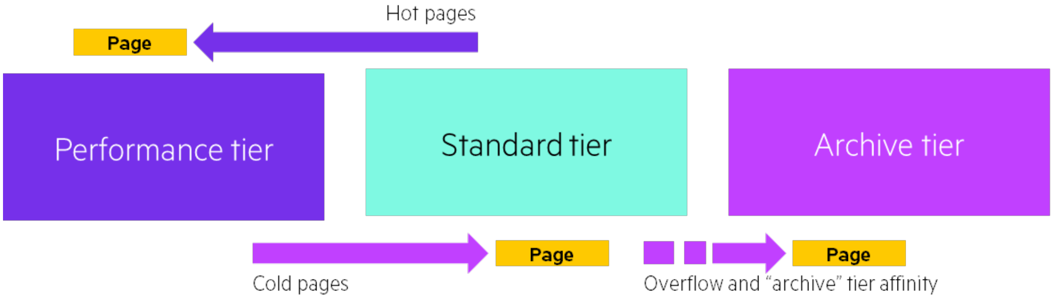


Figure 16. Performance tiering behavior

The migration of a page requires that it be read and then rewritten to another virtual disk group, which consumes performance. Under regular operation, several mechanisms are in place to ensure that reallocations do not negatively impact performance. The only time these rules are circumvented is during the removal of a virtual disk group from a pool. Provided there is sufficient capacity remaining in the pool, removing a disk group will drain pages allocated to it to the remaining disk groups.

General rules are:

- A page will not migrate more often than every 15 minutes.
- Page moves are throttled to not move more than a fixed number of times within a five-second interval.
- 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 the array is configured and workloads are established, daily capacity usage can be evaluated by using the in-array I/O Workload graph. Before production, a safe ratio before an array's installation is between 10% and 15% of the total pool capacity. Although this process might yield positive results under some workloads, HPE recommends that a pool not be configured with a performance tier less than 10% the total pool capacity. If less than 10%, performance would likely become less predictable over time as the pool fills with data. Conversely, increasing the performance tier beyond these recommendations will 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 or linear storage might want to pin a volume to a specific tier, such as 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, it is unnecessary to do this. 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 cripple 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[8]. The ability of an HPE MSA storage system to use automated tiering to achieve high performance with database and Microsoft® Exchange™ workloads was the subject of a Demartek investigation, which resulted in the following reports:

- Accelerating Exchange Server with the HPE MSA 2050 Storage with Built-in Flash
- Accelerating Database Workloads with the HPE MSA 2050 Storage with Built-in Flash

Nevertheless, in some scenarios, the ability to locate a volume within a specific tier provides advantages, and the Tier Affinity setting provides a mechanism that attempts to do so.

Tier affinity is a per-volume setting that can be applied at any time from within the SMU. Its function is to modify the default page ranking algorithm so that pages for that 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 • Overflow[9] = 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.

[8] Provided the proportion of SSD within the pool is appropriately sized
[9] 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.

Note

Changes to the tier affinity setting do not produce measurable results immediately because the setting does not act on existing pages until the page is ranked for promotion (performance affinity) or is needed to make space in a higher tier (archive affinity).

The No Affinity setting is the system default and the recommended setting for most volumes. The No Affinity setting attempts to locate the hottest pages on the highest tier, and all other pages on the next highest tier with free capacity. With three tiers configured, the archive tier remains unused until no capacity remains on the tiers above it because this would otherwise needlessly reduce performance. The archive tier is first accessed for page overflow and is used to store only the coldest pages or new writes when there is nowhere else to write data.

The Performance tier affinity setting is an attractive setting for any volume but should be applied sparingly. The typical application of the Performance setting is for application data frequently in need of flash performance but that has extended periods between heightened read activity. In a balanced system, it should not be possible to migrate the contents of all volumes to the performance tier via the Performance tier affinity setting because there would be insufficient capacity, which reverses its effectiveness. If all volumes are required to be in the performance tier, it would be more appropriate to have a dedicated all-flash pool or to consider an all-flash array from elsewhere in the HPE Storage portfolio.

More often, the most useful tier affinity setting is Archive, which allows the use of the archive tier even when capacity remains in the other tiers. As 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. As examples, data such as disk images, test data, or CCTV video streams are 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.

SSD read cache

As a rule, performance tiering is the best route 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 towards 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 hard disk drives 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:

- Minimum of one drive per pool (NRAID)[10]
- Maximum of two drives per pool (RAID 0)[11]
- A maximum addressable read cache capacity of 4 TB per pool
- Cannot be used with performance tiering within the same pool
- One pool configured with SSD read cache and another as a capacity disk group and performance tiering is supported

Important

Ideally, SSD read cache should account for 80% or more of typical I/O within 24 hours. After the array is configured and workloads are established, daily capacity usage can be evaluated by reviewing the in-array I/O Workload graph. Before production, a safe rule is that SSD read cache should be 10% to 15% of the total pool capacity. However, HPE recommends that SSD read cache not total less than 10% the total pool capacity. If less than 10%, performance would likely become less predictable over time as the pool fills with data. Conversely, increasing the SSD read cache beyond these recommendations will increasingly devalue the cost savings made possible by the SSD read cache mechanism.

With a maximum addressable capacity of 4 TB per pool, the most practical SSD to use in today’s HPE MSA portfolio is the 1.92 TB read intensive (RI) SSD. Two RI drives provide 3.84 TB of SSD read cache, allowing for up to 38.4 TB of capacity per pool while maintaining the recommended ratio of flash to mechanical disk drives.

Figure 17 illustrates how the array reads data when using SSD read cache. However, except for Step 4a, the process is the same regardless of the pool’s configuration. The array will always attempt to complete a read I/O using the fastest storage medium holding a copy of the data.

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

[11] RAID 0 is only applicable to read cache disk groups.

When reading from a disk (as shown in Step 4b), SSD remains a possible source drive type when using performance auto-tiering instead of SSD read cache.

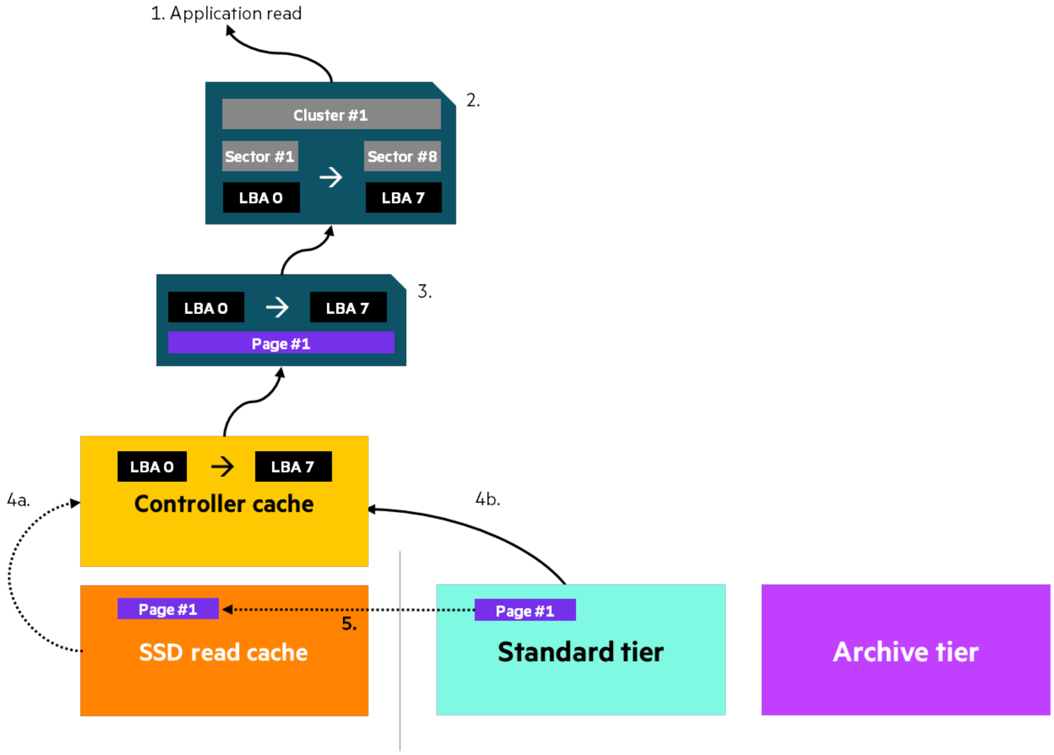


Figure 17. Read mechanism

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, the data must be read from disk:
    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 the page is not in SSD read cache or if SSD read cache is not configured, then the LBAs will be read from the virtual disk group that holds the page, including from the performance tier if configured.
5. Virtual storage uses 'hints' from controller cache regarding whether to copy the page into the SSD read cache. For example, a page read randomly 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 are 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 white paper is to explain HPE MSA technology well enough to understand those best practices. For specific details, review the HPE MSA 1050/2050/2052 Best practices white paper.

Virtual storage uses a controller's general-purpose CPU to track page metadata including the location on disk groups, whereas RAID is a disk group level function responsible for the distributing of a page across individual disks using a specialized ASIC. These functions are carried out independently of each other and are not directly related. 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[12]
RAID 1	Mirroring	Excellent random read/write performance	Performance
RAID 10	Mirroring and striping	Same performance benefits as RAID 1 but allows for up to 256 drives per pool compared to 16 maximum drives with RAID 1.	Performance
RAID 5	Distributed parity (D+P)	Great for sequential performance	Performance Standard
RAID 6	Advanced Data Guarding (D+P+Q)	Great availability Recommended for drive capacities more than 2 TB	Standard Archive

Figure 18 illustrates the relationship of a physical disk, disk group, a tier, and a pool. Note that each disk group can support a different number of drives and RAID protection scheme, ideally based on what is appropriate for a drive type, capacity, and performance characteristics.

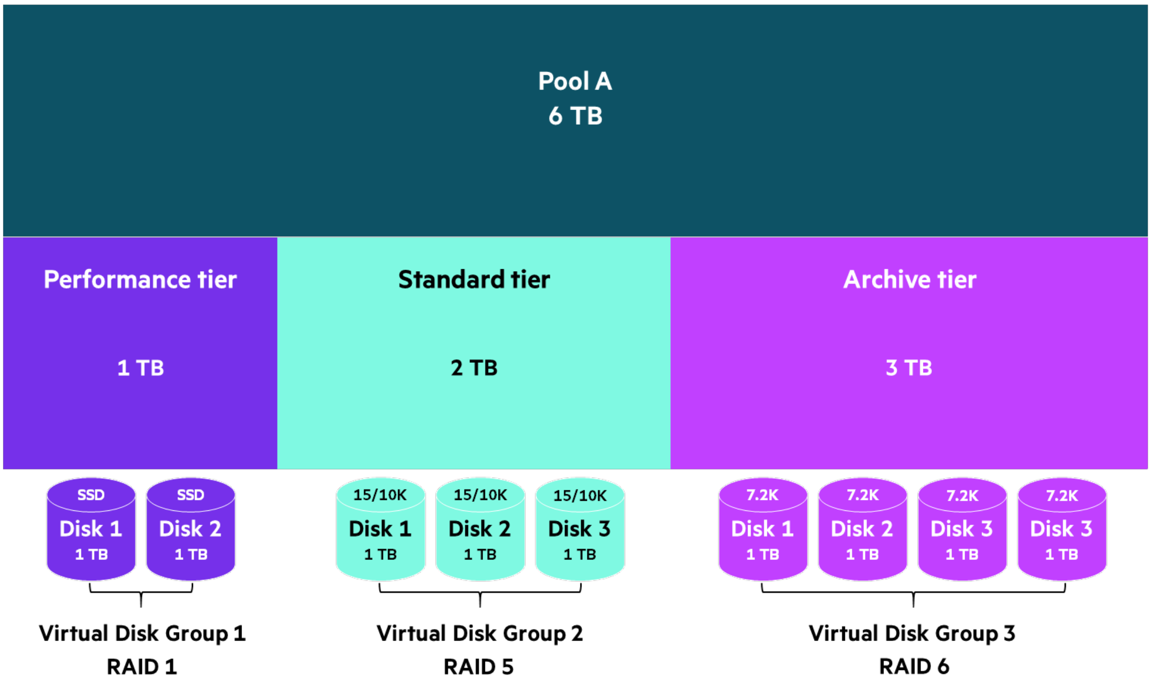


Figure 18. Representation of a multi-tiered pool layout

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 reducing the effects of latency and by multiplying data transfer rates.

[12] Based on drive capacity and potential rebuild times.

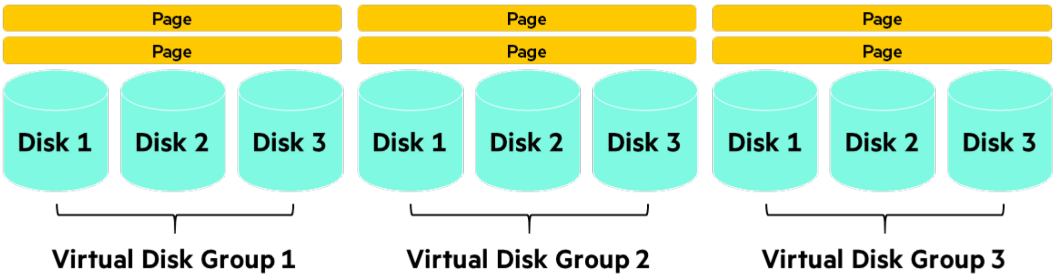


Figure 19. Wide-striped page distribution

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.

However, the longer the platter takes to revolve, the longer I/O must wait to complete. The time it takes for a head to move to the required track and the disk to rotate so that it is aligned with a sector is known as the *access time*. Access times translate into latency, and the higher the latency, the more slowly an application will perform. Random I/O is susceptible to high latencies, but sequential I/O is more tolerant. When a sector is accessed, data is read or written as a stream and not usually subject to further dramatic head movements. Additionally, if they are not disabled, controller cache prefetching algorithms increase sequential read performance by caching the next range of data ahead of time from the next disk.

Figure 20 uses orange arcs 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.

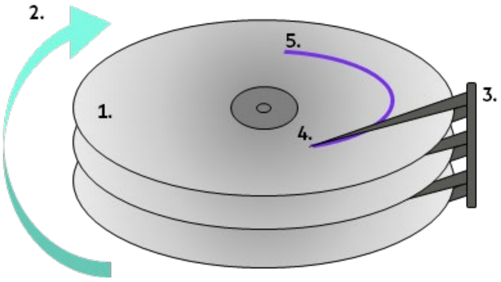


Figure 20. Internal view of a disk drive processing a sequential read/write operation

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 read or written
5. Last sector where data is to be read or written

In contrast, data accessed randomly requires a great deal of physical movement of both read/write heads and the platter, leading to longer access times.

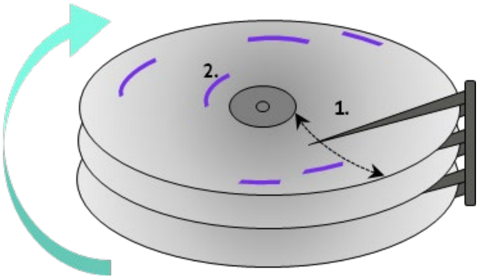


Figure 21. Internal view of a disk drive processing a series of random read/write operations

Figure 21 shows:

1. Movement direction of read/write heads
2. Randomly accessed data

Sequential write optimization

The Power of 2 rule is a highly recommended best practice to ensure sequential write performance does not degrade as a result of a partial page write when configuring virtual disk groups with parity RAID. 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 is 16 including parity, this means either 2, 4, or 8 data drives.

Table 8. 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 18, 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 allocates a 512 KB chunk 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.

In the example illustrated in Figure 22, each stripe uses two data chunks to store 1 MB of data. The parity chunk consumes the same amount of disk space as a data chunk but does not count toward capacity. In this example, four stripes are required 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 the full page.

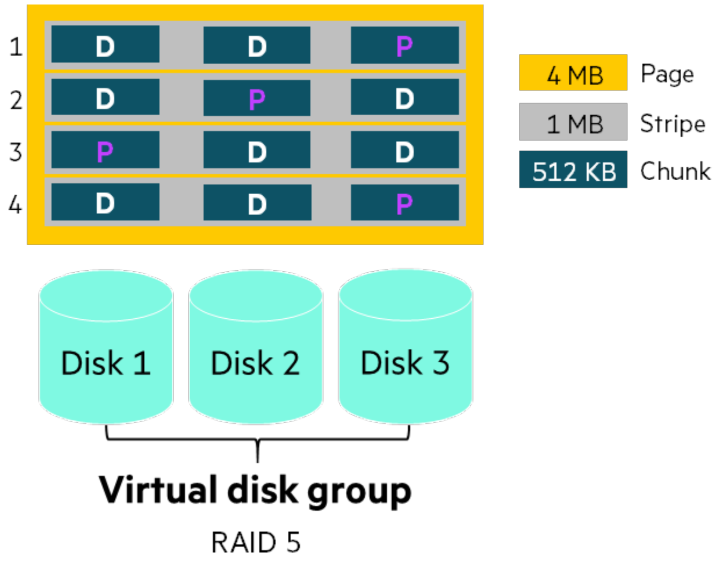


Figure 22. Example of how a page is distributed across physical disks when following the Power of 2 rule

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 9, for each write to a disk group, additional I/Os are required 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:

1. Read the original data chunk
2. Read the original parity chunk
3. Write the new data chunk
4. Write the new parity chunk

Table 9. 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

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, but 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.

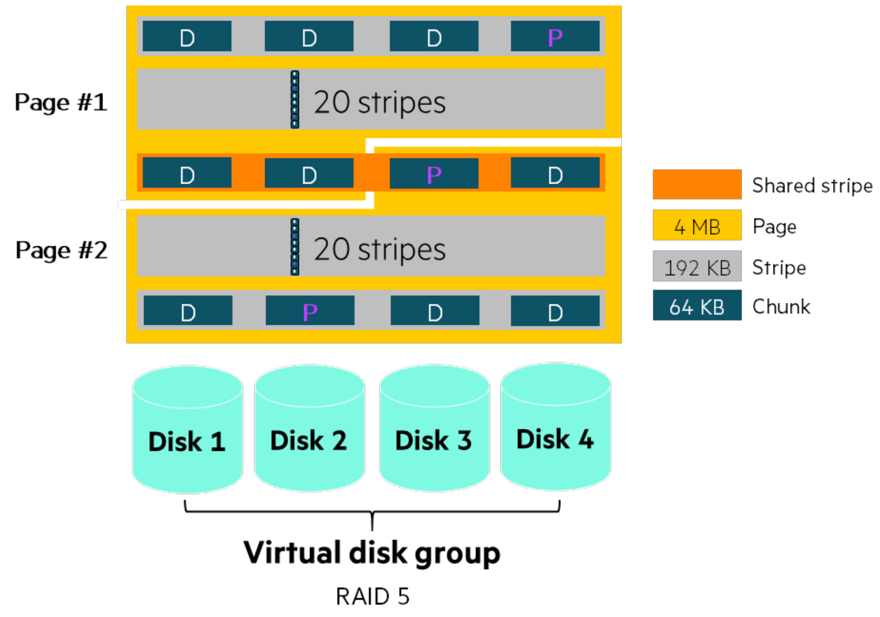


Figure 23. Example of how a page is laid out across physical disks when not following the Power of 2 rule

Capacity expansion

The capacity of a pool is expanded online by the addition of disk groups to a pool, not by expanding an individual disk group. An important consideration is to ensure that when adding hard disk drives to the standard or archive tiers, the performance tier is also increased to maintain a satisfactory proportion of flash storage to mechanical.

Figures 24 and 25 illustrate the before and after tier proportions of a pool from capacity expansion.

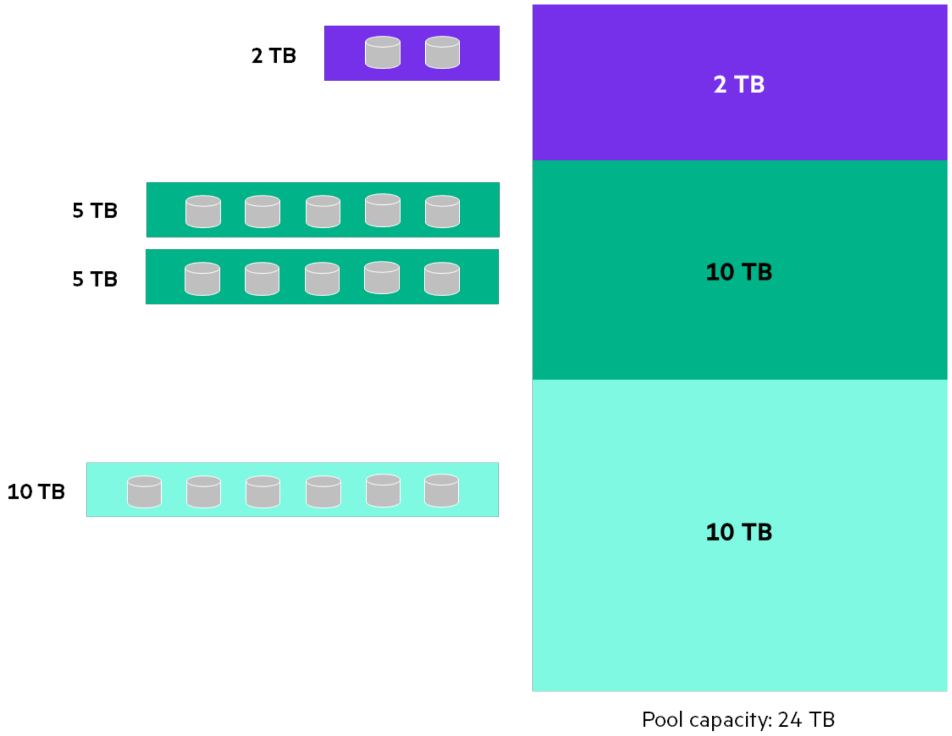


Figure 24. Example pool layout before capacity expansion

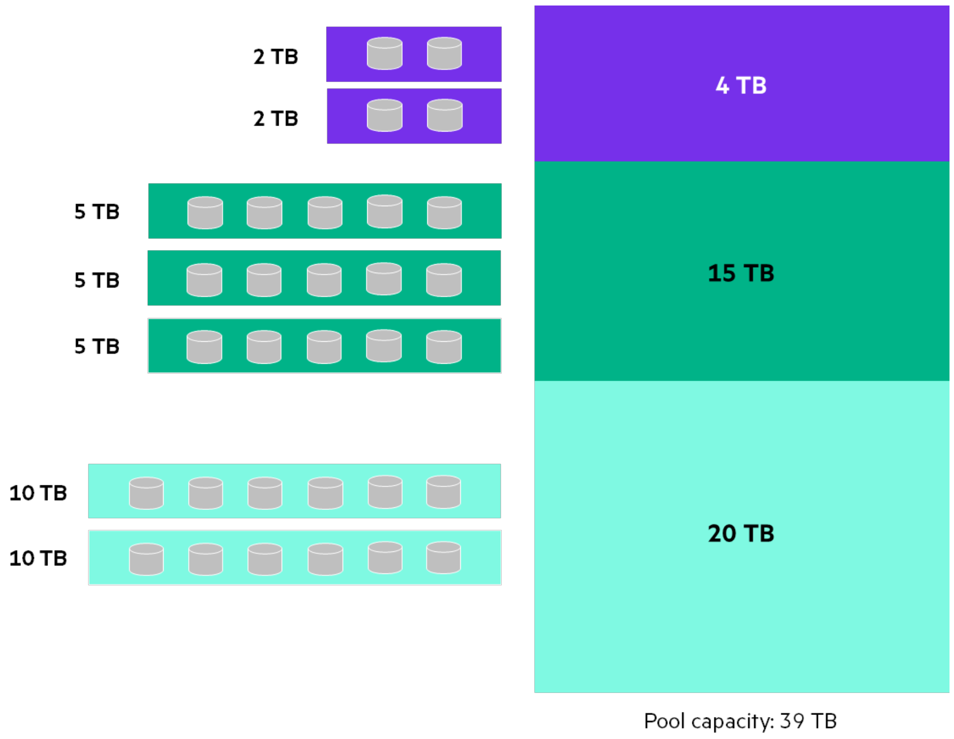


Figure 25. Example pool layout after capacity expansion

SSD drive endurance

Another benefit of virtual storage, automated tiering, and SSD read cache is the ability to reduce wear of SSDs. Memory cells in SSDs eventually wear out because of the physical limitations of the current memory technology used when writing data. Several techniques used in an SSD prevent accelerated wear, and some storage arrays further reduce wear by creative or complex write strategies. The effectiveness of HPE MSA page ranking algorithms, together with data wide striping eliminate this concern. A fortunate by-product of reduced wear is the ability to use lower-cost RI SSDs in an HPE MSA storage system without affecting the useful life of the array. HPE has qualified RI drives for all workloads and in any configuration.

Table 10. Example of SSD drive wear with exaggerated random-write workloads

1TB 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 • Two RAID 1 virtual disk groups	• 18 x 1.8 TB 10K enterprise SAS • Two 9-drive RAID 5 virtual disk groups
Tier capacity	3.84 TB	28.8 TB
Proportion of pool	13.3%	86.7%
Daily writes to tier	810 GB[13]	190 GB
Daily drive writes	405 GB[14]	N/A
Years to drive wear-out	23.7[15]	N/A

Table 10 demonstrates how both the automated tiering engine and wide striping decouples front-end I/O from the back-end I/O, thus reducing drive writes. In this example, 90% of the 1 TB of data flowing through a pool within 24 hours is random, and of that 90% is in the form of writes, so an SSD drive must write 405 GB during that period. Although 405 GB might seem to be a lot, it accounts for only 21% of

[13] 90% of 90%
[14] 21% of DWPD
[15] 20.7 years after warranty expiration

the 1.92 TB of data that can be written every day for five years before the drive would wear out. At this rate, each SSD would have 2.76 PB of unused wear after five years.

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, but not physically allocating pages within the pool. Only when writing data to a volume will a page be allocated and physical capacity consumed. 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. These mechanisms enable a pool to grow and contract as required and results 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 will be released.
- For older operating systems that do not support UNMAP, or where a file system has been fully formatted, periodic zero detection occurs as part of the disk scrubbing maintenance task. Pages that only contain zeroes are released back to the pool for use by another part of the same volume or any other volume within the pool.

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 sends alerts to system administrators who have been configured to receive them, as well as monitoring applications such as Arxscan ArxView® or other similar SNMP/SMI-S based solutions. 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 rebuild

Because unallocated pages do not consume physical disk space, recovery from degraded disk group is faster. If a disk fails or goes offline for any reason and a compatible spare is available, 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 fifth-generation systems. It includes:

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

Note
The ADS Suite is optional for HPE MSA 1050 and 2050 arrays. It is included with HPE MSA 2052 systems, along with two 800 GB SSDs.

Important
A license is required when configuring SSDs as capacity together with hard disk drives within the same system, even if they are not in the same pool, and even if performance tiering is not in use.

Table 11 makes clear when a license is required based on the combination of drives and the model of HPE MSA array.

Table 11. License requirements

Array	Single drive type (HDD or SSD)	Mixing any HDD[16] types within the same system[17]	Coexistence of both a performance tier and any other tier within the same system	SSD read cache
MSA 1050	No	No	Yes	No
MSA 2050	No	No	Yes	No
MSA 2052	No	No	Included	No

[16] 15K, 10K or 7.2K hard disk drives
[17] The term *system* refers to the HPE MSA array and includes either pool.

Data protection

An HPE MSA array provides several technologies in addition to RAID to provide increased resiliency and recovery from failure. This white paper does not address Remote Snap Replication (RSR) because it has only a few use cases and is not flexible enough to be used in all situations. HPE recommends the HPE Remote Snap Replication white paper to review whether RSR would be a good fit for a particular scenario. If not, consider Zerto through the HPE Complete Program.

Snapshots

The volume snapshot mechanism within the HPE MSA improved when moving from linear storage, which used copy on write (CoW) snapshots, to the much more efficient redirect on write (RoW) method. CoW snapshots require the potentially wasteful reservation of useable capacity for snapshot data, but RoW snapshots use pointers to reference shared pages. RoW snapshots also eliminate the performance penalty caused by copying data within the system to protect it.

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 will be 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.

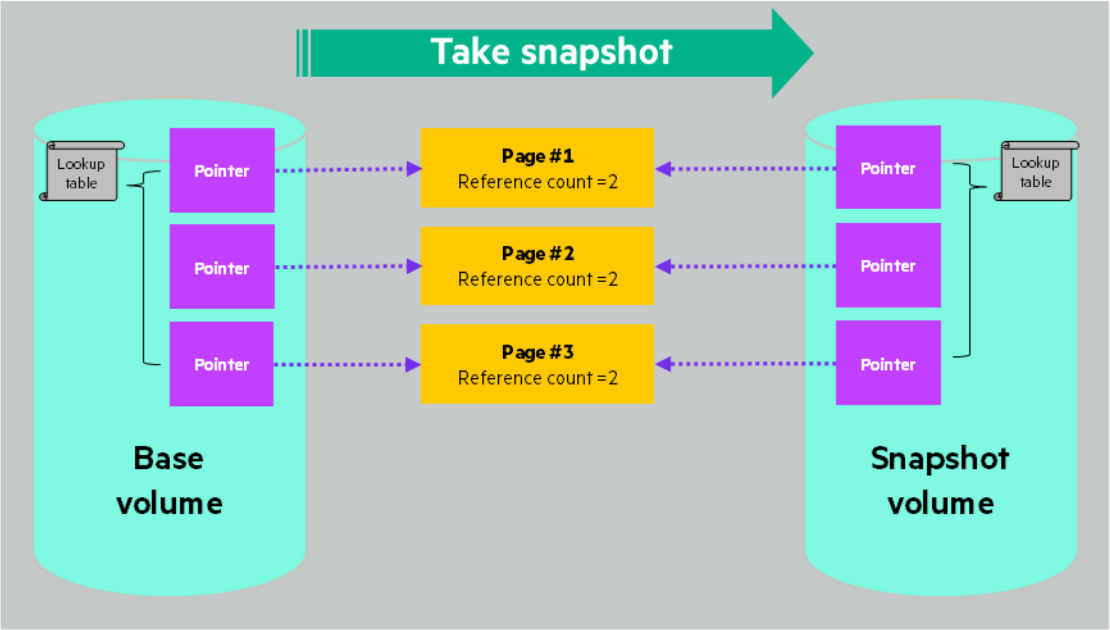


Figure 26. RoW snapshot mechanism

Conclusion

The HPE MSA saw substantial advancements in its design during the fourth-generation[18] product life-cycle, enhancing both its capabilities and its value. These design improvements exist entirely within controller software, adding both a significant increase in performance[19] and advanced new virtual storage architecture[20] but without the cost of new hardware. Supporting older and more modern architectures simultaneously on the same platform provided the opportunity for a business to adopt and adapt to the future of the MSA with minimal disruption.

Virtual storage provides the foundation for a multitude of essential features and capabilities that tackle the challenges faced by small and medium-sized customers. With virtual storage, fifth-generation MSA alleviates management overhead and provides advanced technology to reduce both initial and ongoing costs.

[18] HPE MSA fourth-generation arrays includes HPE MSA 1040, 2040, and 2042 SAN arrays with either SAN or SAS controllers.
[19] Array firmware GL210
[20] Array firmware GL200

Resources

HPE MSA 1050 Storage QuickSpecs
hpe.com/support/MSA1050QuickSpecs

HPE MSA 2050 Storage QuickSpecs
hpe.com/support/MSA2050QuickSpecs

HPE MSA 2052 Storage QuickSpecs
hpe.com/support/MSA2052QuickSpecs

HPE MSA 1050/2050 SMU Reference Guide
https://support.hpe.com/hpsc/doc/public/display?docId=a00017707en_us

HPE MSA 1050/2050/2052 Best practices
https://h20195.www2.hpe.com/v2/getdocument.aspx?docname=a00015961enw

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

Learn more at HPE MSA Storage

hpe.com/storage/msa

Share now

Get updates

a00087404ENW, January 2020, Rev. 1