Technical Report

Storage Subsystem Configuration Guide

Jay White, NetApp
April 2012 | TR-3838

ABSTRACT

This document provides technical recommendations and best practices applicable to the configuration of the NetApp® storage subsystem. This document targets those readers who want to better understand capacity, limits, RAID group sizing, aggregate configuration, shelf technologies, and more.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1   INTRODUCTION

Much has changed since this document's predecessor, the "Aggregate and RAID Group Sizing Guide," was made available in the fall of 2005. The introduction of 64-bit aggregates, larger drive capacities, new shelf technologies, and more warrants a refresh of the information that is the basis of understanding these innovations.

This document provides a deeper understanding of the following areas of the NetApp storage subsystem:

- Storage capacity and limits
- Aggregates and flexible volumes
- RAID layer functionality and configuration
- Drive operations and features
- Shelf architecture and components

By understanding these factors, you will be more effective in addressing the common disinformation that generally surrounds these topics, such as how NetApp competitors and drive vendors state capacity information differently than NetApp.

2   UNDERSTANDING CAPACITY

In order to be successful in sizing any storage configuration, it is necessary to have a firm grasp of what a given storage configuration will yield in terms of usable capacity. Tools are available today that can assist you in determining the capacity of a storage configuration (for example, NetApp Synergy). Tools should be used only to make the process of sizing faster and easier, and they do not replace knowledge and expertise.

2.1 MARKETED CAPACITY

There are two primary reasons for understanding what marketed capacity is:

- To understand how competitors might inappropriately use marketed capacity to imply that they have an advantage over NetApp storage solutions.
- To know why marketed capacities should never be used for determining actual storage capacity.

Marketed capacities are based on the decimal (base 10) representation of the drive capacity. The following factors make the marketed capacity misleading:

- Computer systems use binary (base 2) calculations.
- The marketed capacity does not represent the actual number of physical blocks that exist on the drive, which varies for like-capacity drives from different manufacturers.
- Even within the decimal calculation, drive capacities are rounded up to the nearest common marketed capacity point.
- At a minimum, additional capacity will be consumed by formatting that is necessary to make the drive useful in a computer or storage system.

Referring to drives by their marketed capacity allows different drive manufacturers to group like-capacity drives into common capacity points that can be understood by the average person. Drive manufacturers also get the added benefit of giving the impression that their drive has more usable capacity than it actually does.

When referring to capacities in decimal, the International System of Units (SI) prefix is used to indicate base 10 calculations. SI prefixes have been commonly used to refer to capacity points in computer systems for some time: You know these as TB (terabyte), GB (gigabyte), MB (megabyte), and so on. Binary prefixes are now used to indicate that a represented capacity uses base 2 calculations. Binary prefixes are TiB (tebibyte), GiB (gibibyte), MiB (mebibyte), and so on.

Table 1) Base 10 capacities versus base 2 equivalents.

Base 10 Versus Base 2			
SI Prefix (Base 10)	Equivalent Bytes	Binary Equivalent (Base 2)	Equivalent Bytes
1TB	1 x $1{,}000^4$ bytes	0.9095TiB	0.9095 x $1{,}024^4$ bytes
1GB	1 x $1{,}000^3$ bytes	0.9313GiB	0.9313 x $1{,}024^3$ bytes
1MB	1 x $1{,}000^2$ bytes	0.9537MiB	0.9537 x $1{,}024^2$ bytes

2.2 MANUFACTURER'S PHYSICAL CAPACITY

Manufacturer's physical capacity is important to understand for the following reason:

- Individual drive manufacturers publish their own version of physical drive capacities that are different from what is seen in Data ONTAP®.

This type of physical capacity is based purely on the number of sectors present on the drive before it is used by any system. It is not uncommon to see variations in the number of sectors on like-capacity drives from different manufacturers. Generally SATA drives and the current solid-state drives (SSDs) use 512-byte sectors, whereas FC and SAS drives use 520-byte sectors.

In addition to publishing the physical (or "raw") number of sectors available for any drive, drive manufacturers also use base 10 calculations and not base 2. For example, if we look at a Western Digital RE4 2TB drive's specification (www.wdc.com/wdproducts/library/SpecSheet/ENG/2879-701338.pdf), we see the following stated:

- Sectors per drive = 3,907,029,168
- Formatted capacity = 2,000,398MB

Apply the math (remember base 10 in this case):

- 3,907,029,168 sectors x 512 bytes = 2,000,398,934,016 bytes
- 2,000,398,934,016 bytes / 1,000 (base 10) = 2,000,398,934kB / 1,000 = 2,000,398MB

If you consider the preceding is base 2, you get the following:

- 3,907,029,168 sectors x 512 bytes = 2,000,398,934,016 bytes
- 2,000,398,934,016 bytes / 1,024 (base 2) = 1,953,514,584kiB / 1,024 = 1,907,729MiB

Customers should not be exposed to this information under regular circumstances, but this information is publically accessible from the various drive manufacturers' Web sites.

When considering capacity "overhead" for storage systems, you should never consider the delta between the usable capacity and marketed capacity of the drive but rather the delta between the base 2 manufacturer's physical capacity and usable capacity of the drive.

2.3 PHYSICAL CAPACITY

Physical capacity is important to understand for the following reasons:

- Physical capacity is the factor that determines if a system has exceeded maximum system capacity.
- This is a more realistic calculation that can be used as a basis for the physical (or "raw") capacity of a storage configuration.

The physical capacity of a drive as seen in Data ONTAP is based on the following:

- The physical number of sectors present for the drive
- Application of block checksum scheme (BCS) formatting for FC and SAS drives or BCS 8/9 formatting for SATA drives (explained in more detail later)

FC and SAS drives use 520-byte sectors. FC and SAS drives are formatted using the block checksum scheme (BCS) method. Each 4KB WAFL® (Write Anywhere File Layout) block is made up of eight sectors. The last eight bytes of each sector are reserved to store the block checksum. This means that each 520-byte sector is effectively 512 bytes of usable capacity. In doing the math, you end up with 512 bytes x 8 = 4,096 bytes, or a 4kB WAFL block.

SATA drives start with 512-byte sectors. In order to store block checksum information, Data ONTAP uses the BCS 8/9 checksum method when formatting the drive. This means that each 4kB WAFL block is made up of nine sectors, eight that are used to store data and the ninth sector used to store the checksum (8/9).
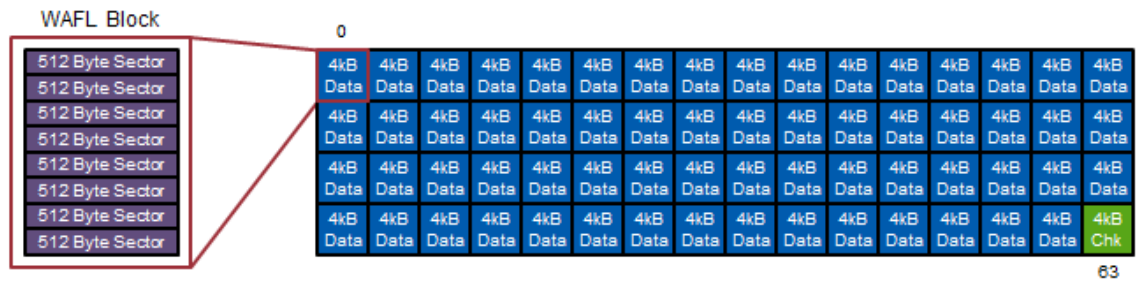
SSDs currently use 512-byte sectors. As a result, SSD is subject to the BCS 8/9 checksum format that is described earlier for SATA drives.

4kB WAFL Block (520B sectors)

520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum
520 Byte Sector = 512 Bytes Data | 8 Bytes Checksum

4kB WAFL Block (512B sectors)

512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 512 Bytes Data
512 Byte Sector = 64 Bytes Checksum

Figure 1) Block checksum scheme for 520-byte and 512-byte sectors.

The `sysconfig -r` command will always display the number of sectors (reported as blks) that are available on the drive (after applying BCS or 8/9 formatting) as well as the physical capacity of the drive in MiB (reported as MB). An example of a `sysconfig -r` output for a 2TB SATA drive is as follows:

```
Phys (MB/blks)
--------------
1695759/3472914816
```

The following math demonstrates two things:

- How the `blks` result is determined
- How the `MB` result is determined

To keep consistency, we use the same drive example as is used in section 2.2, "Manufacturer's Physical Capacity." In that example we used the Western Digital RE4 2TB drive specification (www.wdc.com/wdproducts/library/SpecSheet/ENG/2879-701338.pdf), which states that it contains 3,907,029,168 sectors. So why does Data ONTAP report that there are only 3,472,914,816 sectors (blks)?

Apply the math:

- BCS 8/9 method multiplier = 0.88888888888888888888888888888889
- 3,907,029,168 sectors x .88888888888888888888888888888889 = 3,472,914,816 sectors (blks)
- 3,472,914,816 blks x 512 bytes = 1,778,132,385,792 bytes
- 1,778,132,385,792 bytes / 1,024 = 1,736,457,408kiB / 1,024 = 1,695,759MiB

Even though all calculations in Data ONTAP are done in base 2, they are shown in command output using the SI prefix (not the binary prefix).

ZONE CHECKSUM SCHEME

FAS systems running Data ONTAP have not used the zone checksum scheme (ZCS) for some time (they started using BCS some time ago). ZCS is still an option for V-Series systems that can use third-party storage. This checksum scheme is more capacity friendly than the BCS 8/9 checksum scheme, but it suffers from a performance impact for read operations and the inability to work with some resiliency features of Data ONTAP (such as lost write protection).

Figure 2) Zone checksum scheme (ZCS).

This checksum scheme is very different from BCS. BCS deals primarily at the individual sector level, whereas ZCS works with a collection of 4kB WAFL blocks. As you can see in Figure 2, there is a 4kB checksum block for every 63 "data" WAFL blocks. The checksum block is located in the last block position, which is the primary reason for the performance impact, since the checksum information is located in an entirely different WAFL block from the data block (adding seek time). Checksums for BCS are embedded in the WAFL block, which eliminates the seek time for checksums.

2.4 USABLE CAPACITY

Usable capacity is the most important capacity point to understand for the following reasons:

- It is the usable capacity that determines if you have exceeded the maximum aggregate size.
- Usable capacity is the actual storage capacity point (minus parity and hot spare drives) that can be used to store data.

Since each drive manufacturer ships like-capacity drives with different numbers of sectors, it is important that the file system format normalizes the actual size of like-capacity drives. The process of normalization is known as right-sizing. The following factors are accounted for during the drive right-sizing process:

- Sector normalization occurs across like-capacity drives.
- Space is reserved for RAID labels, mailbox, and core zone.

Sector normalization is an important process that allows Data ONTAP to standardize the usable number of sectors available from any manufacturer's like-capacity drives. With normalizing, any manufacturer's like-capacity drive can be used interchangeably with another. This affords consistency in capacity sizing; greater hot spare flexibility, because there is no need to maintain hot spares by manufacturer; and more.

In order to maximize the usable space of SSDs, Data ONTAP does not reserve space for the core zone for SSDs. SSD systems require HDD (rotating media) to be attached (no homogeneous SSD systems) so the HDD can provide the core zone for SSDs.

Once the right-sizing process is complete, the resulting usable capacity can be referred to as the right-sized capacity of the drive. The `sysconfig -r` command will always display the right-sized usable capacity of a drive in MiB (reported as MB) based on the resulting usable number of usable sectors (reported as blks). An example of a `sysconfig -r` output for a 500GB SATA drive is as follows:

```
Used (MB/blks)
--------------
423111/866531584
```

Apply the math:

- 866,531,584 sectors x 512 bytes = 443,664,171,008 bytes
- 443,664,171,008 bytes / 1,024 = 433,265,792kiB / 1,024 = 423,111MiB

2.5 THE NETAPP GIGABYTE

Some might be confused by a calculation known in Data ONTAP as the NetApp gigabyte (nGB). An nGB is equal to 1,000 x 1,024 x 1,024 bytes. This calculation method is used by Data ONTAP only when reporting capacities in GB, and it is not used for actual sizing, which is done completely using base 2 calculations. The

most common application of this calculation is with listed drives on a system. For example, a 300GB drive is 272,000MiB for actual sizing. When this drive is listed in nGB, it is 272nGB (but is displayed as 272GB in Data ONTAP).

This can be confusing because Data ONTAP shows sizing calculations using the SI prefix (decimal), and when extrapolating capacity to GB, it uses nGB. For example, a 600GB drive will display 560,000MB as the usable capacity in the output of the `sysconfig -r` command, but it actually is 560,000MiB.

2.6 AGGREGATE AND FLEXIBLE VOLUME CAPACITY

Without taking into consideration any space reservations, it is generally stated that the maximum flexible volume size is the same as the maximum aggregate size. For example, on a FAS6280 the maximum aggregate size is 100TiB, which would result in the largest flexible volume size being 100TiB. However, this is not accurate.

The first issue to understand is that, just because the maximum-sized aggregate is a specific capacity, it does not mean that the drive capacities used to populate that aggregate result in a capacity exactly equal to the maximum aggregate size. For example, consider a 100TiB aggregate populated entirely using 2TB drives:

- 61 x 2TB data drives can fit in a 100TiB aggregate
- 61 x 2TB drives (1,695,466MiB each) equals 103,423,426MiB, or 98.63TiB

Note that all calculations are conducted in MiB. If a value is not stated in MiB, it is reduced or increased as appropriate. Only after the result is determined is the resulting value scaled up or down from MiB.

In addition to the number of data drives that fit into the aggregate, it is necessary to account for 20.5MiB from each data drive in the aggregate (the space is reserved on all drives, but for capacity only data drives matter). This space is reserved for bootstrap, kernel, and some other disk-related data. Therefore, continuing the previous example:

- 61 x 20.5MiB = 1,250.5MiB subtract from 103,423,426MiB = 103,422,175MiB or 98.63TiB

The following table shows the resulting maximum aggregate size by disk capacity based on the number of data drives that can fit into the maximum aggregate capacities supported on Data ONTAP 8.0.x and 8.1.x.

Table 2) Data ONTAP 8.0.x and 8.1.x resulting maximum aggregate size by drive capacity.

Resulting Maximum Aggregate Size by Drive Capacity					
Drive Capacity	**Maximum Aggregate Size**				
	16TiB (16,777,216MiB)	**30TiB** (31,457,280MiB)	**50TiB** (52,428,800MiB)	**60TiB** (62,914,560MiB)	**70TiB** (73,400,320MiB)
300GB SAS/FC	(61 drives) 16,592,000MiB (15.82TiB)	(115 drives) 31,280,000MiB (29.83TiB)	(192 drives) 52,224,000MiB (49.80TiB)	(231 drives) 62,832,000MiB (59.52TiB)	(269 drives) 73,168,000MiB (69.77TiB)
450GB SAS/FC	(40 drives) 16,720,000MiB (15.94TiB)	(75 drives) 31,350,000MiB (29.90TiB)	(125 drives) 52,250,000MiB (49.83TiB)	(150 drives) 62,700,000MiB (59.79TiB)	(175 drives) 73,150,000MiB (69.76TiB)
600GB SAS/FC	(29 drives) 16,240,000MiB (15.49TiB)	(56 drives) 31,360,000MiB (29.91TiB)	(93 drives) 52,080,000MiB (49.67TiB)	(112 drives) 62,720,000MiB (59.81TiB)	(131 drives) 73,360,000MiB (69.96TiB)
500GB SATA	(39 drives) 16,501,329MiB (15.74TiB)	(74 drives) 31,310,214MiB (29.86TiB)	(123 drives) 52,042,653MiB (49.63TiB)	(148 drives) 62,620,428MiB (59.72TiB)	(173 drives) 73,198,203MiB (69.80TiB)
1TB SATA	(19 drives) 16,103,545MiB (15.36TiB)	(37 drives) 31,359,535MiB (29.91TiB)	(61 drives) 51,700,855MiB (49.30TiB)	(74 drives) 62,719,070MiB (59.81TiB)	(86 drives) 72,889,730MiB (69.51TiB)
2TB SATA	(9 drives) 15,259,194MiB (14.55TiB)	(18 drives) 30,518,388MiB (29.10TiB)	(30 drives) 50,863,980MiB (48.51TiB)	(37 drives) 62,732,242MiB (59.83TiB)	(43 drives) 72,905,038MiB (69.53TiB)
3TB SATA	(6 drives) 15,231,276MiB (14.53TiB)	(12 drives) 30,462,552MiB (29.05TiB)	(20 drives) 50,770,920MiB (48.42TiB)	(24 drives) 60,925,104MiB (58.10TiB)	(28 drives) 71,079,288MiB (67.79TiB)

Resulting Maximum Aggregate Size by Drive Capacity					
Drive Capacity	**Maximum Aggregate Size**				
	75TiB (78,643,200MiB)	**90TiB (94,371,840)**	**100TiB (104,857,600)**	**105TiB (110,100,480)**	**162TiB (169,869,312)**
300GB SAS/FC	(289 drives) 78,608,000MiB (74.96TiB)	(346 drives) 94,112,000MiB (89.75TiB)	(385 drives) 104,720,000MiB (99.86TiB)	(404 drives) 109,888,000MiB (104.79TiB)	(624 drives) 169,728,000MiB (161.85TiB)
450GB SAS/FC	(188 drives) 78,584,000MiB (74.94TiB)	(225 drives) 94,050,000MiB (89.69TiB)	(250 drives) 104,500,000MiB (99.65TiB)	(263 drives) 109,934,000MiB (104.84TiB)	(406 drives) 169,708,000MiB (161.84TiB)
600GB SAS/FC	(140 drives) 78,400,000MiB (74.77TiB)	(168 drives) 94,080,000MiB (89.72TiB)	(187 drives) 104,720,000MiB (99.87TiB)	(196 drives) 109,760,000MiB (104.67TiB)	(303 drives) 169,680,000MiB (161.81TiB)
500GB SATA	(185 drives) 78,275,535MiB (74.65TiB)	(223 drives) 94,353,753MiB (89.98TiB)	(247 drives) 104,508,417MiB (99.66TiB)	(260 drives) 110,008,860MiB (104.91TiB)	(401 drives) 169,667,511MiB (161.80TiB)
1TB SATA	(92 drives) 77,975,060MiB (74.36TiB)	(111 drives) 94,078,605MiB (89.72TiB)	(123 drives) 104,249,265MiB (99.42TiB)	(129 drives) 109,334,595MiB (104.27TiB)	(200 drives) 169,511,000MiB (161.65TiB)
2TB SATA	(46 drives) 77,991,436MiB (74.38TiB)	(55 drives) 93,250,630MiB (88.93TiB)	(61 drives) 103,423,426MiB (98.63TiB)	(65 drives) 108,509,824MiB (103.48TiB)	(100 drives) 169,546,600MiB (161.69TiB)
3TB SATA	(30 drives) 76,156,380MiB (72.63TiB)	(37 drives) 93,926,202MiB (89.57TiB)	(41 drives) 104,080,386MiB (99.26TiB)	(43 drives) 109,157,478MiB (104.10TiB)	(66 drives) 167,544,036MiB (159.78TiB)

To create a maximum-sized volume within a given aggregate capacity, two capacity calculations should be made to determine the maximum size of the volume. This accounts for the mandatory capacity reservations that are made and does not consider additional capacity factors such as deduplication, volume Snapshot™ reserves, and more. First, take the 10% WAFL reserve out of the remaining capacity. You then need 100.5% of the remaining space to create the volume, which means that you further reduce the capacity by 0.005%. You then have the maximum-sized volume that can be created within the aggregate.

Using the 2TB drive example from above (for a 100TiB aggregate), you get the following:

- 103,423,426MiB – 10% = 93,081,083.4MiB
- 93,081,083.4MiB – 0.005% = 92,615,677.983MiB / 1024 / 1024 = 88.32TiB

The following table gives the estimated maximum usable FlexVol® size for the currently supported 64-bit aggregate capacities and shipping drive capacities. Note that this calculation only takes into account mandatory space reservations and does not consider user-configurable space reservations (for example, Snapshot reserve) or storage efficiency features (for example, deduplication).

Table 3) Data ONTAP 8.0.x and 8.1.x maximum volume size by drive capacity.

Data ONTAP 8.0.x and 8.1.x Estimated Volume Size by Drive Capacity							
Drive Capacity	**Maximum FlexVol Volume Size**						
	16TiB	**30TiB**	**50TiB**	**60TiB**	**70TiB**	**100TiB**	
300GB SAS or FC	14.17TiB	26.71TiB	44.60TiB	53.66TiB	62.48TiB	89.43TiB	Usable Vol Size
450GB SAS or FC	14.28TiB	26.77TiB	44.62TiB	53.54TiB	62.47TiB	89.24TiB	
600GB SAS or FC	13.87TiB	26.78TiB	44.48TiB	53.56TiB	62.65TiB	89.43TiB	
500GB SATA	14.09TiB	26.74TiB	44.44TiB	53.48TiB	62.51TiB	89.25TiB	
1TB SATA	13.75TiB	26.78TiB	44.15TiB	53.56TiB	62.25TiB	89.03TiB	
2TB SATA	13.03TiB	26.06TiB	43.44TiB	53.57TiB	62.26TIB	88.32TiB	
3TB SATA	13.01TiB	26.02TiB	43.36TiB	52.03TiB	60.70TiB	88.89TiB	

The preceding table addresses systems where the maximum FlexVol capacity is equal to the maximum aggregate capacity. With the introduction of Data ONTAP 8.1, many systems' maximum FlexVol capacity is less than the maximum aggregate capacity. On these systems it is possible to yield larger usable volume capacity as some of the aggregate-level capacity reservations are pushed into the remaining aggregate

capacity that is beyond the maximum FlexVol capacity. On these systems you are able to create a larger flexible volume (very near full FlexVol size of usable capacity) than is listed in the preceding table.

2.7 DRIVE CAPACITY COMPARISON

We have now addressed the specifics of marketed, physical, and usable capacity. Table 4 provides a comparison of these different types of capacity as they relate to the common drive capacities shipping today.

Table 4) Physical and usable capacities for commonly shipping drives.

Base 10 Versus Base 2		
Marketed Capacity	Physical Capacity (MiB and GiB)	Usable Capacity (MiB and GiB)
100GB SSD (X441A-R5)	95,396MiB 93.1GiB	95,146MiB 92.9GiB
100GB SSD (X442A-R5)	84,796MiB 82.7GiB	84,574MiB 82.5GiB
300GB SAS/FC	280,104MiB 273.5GiB	272,000MiB 265.6GiB
450GB SAS/FC	420,156MiB 410.3GiB	418,000MiB 408.2GiB
500GB SATA	423,946MiB 414.0GiB	423,111MiB 413.2GiB
600GB SAS/FC	560,208MiB 547.1GiB	560,000MiB 546.9GiB
1TB SATA	847,884MiB 828.0GiB	847,555MiB 827.7GiB
2TB SATA	1,695,702MiB 1,655.9GiB	1,695,466MiB 1,655.7GiB
3TB SATA	2,543,634MiB 2,484.0GiB	2,538,546MiB 2,479.0GiB

Physical drive capacities vary by individual drive manufacturer. The capacities listed in Table 4 are based on actual `sysconfig -r` command output for the marketed drive capacities listed.

ADDITIONAL FACTORS

Several additional factors affect usable capacity when additional configuration, such as creating aggregates and flexible volumes, is applied to a storage configuration. Some of these factors include:

- 10% WAFL system reserve
- 5% aggregate Snapshot reserve
- 20% default volume Snapshot reserve

Configuration of aggregates and flexible volumes, and how that affects capacity, is discussed in more detail in the "Aggregates and RAID Groups" section of this document.

3 SYSTEM LIMITS

Two primary system limits in place today are enforced by Data ONTAP. The vast majority of the time the maximum spindle (drive) limit is the limiting factor in any storage configuration. Your storage configuration is considered limited once either one of the aforementioned limits is encountered.

It is important to note that, depending on your configuration, you might need to consider other types of limits when designing your storage configuration. Examples are:

- Volume size limits
- Maximum number of volumes
- Deduplication volume limits

3.1 MAXIMUM SPINDLES (DRIVES)

The maximum spindle (drive) limit is a system limit that is enforced by Data ONTAP. If you are planning a maximum-capacity storage configuration, it is very likely that maximum spindles will be the limiting factor. Maximum spindle limits are:

- Specific to each platform
- Independent of shelf type
- Specific to the version of Data ONTAP

Consult the latest platform and system documentation for the most up-to-date information on maximum spindle limits for each platform. Table 5 shows maximum spindle limits for Data ONTAP 8.0.x and 8.1.x.

Table 5) Maximum spindles by platform.

Data ONTAP 8.0.x and 8.1.x Maximum Spindles by Platform		
Platform	Maximum Spindles 8.0.x	Maximum Spindles 8.0.x
FAS2040	136 drives	136 drives
FAS2240-2	Unsupported	144 drives
FAS2240-4	Unsupported	144 drives
FAS/V3040	336 drives	336 drives
FAS/V3140	420 drives	420 drives
FAS/V3070	504 drives	504 drives
FAS/V3160	672 drives	672 drives
FAS/V3170	840 drives	840 drives
FAS/V3210	240 drives	240 drives
FAS/V3240	600 drives	600 drives
FAS/V3270	960 drives	960 drives
FAS/V6030	840 drives	840 drives
FAS/V6040	840 drives	840 drives
FAS/V6070	1,008 drives	1,008 drives
FAS/V6080	1,176 drives	1,176 drives
FAS/V6210	1,200 drives (1,008 FC drives)	1,200 drives (1,008 FC drives)
FAS/V6240	1,440 drives (1,176 FC drives)	1,440 drives (1,176 FC drives)
FAS/V6280	1,440 drives (1,176 FC drives)	1,440 drives (1,176 FC drives)

When using 2TB drives with Data ONTAP 7.3.x, the maximum spindle count is halved. See section 3.2, "Maximum System Capacity," for more information on using 2TB drives with Data ONTAP 7.3.x.

Maximum spindle counts might not always divide evenly by the number of drives that are populated in DS14 and DS4243 storage shelves. In some cases, internal drives and a mix of different external storage shelf densities are factored into spindle count limits. Some factors to keep in mind when sizing your configuration are:

- FAS2000 series internal drives (12 drives)
- DS4243 half storage shelves (12 drives)
- Mixing of DS14 (14 drives) and DS4243 (24 drives) storage shelves

3.2 MAXIMUM SYSTEM CAPACITY

The maximum system capacity limit is a system limit that is enforced by Data ONTAP. When major Data ONTAP versions are released, the maximum system capacities are determined based on the largest capacity drive that is shipping at release time, in relation to the maximum number of supported spindles for each platform.

This is a multiplication of the marketed capacity of the drive by maximum supported spindles, which is then converted into base 2. This leaves a gap between the base 2 capacities that can be added to a system based on maximum spindle count. This gap can be misleading because it is not representative of the actual capacity that can be added to a storage system. The term "full system capacity" refers to the resulting capacity of a storage system when a maximum number of drives of a specific capacity point are attached to a storage system. As a result, full system capacity is a more accurate representation of the true maximum amount of capacity that can be attached to a system.

In order to achieve accurate calculations, it is important to always reduce calculations of capacity down to MiB, because this is what Data ONTAP reports for physical and usable drive capacities. The larger the capacity size (GiB, TiB, and so on), the larger the resulting delta in actual capacity is should you not reduce the number down to MiB prior to multiplication or division, because larger capacities are normally stated after having been rounded or with decimal places completely removed or reduced. As you can see from the preceding example, maximum spindle count will be encountered before the maximum system capacity.
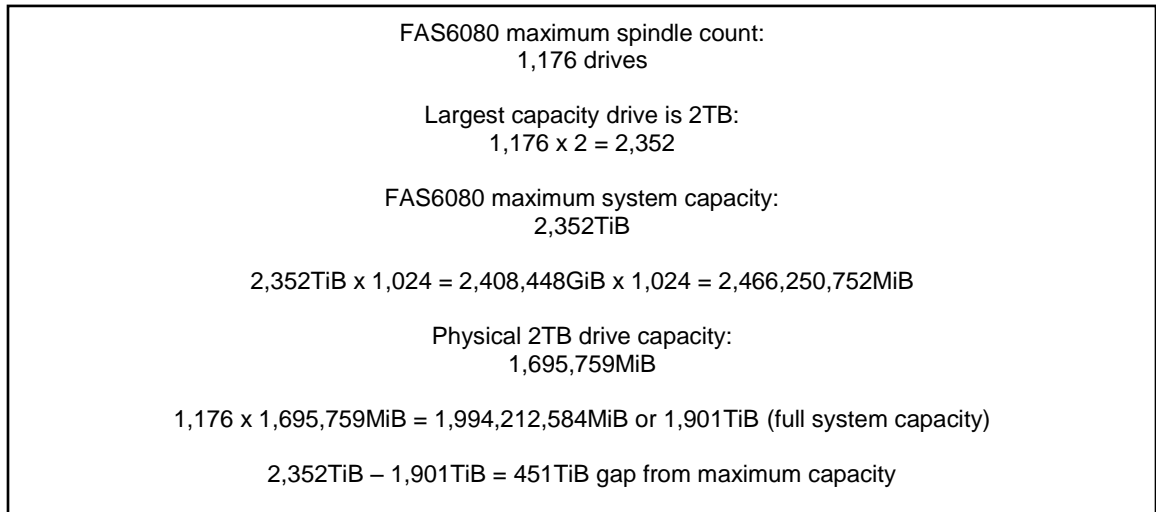
Figure 3 shows an example with Data ONTAP 8.0.

FAS6080 maximum spindle count:
1,176 drives

Largest capacity drive is 2TB:
1,176 x 2 = 2,352

FAS6080 maximum system capacity:
2,352TiB

2,352TiB x 1,024 = 2,408,448GiB x 1,024 = 2,466,250,752MiB

Physical 2TB drive capacity:
1,695,759MiB

1,176 x 1,695,759MiB = 1,994,212,584MiB or 1,901TiB (full system capacity)

2,352TiB – 1,901TiB = 451TiB gap from maximum capacity

Figure 3) Why maximum spindle count is normally the limiting factor in a storage configuration.

When a larger capacity drive is supported, older versions of Data ONTAP might end up supporting new capacity points without the maximum system capacities increasing for that specific version of Data ONTAP. This is the one case in which maximum system capacity becomes the limiting factor for platforms. An example is Data ONTAP 7.3.2, which supports the 2TB drive capacity point. The maximum system capacities for platforms running Data ONTAP 7.3.2 have not increased and therefore are still based on 1TB drives. For example, a FAS6080 running Data ONTAP 7.3.2 has a maximum system capacity of 1,176TiB. As seen in the previous example, 1,176 drives at 2TB yield 1,901TiB, which is far over the maximum system capacity. Only in this rare circumstance will it be necessary to calculate system limits based on maximum system capacity.

As an alternative to attempting to squeeze every last bit of capacity out of a configuration, a simplified approach can be taken. Using the earlier 2TB drive example, you could simply divide the maximum spindle count for a platform, in this case a FAS6080, by 2. This would yield a maximum spindle count of 588 (1,176 / 2), which would be a guideline since it would not technically be enforced by Data ONTAP.

A situation that warrants using a simplified approach to make this easier to digest is having mixed storage configurations. For example, if I had a storage configuration with a mix of 500GB, 1TB, and 2TB drives and I wanted to implement a maximum capacity system, I could do the math and wrack my brain, or I could use a simplified approach:

- Multiply the number of 2TB drives by 2.
- Subtract the result from the maximum spindles supported for the platform.
- The result would be the remaining spindle count that could be used for non-2TB-capacity drives.

These simplified suggestions only apply when system capacity becomes the limiting factor in sizing your storage configuration.

Be aware of any stated maximum spindle limitations for these situations. To avoid forcing the use of a manual sizing method when a situation is encountered in which maximum system capacity becomes the enforced maximum, product management will normally state a spindle limitation. This stated limit is not enforced by Data ONTAP but is meant to provide ease of configuration. For example, a stated spindle maximum for 2TB drives deployed on Data ONTAP 7.3.x systems is that normal spindle maximums are halved.

Consult the latest platform and system documentation for the most up-to-date information on maximum system capacity limits for each platform. Table 6 shows maximum system capacity limits for Data ONTAP 7.3.x, 8.0.x, and 8.1.x.

Table 6) Maximum system capacity by platform for Data ONTAP 7.3.x, 8.0.x, and 8.1.x.

Data ONTAP 7.3.x, 8.0.x and 8.1.x Maximum Capacity by Platform			
Platform	Maximum System Capacity		
	Data ONTAP 7.3.x	Data ONTAP 8.0/8.0.1	Data ONTAP 8.0.2/8.1.x
FAS2040	136TiB	272TiB	408TiB
FAS2240-2	Unsupported	Unsupported	374TiB
FAS2240-4	Unsupported	Unsupported	432TiB
FAS/V3040	336TiB	672TiB	672TiB
FAS/V3140	420TiB	840TiB	1,260TiB
FAS/V3070	504TiB	1,008TiB	1,008TiB
FAS/V3160	672TiB	1,344TiB	2,016TiB
FAS/V3170	840TiB	1,680TiB	2,520TiB
FAS/V3210	240TiB	480TiB	720TiB
FAS/V3240	600TiB	1,200TiB	1,800TiB
FAS/V3270	900TiB	1,920TiB	2,880TiB
FAS/V6030	840TiB	1,680TiB	2,520TiB
FAS/V6040	840TiB	1,680TiB	2,520TiB
FAS/V6070	1,008TiB	2,016TiB	3,024TiB
FAS/V6080	1,176TiB	2,352TiB	3,528TiB
FAS/V6210		2,400TiB	3,600TiB
FAS/V6240	Unsupported	2,880TiB	4,320TiB
FAS/V6280		2,880TiB	4,320TiB

4   AGGREGATES AND RAID GROUPS

Since their implementation in Data ONTAP 7.0, aggregates have become the primary storage containers for NetApp storage configurations. An aggregate consists of one or more RAID groups consisting of both data and parity drives. With larger capacity drives and the introduction of 64-bit aggregates, we need to update our practices around sizing and implementation.

4.1 ROOT VOLUMES

In Data ONTAP 8.0 and earlier the aggregate containing the root FlexVol volume must be a 32-bit aggregate. Starting with Data ONTAP 8.0.1 the root aggregate can be a 64-bit aggregate. Table 7 outlines the minimum root FlexVol volume sizes for each platform for Data ONTAP 7.3.x, 8.0.x, and 8.1.x. Note that the numbers in the table are confirmed through actual system testing.

The various Data ONTAP system administration guides might not match Table 7. This is a known issue that will be resolved in future revisions of the Data ONTAP documentation.

The recommended root volume size is based on enabling adequate space to be reserved for system files, log files, and core files. If system problems occur, files contained on the root volume will be needed by technical support to assist with troubleshooting. User data should never be stored in the root volume.

When a volume is created in Data ONTAP, it reserves 20% of the capacity for volume Snapshot copies by default. This capacity overhead is not included in the recommended root volume size. The recommended root volume size is for the actual usable volume capacity. Note that it is not necessary to keep such a large volume Snapshot reserve for a root volume. As a result, administrators should consider reducing the volume Snapshot reserve for root volumes.

Table 7) Data ONTAP recommended root FlexVol volume size by platform.

Data ONTAP Recommended Root FlexVol Volume Size by Platform			
Platform	Root Volume Size 7.3.x	Root Volume Size 8.0.x	Root Volume Size 8.1.x
FAS2040	13GiB	133GiB	132GiB
FAS2240-2	Unsupported	Unsupported	159GiB
FAS2240-4	Unsupported	Unsupported	159GiB
FAS/V3040	16GiB	141GiB	134GiB
FAS/V3140	16GiB	141GiB	141GiB
FAS/V3070	23GiB	213GiB	205GiB
FAS/V3160	16GiB	141GiB	141GiB
FAS/V3170	40GiB	250GiB	250GiB
FAS/V3210	17GiB	151GiB	151GiB
FAS/V3240	22GiB	205GiB	195GiB
FAS/V3270	44GiB	250GiB	250GiB
FAS/V6030, 6040	37GiB	250GiB	250GiB
FAS/V6070, 6080	69GiB	250GiB	250GiB
FAS/V6200 series	Unsupported	250GiB	250GiB

4.2 AGGREGATE PERFORMANCE

NetApp recommends using the Sizer tools to determine the expected level of performance for your given storage configuration. Some basic best practices are applicable to all storage configurations, such as:

- RAID group size should be set to the highest optimal size value within the recommended RAID group size range.
    - For example, the recommended RAID group size range is from 12 through 20 for SAS-based RAID groups. If a RAID group size of both 12 and 18 provides an optimal RAID group layout (even layout), NetApp recommends selecting the higher of the two values to enable optimal spindle count.
- Use 64-bit aggregates versus 32-bit aggregates using large SATA drives.
    - Sequential reads/writes = same performance
    - Random writes = same performance
    - Random reads = same or worse performance with 64-bit aggregates
        - Use the appropriate sizing tools to determine if PAM is beneficial for your random read workloads.
- Do not mix different speed drives within the same aggregate.
- Determine that no single RAID group is deficient by more than a single drive from the RAID group size set for the aggregate.
- Do not spread RAID groups from the same aggregate across different shelf technologies.
    - For example, spreading an aggregate RAID group across both DS4243 and DS14Mk2AT shelves affects aggregate performance (and resiliency) as a whole.

In some cases, such as with 2TB drives in Data ONTAP 7.3.2, the largest sized aggregate results in a suboptimal configuration. In Data ONTAP 7.3.2 the maximum aggregate size is 16TB, which results in a single RAID group of nine data drives when using 2TB drives. In this case migrating to Data ONTAP 8.0 to enable 64-bit aggregates is warranted in order to increase spindle counts for the aggregate.

4.3 32-BIT AGGREGATES

All Data ONTAP 7.x.x aggregates are 32-bit aggregates and have a maximum capacity of 16TiB. Determining how many drives can fit into a maximum-sized 32-bit aggregate depends on the drive capacity point and the version of Data ONTAP:

- In Data ONTAP 7.3.x and later, the sum of the right-sized (usable) capacity of all data drives is used to determine the number of drives that can fit into a 16TiB aggregate.
- In Data ONTAP 7.2.x and earlier, the sum of the physical capacity of all data and parity drives is used to determine the number of drives that can fit into a 16TiB aggregate.

When upgrading from Data ONTAP 7.2.x to Data ONTAP 7.3.x or later, it is very likely that you will be able to add additional drives into any existing maximum-size aggregates. This depends on the delta in usable versus physical drive capacity and the number of parity drives utilized by the RAID group(s) in the

aggregate. When drives are added to an existing aggregate, there is a potential performance effect. NetApp recommends using `reallocate` to balance data across all RAID groups in this situation.

In Data ONTAP 8.0 NetApp requires that the root aggregate is a 32-bit aggregate. Starting with Data ONTAP 8.0.1 the root aggregate can be a 64-bit aggregate.

4.4 64-BIT AGGREGATES

Starting with Data ONTAP 8.0, aggregates can be either 32-bit or 64-bit aggregates. The primary benefits of 64-bit aggregates are:

- Aggregates are larger than the 16TiB limit for 32-bit aggregates
- Better volume performance, because there are more drives in the aggregate
- Higher storage efficiency, especially for large-capacity SATA drives
- Fewer aggregates on a system are easier to manage than many small aggregates
- Better scalability for future growth

64-bit aggregate capacity limits vary by platform, as shown in Table 8.

Table 8) Maximum 64-bit aggregate capacity by platform.

Maximum 64-Bit Aggregate Size by Platform and Data ONTAP Version		
Platform	Data ONTAP 8.0.x	Data ONTAP 8.1.x
FAS/V2040	30TiB	50TiB
FAS2240-2	Unsupported	60TiB
FAS2240-4	Unsupported	60TiB
FAS/V3040	50TiB	50TiB
FAS/V3070	50TiB	50TiB
FAS/V3140	50TiB	75TiB
FAS/V3160	50TiB	90TiB
FAS/V3170	70TiB	105TiB
FAS/V3210	50TiB	75TiB
FAS/V3240	50TiB	90TiB
FAS/V3270	70TiB	105TiB
FAS/V6030	70TiB	105TiB
FAS/V6040	70TiB	105TiB
FAS/V6070	100TiB	162TiB
FAS/V6080	100TiB	162TiB
FAS/V6210	70TiB	162TiB
FAS/V6240	100TiB	162TiB
FAS/V6280	100TiB	162TiB

Since 64-bit aggregates are available in Data ONTAP 8.0 and later, the sum of the right-sized (usable) capacity of all data drives is used to determine the number of drives that can fit into a 64-bit aggregate.

The following recommendations apply to storage configurations that use 64-bit aggregates:

- NetApp highly recommends using 64-bit aggregates for aggregates that consist of large-capacity SATA drives.
- If your total usable system capacity for all data drives in a given storage configuration is not larger than 16TiB, NetApp recommends using 32-bit aggregates (primarily applicable to entry-level systems).
- The use of 64-bit aggregates with random read–intensive workloads might not be optimal.
- NetApp highly recommends using the appropriate sizing tools to determine if PAM is beneficial for your random read workloads.

The ability to do in-place 32-bit to 64-bit aggregate upgrades was introduced in Data ONTAP 8.1 and later. In Data ONTAP 8.0.x, it is not possible to convert 32-bit aggregates into 64-bit aggregates. This is important for any system that is being upgraded from Data ONTAP 7.3.x to 8.0, because all existing aggregates are considered 32-bit aggregates in Data ONTAP 8.0.x.

NetApp recommends reading TR-3786, "A Thorough Introduction to 64-Bit Aggregates," to gain additional insight into 64-bit aggregates.

4.5 FLEXIBLE VOLUMES

The addition of 64-bit aggregates in Data ONTAP 8.0 means that several new maximum-capacity points for aggregates are now supported by platform. Several factors must be kept in mind when creating flexible volumes (FlexVol volumes) in your aggregates:

- Maximum FlexVol volume size
- Deduplication FlexVol volume size limits
- Maximum number of systemwide FlexVol volumes

Section 2.6 of this document, "Aggregate and Flexible Volume Capacity," discusses maximum FlexVol volume size.

When using deduplication on a FlexVol volume, it is necessary to be aware of FlexVol size limits that are supported by deduplication. NetApp recommends reading TR-3505, "NetApp Deduplication for FAS and V-Series Deployment and Implementation Guide," for additional information.

Each controller has a maximum number of FlexVol volumes that are supported systemwide. In Data ONTAP 8.0.x, up to 500 FlexVol volumes are supported across all aggregates on all supported platforms, except the FAS2040, which supports 200 FlexVol volumes. In Data ONTAP 7.3.x and earlier, the number of supported FlexVol volumes is different for some platforms, for example, the FAS2000 series.

There are additional configuration considerations related to WAFL that you might need to consider when creating FlexVol volumes that are outside the scope of this document. NetApp recommends that you do additional research to understand WAFL configuration details and maximums.

4.6 RAID GROUP SIZING

The previous approach to RAID group and aggregate sizing was to use the default RAID group size. This no longer applies, because the breadth of storage configurations being addressed by NetApp products is more comprehensive than it was when the original sizing approach was determined. Sizing was also not such a big problem with only 32-bit aggregates that are limited to 16TB: You can only fit so many drives and RAID groups into 16TB. The introduction of 64-bit aggregates delivers the capability for aggregates to contain a great number of drives and many more RAID groups than were possible before. The compounds have the opportunity for future expansion as new versions of Data ONTAP support larger and larger aggregates.

Aggregates do a very good job of masking the traditional performance implications that are associated with RAID group size. The primary point of this policy is not bound to performance concerns but rather to establishing a consistent approach to aggregate and RAID group sizing that:

- Facilitates ease of aggregate and RAID group expansion
- Establishes consistency across the RAID groups in the aggregate
- Reduces parity tax to help maximize "usable" storage
- Reduces CPU overhead associated with implementing additional RAID groups that might not be necessary
- Considers both the time it takes to complete corrective actions and how that associates with the actual reliability data available for our drives

These recommendations apply to aggregate and RAID group sizing for RAID-DP®. RAID-DP is the recommended RAID type to use for all NetApp storage configurations. In Data ONTAP 8.0.1 the maximum SATA RAID group size for RAID-DP has increased from 16 to 20.

For HDD (SATA, FC, and SAS) the recommended sizing approach is to establish a RAID group size that is within the range of 12 (10+2) to 20 (18+2) and that achieves an even RAID group layout (all RAID groups containing the same number of drives). If multiple RAID group sizes achieve an even RAID group layout, NetApp recommends using the higher RAID group size value within the range. If drive deficiencies are unavoidable (as is the case sometimes), NetApp recommends that the aggregate not be deficient more than the number of drives equal to one less than the number of RAID groups (otherwise you would just pick the next-lowest RAID group size). Drive deficiencies should be distributed across RAID groups so that no single RAID group is deficient more than a single drive.

Given the added reliability of SAS and FC drives, it might sometimes be justified to use a RAID group size that is as large as 24 (22+2) if this aligns better with physical drive count and storage shelf layout.

SSD is slightly different. The default RAID group size for SSD is 23 (21+2), and the maximum size is 28. For SSD aggregates and RAID groups NetApp recommends using the largest RAID group size in the range of 20 (18+2) to 28 (26+2) that affords the most even RAID group layout, as with the HDD sizing approach.

4.7 MIXED CONFIGURATIONS

Although there are various mixed configurations that are supported by NetApp (and many that are not), there are some general best practices to keep in mind when considering, or attempting to avoid, mixed configurations:

- Always use homogeneous drives within aggregate configurations.
- If mixed configurations are necessary, minimize the deltas.
- Determine that an appropriate number of hot spares are on hand for the different types of drives populating the system.

NetApp highly recommends maintaining homogeneous drive configurations within aggregates. Keeping homogeneous aggregate configurations provides consistent performance from the aggregate and helps simplify administration. For example, it is then easier to understand how many hot spares of each drive type to keep on hand for the system.

If an aggregate does contain a mix of drives, it is best to minimize the number of differences. This is focused around drive speed and capacity. The more you mix different types of drives into an aggregate, the more variability you add into the possible issues that aggregate can face. For example, an aggregate can really only perform as well as its weakest link. If you mix 15,000-rpm drives with 10,000-rpm drives, you limit the performance of the aggregate as a whole. Although drive manufacturers sometimes state that the performance of fundamentally different drives is similar, when drives are actually deployed in a system the resulting performance is often not the same.

If hot spares of a specific drive type are not available, Data ONTAP might select a compatible drive to replace the failed drive. For example, if a 300GB drive fails and no 300GB hot spares are available on the system but a 450GB drive is, that 450GB drive will be used to replace the 300GB drive, and the 450GB drive will be artificially downsized to 300GB. If Data ONTAP is unable to find a hot spare that is compatible, the aggregate will not go into reconstruction and will continue to operate in a degraded mode.

4.8 SPARES POLICY

Spares recommendations vary by configuration and situation. In the past NetApp has based spares recommendations purely on the number of drives attached to a system. This is certainly an important factor but not the only consideration. NetApp storage systems are deployed in a wide breadth of configurations. This warrants defining more than a single approach to determining the appropriate number of spares to maintain in your storage configuration.

Depending on the requirements of your storage configuration you can choose to tune your spares policy toward:

- **Minimum spares.** In configurations in which drive capacity utilization is a key concern, the desire might be to use only the minimum number of spares. This option allows you to survive the most basic failures. If multiple failures occur, it might be necessary to manually intervene to make sure of continued data integrity. This recommendation is based on having spares available to address a double-disk failure situation.
- **Balanced spares.** The configuration is the middle ground between minimum and maximum. This assumes you will not encounter the worst-case scenario and will provide sufficient spares to handle most failure scenarios. This recommendation is based on having spares available to address two RAID groups that are doubly degraded (up to four disk failures of which no more than two have occurred in any RAID group).
- **Maximum spares.** This option makes sure that enough spares are on hand to handle a failure situation that would demand the maximum number of spares that could be consumed by a system at a single time. Using the term "maximum" is not stating that the system might not operate with more than this recommended number of spares. You can always add additional hot spares within spindle limits as you deem appropriate. This recommendation is based on having spares available to address two singly degraded RAID groups and two doubly degraded RAID groups. In this situation the two singly degraded RGs will have their reconstructions paused awaiting the completion of reconstructions in the two doubly

degraded RAID groups. This is the result of RAID reconstruction prioritization, which prioritizes doubly degraded RAID groups that are at higher risk of data loss for reconstruction over singly degraded RAID groups (at less risk of data loss compared to those doubly degraded RAID groups).

Selecting any one of the preceding approaches is considered to be the best practice recommendation within the scope of your system requirements. The majority of storage architects will likely choose the balanced approach, although customers who are extremely sensitive to data integrity might warrant taking a maximum spares approach. Given that entry platforms use small numbers of drives, a minimum spares approach would be reasonable for those configurations.

Note that spares of any kind are just one type of protection that can be used to provide increased storage system resiliency. Spares are not a replacement for the use of redundant components or mirrored technologies such as SyncMirror® (which protects against full shelf failures or triple+ disk failures within the same RAID group).

For RAID-DP configurations, consult Table 9 for the recommended number of spares.

Table 9) Determining recommended spares.

Recommended Spares		
Minimum	**Balanced**	**Maximum**
Two per controller	Four per controller	Six per controller
Special Considerations		
Entry platforms	Entry-level platforms using only internal drives can be reduced to using a minimum of one hot spare.	
RAID groups	Systems containing only a single RAID group do not warrant maintaining more than two hot spares for the system.	
Maintenance Center	Maintenance Center requires a minimum of two spares to be present in the system.	
>48-hour lead time	Remotely located systems have an increased chance of encountering multiple failures and completed reconstructions before manual intervention can occur. Spares recommendations should be doubled for these systems.	
>1,200 drives	For systems using greater than 1,200 drives, an additional two hot spares should be added to the recommendations for all three approaches.	
<300 drives	For systems using fewer than 300 drives, you can reduce spares recommendations for a balanced and maximum approach by two.	

Additional notes regarding hot spares:

- Spares recommendations are for each drive type installed in the system. See section 5.4, "Mixed Configurations," for more information.
- Larger capacity drives can serve as spares for smaller capacity drives (they will be downsized).
- Slower drives replacing faster drives of the same type will affect RAID group and aggregate performance. For example, if a 10k rpm SAS drive (DS2246) replaces a 15k rpm SAS drive (DS4243), this results in a nonoptimal configuration.
- Although FC and SAS drives are equivalent from a performance perspective, the resiliency features of the storage shelves in which they are offered are very different. By default Data ONTAP will use FC and SAS drives interchangeably. This can be prevented by setting the RAID option `raid.disk.type.enable` to on. See section 5.4, "Mixed Configurations," for more information.

HOT AND COLD SPARES

NetApp does not discourage administrators from keeping cold spares on hand. NetApp recommends removing a failed drive from a system as soon as possible, and keeping cold spares on hand can speed the replacement process for those failed drives. However, cold spares are not a replacement for keeping hot spares installed in a system.

Hot spares are also present to replace failed drives, but in a different way. Cold spares can replace a failed part (speeding the return/replace process), but hot spares serve a different purpose. That is to respond in real time to drive failures by providing a target drive for RAID reconstruction or rapid RAID recovery actions.

It is hard to imagine an administrator running into a lab to plug in a cold spare when a drive fails. Cold spares are also at greater risk of being "dead on replacement," as drives are subjected to the increased possibility of physical damage when not installed in a system. For example, handling damage from electrostatic discharge is a form of physical damage that can occur when retrieving a drive to install in a system.

Given the different purpose of cold spares versus hot spares, you should never consider cold spares as a substitute for maintaining hot spares in your storage configuration.

ENFORCING MINIMUM SPARES

The RAID option `raid.min_spare_count` can be used to specify the minimum number of spares that should be available in the system. This is effective for Maintenance Center users since when set to the value 2 it effectively notifies the administrator if the system falls out of Maintenance Center compliance. NetApp recommends setting this value to the resulting number of spares that you should be maintaining for your system (based on this spares policy) so the system notifies you when you have fallen below the recommended number.

4.9 DEGRADED AGGREGATES

An aggregate is considered in degraded mode if it has at least one RAID group that contains a failed drive that is not reconstructing. The primary reason an aggregate becomes degraded is that Data ONTAP is unable to find a compatible hot spare in the system for the reconstruction operation.

FlexVol performance for those FlexVol volumes that exist within the degraded aggregate will be affected while the aggregate remains degraded. This occurs because each read request requires an extra read from parity in order to calculate the missing block data for the affected RAID group. Because each read request equals two read requests, performance can be significantly affected by a degraded-mode aggregate. As a result, the more read intensive your workload is, the more of a performance effect you will see.

NetApp recommends maintaining an appropriate number of hot spares on a system so that degraded-mode situations are avoided. The section "Hot and Cold Spares" provides guidelines on the number of hot spares you should maintain in your storage configuration.

4.10 RAID GROUP TOPOLOGY

With DS14 it is reasonable to create a storage configuration that enables no more than two drives from each RAID group to be present in any shelf. This protects the RAID groups from a single shelf failure, because each RAID group (RAID-DP) would only be at risk of losing two drives in such a case. As denser shelves are offered by NetApp, such as the DS4243, this approach becomes increasingly more difficult to implement and maintain. NetApp highly recommends using SyncMirror to protect against storage shelf failures.

5 DRIVE CHARACTERISTICS

NetApp now ships FC, SAS, SSD, and SATA drives. It is important to understand some of the differences and similarities between these drives in order to select the right drives for your storage configuration. For detailed drive specifications, NetApp recommends acquiring the product datasheet from the specific drive manufacturer, because this information is the most up-to-date source of publicly accessible information for specific drives.

5.1 FC, SAS, AND SATA (HDD)

From a drive perspective, performance is primarily a function of speed and not capacity (at least initially). For example, a 2TB SATA drive spinning at 7,200 rpm will perform similarly to a 1TB drive spinning at 7,200 rpm.

Capacity becomes a factor when drives become full, because the heads have to travel farther to conduct reads and writes for larger capacity drives. As a result NetApp recommends that drive "fullness" be monitored. If your drives are getting full and a performance impact is noticeable, you might need to consider rebalancing your storage configuration by migrating data or adding additional storage.

Because the primary performance factor is speed, NetApp does not recommend that 10k rpm and 15k rpm drives be mixed in the same aggregate, even though by default mixing is enabled in Data ONTAP.

Aggregate performance is tied to the slower drives, which reduces the increased rotational speed benefit of the faster drives (since all data drives in the aggregate work together).

There is little difference in performance between FC and SAS drives. When planning your storage configuration, SAS drives can be used for the same types of workloads that are traditionally serviced by FC drives. It is important to note that competitors might be shipping like-capacity drives, but with slower speeds. Drives might also have an interface that supports a larger data "pipe," but this is still limited by the rotational speed of the drive.

Up to a 20% performance increase has been experienced when SATA drives use the newer DS4243 storage shelf versus the DS14mk2 AT storage shelf. This performance increase can be attributed to factors such as increased bandwidth, native command queuing (NCQ), and SAS-to-SATA bridging.

It is expected that SAS drive performance will start to improve over FC drive performance (they are equivalent today) as the industry migrates away from FC drives to SAS drives. Since FC drives are no longer revised to newer generations of drive technology and SAS drives are, the performance improvements for the newer drive generations will be seen for SAS drives and not FC drives.

5.2 DRIVE IOPS

Per-drive IOPS should *never* be used for sizing storage configurations. Sizing should always consider all components of a system as a whole. For this reason, all NetApp storage configuration sizing should be done by using the Sizer tools.

Per-drive IOPS can be helpful when observed from an industry standard perspective to help quickly estimate whether stated performance by a competitor is in the realm of possibility. A standard formula can be used in conjunction with drive specifications to determine what each drive's potential per-drive IOPS performance could be. The formula is as follows:

- 1 / ((average seek time / 1,000) + (average rotational latency / 1,000)) = estimated IOPS

Average seek time is the average of 100% random read seek time and 100% random write seek time. Both seek time and rotational latency are stated in milliseconds.

The following table is based on publically available enterprise drive specifications from Seagate (links to source specifications are embedded in the table), combined with the formula given earlier. Specific operating system features, workload, latency, stroke distance per operation, and many more factors ultimately decide what you see in real storage configurations. Note that like drives from different vendors can also vary in performance. The following table uses Seagate to keep the drive manufacturer consistent across the various drive types; however, using the formula you can work out per-drive IOPS for any drive manufacturer, provided you can determine the values needed for the formula.

Table 10) Estimated per-drive IOPS.

Estimated Enterprise Drive IOPS					
Type	Speed	Form Factor	Average Seek Time	Average Rotational Latency	Estimated Drive IOPS
SATA	7.2k rpm	3.5"	9.0	4.16	75
FC	10k rpm	3.5"	4.1	2.98	141
FC	15k rpm	3.5"	3.70	2.0	176
SAS	10k rpm	2.5"	4.10	3.0	140
SAS	15k rpm	3.5"	3.70	2.0	176

IOPS latency for the preceding table is a function of the average seek time and average rotational latency, which is the average time it takes the drive to complete an operation without any system-level overhead (for example, RAID).

SSD is not listed in the preceding table because the formula outlined above does not apply to solid-state technology (for example, no rotational latency). For reference, testing results have shown that NetApp SSD drives are capable of ~8,000 IOPS at 1.5ms or better response times for 4k 100% random read–based workloads. In fact, newer generations of SSD drives perform even better than the original drives that were used to establish the ~8,000 IOPS guidance.

5.3 SOLID-STATE DRIVE (SSD)

The introduction of NetApp SSD brings an entirely new drive type and technology into the storage options available for NetApp solutions. Although SSD specifications look very impressive, they still need to be attached to a system that can take advantage of them. The primary benefits of SSD are for read-intensive workloads (specifically random read). This is not to say there is no advantage to write workloads, but the advantage is not nearly as significant as it is for read workloads.

The drives are capable of much better performance on a per-drive basis, but they are still bound by the throughput capabilities of the systems to which they are attached. Low-latency operations are a key benefit of SSD.

A common question is if a storage configuration should use Flash Cache or SSD. Both can be used on the same system. SSD is automatically segregated from Flash Cache, or, put another way, Flash Cache only works with HDD.

Figure 4 demonstrates the more significant advantage afforded to read-based workloads when using (and paying for) SSD. Since writes are normally acknowledged by system memory for NetApp systems, the write benefit for SSD is potentially further limited when compared to 15k rpm HDD performance capabilities (and cost). Some NetApp competitors will rely more on SSD drives to provide increased write performance because they lack the functionality provided by WAFL or NVRAM.
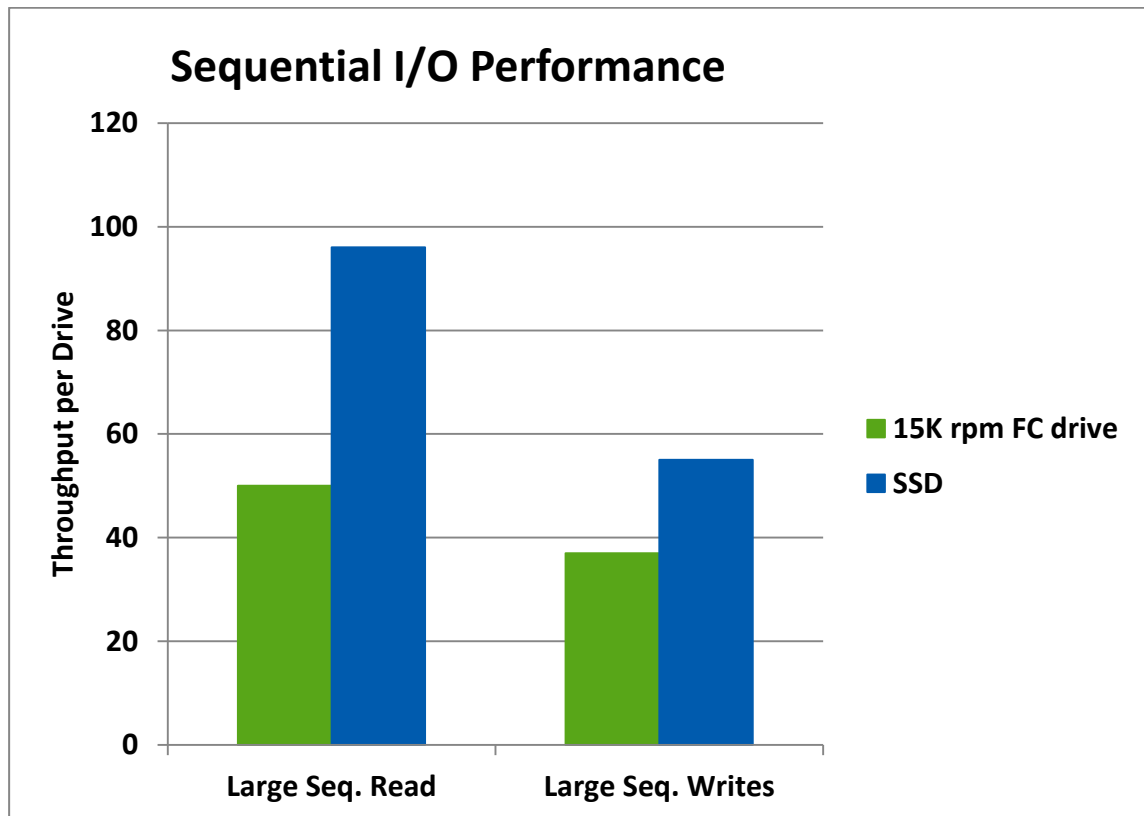

Figure 4) Per-drive sequential I/O performance comparison.

Attaching SSD to a controller does not mean that the maximum throughput for the controller is increased. SSD does provide low-latency I/O, but for throughput SSD configurations are bound to controller maximum throughput, just as with HDD. The difference is the number of drives it takes to achieve peak controller throughput performance. Table 11 shows the comparative drive count needed to achieve peak controller performance on a FAS3170 for various workloads.

Table 11) FAS3170 comparative drive count for peak controller throughput.

FAS3170 Comparative Drive Count for Peak Controller Throughput			
Workload	Peak Throughput	Number of 15k rpm HDDs	Number of SSDs
4k FCP random read	64,000 IOPS	215	8
64kB FCP sequential read	1,000MB/sec	20	8
32kB NFS sequential write	450MB/sec	12	8
OLTP	47,000 IOPS	98	11

SSD RELIABILITY

Although SSDs do not contain any moving parts, they are still relatively new to the enterprise storage industry. As a result the actual (AFR/AAR) reliability data available for SSDs is very limited. There is an expectation that SSDs are more reliable than HDDs during their service life, but a major differentiator of SSDs compared to hard drives (HDDs) is that SSDs have a use-based consumable operating life.

With HDDs, as long as the mechanical aspects of the drive continue to operate and the platters remain good, they can operate well beyond their warranty period (generally five years).

With SSDs there are a limited number of times that each memory block can be written to. Those blocks "wear" each time they are written to, which is known as write wear. For example, a block might be good for a minimum of 100,000 writes, after which it starts to degrade to a point that the block is no longer useful. SSDs are expected to be fully operational within their warranty periods.

NetApp recommends replacing all drives that have exceeded their warranty period. Given the use-based life of SSDs, this is especially important for SSDs that have been deployed in high-I/O configurations.

CHOOSING SSD AND FLASH CACHE

As mentioned earlier, there is no reason why both SSD and Flash Cache cannot be used in the same system. Table 12 shows the key differentiators to consider when determining if SSD or Flash Cache is a better fit for your storage configuration.

Table 12) Comparing SSD and Flash Cache.

Comparing SSD and Flash Cache	
SSDs in DS4243 **(Persistent Storage)** **Good Fit When …**	**Flash Cache** **(Intelligent Read Caching)** **Good Fit When …**
Random read–intensive workload	Random read–intensive workload
Every read must be fast	Improving average response time is adequate
Active data is known and fits into SSD capacity	Active data is unpredictable or unknown
Active data is known, is dynamic, and ongoing administration is okay	Administration-free approach desired
Upside of write acceleration desired	Minimizing system price is important
Performance must be consistent across failover events	Accelerating an existing HDD configuration is desired

5.4 MEAN TIME BETWEEN FAILURE (MTBF)

Drive mean time between failure (MTBF) is the predicted elapsed time between inherent failures of a drive during operation. MTBF information is readily available from manufacturer Web sites but is consistent across the common disk types shipped today.

Table 13) Manufacturer-stated MTBF by drive type.

Drive MTBF by Drive Type	
Drive Type	**MTBF**
SATA	1.2 million hours
FC	1.6 million hours
SAS	1.6 million hours
SSD	2.0 million hours

Drive manufacturers use this as a metric to determine the quality of a drive before it enters into production. After a drive is put into production, a better measure of its quality is the annualized failure rate (AFR), which is based on real-world drive failures.

[TR-3437: Storage Subsystem Resiliency Guide](#) contains more information about MTBF and the reliability of hardware components.

6   DRIVE-RELATED OPERATIONS

This section addresses some of the key drive-related operations that are core to Data ONTAP today.

6.1 ZEROING

Drives are zeroed prior to being added to new or existing aggregates (unless they are already zeroed) or when the storage administrator executes the `disk zero spares` command. Zeroing times for each drive that is qualified by NetApp are recorded during the qualification process. Although drive zeroing times vary by drive manufacturer, you can use Table 14 as a guideline for estimating drive zeroing times.

Table 14) Estimated drive zeroing time by drive type, speed, and capacity.

Drive Zeroing Time Estimates			
Capacity	**Type**	**Speed**	**Estimated Zeroing Time (hrs)**
100GB	SSD	-	0.1
300GB	FC	15k rpm	1.5
450GB			2.2
600GB			2.5
300GB	SAS	15k rpm	1.5
450GB			2.2
600GB			2.5
450GB	SAS	10k rpm	2.3
600GB			2.6
500GB	SATA	7.2k rpm	2.5
1TB			4.3
2TB			5.6
3TB			7.0

Zeroing hot spares is not a prerequisite for RAID reconstruction or rapid RAID recovery operations.

6.2 RAPID RAID RECOVERY

Data ONTAP always monitors drives for many reasons. A component of the result of this monitoring is to determine when a drive is about to encounter a failure based on various predetermined thresholds and alerts. This process is known as predictive failure analysis. A failed drive will trigger a RAID reconstruction that might take some time to complete. If the system determines that a drive is about to fail, a drive copy operation will complete more quickly than a RAID reconstruction that has the added requirement of calculating RAID parity information for the failed drive.

Rapid RAID recovery is a feature of Data ONTAP that is triggered by predictive failure analysis. The rapid RAID recovery process does a block-level copy of the identified drive's data to a hot spare. This process is similar to a RAID reconstruction but without the need to calculate RAID parity information, because this occurs before the drive fails. Similar to a RAID reconstruction, this process affects foreground I/O. This performance effect can be managed using the same `raid` options that apply to RAID reconstructions.

Rapid RAID recovery operations are independent of RAID recovery operations, although they do draw from the same pool of hot spares. The number of RAID groups that can undergo a rapid RAID recovery is four. The option `raid.disk.copy.max_count` can be used to change the number of RAID groups that can undergo a rapid RAID recovery. The default value is four and can be changed to from one through four. Rapid RAID recovery can be enabled and disabled using the `raid.disk.copy.auto.enable` option. By default this feature is enabled (option set to `on`).

Table 15 can be used as a guideline for rapid RAID recovery times on an idle system. Several factors can determine the timing of these types of drive operations such as system configuration, I/O workload, RAID options, and more.

Table 15) Estimated rapid RAID recovery time by drive type, speed, and capacity.

Rapid RAID Recovery Time Estimates			
Capacity	Type	Speed	Estimated Rapid RAID Recovery Time (hrs)
100GB	SSD	-	0.2
300GB	FC	15k rpm	1.0
450GB			1.5
600GB			1.8
300GB	SAS	15k rpm	1.0
450GB			1.5
600GB			1.8
450GB	SAS	10k rpm	2.5
600GB			3.8
500GB	SATA	7.2k rpm	2.8
1TB			5.5
2TB			12.8
3TB			18.3

6.3 RAID RECONSTRUCTION

A RAID reconstruction occurs when a drive fails, provided a hot spare is available on the system. If no hot spares are available on the system or a compatible hot spare cannot be located for the failed drive, a RAID reconstruction process will not initiate. An aggregate containing a RAID group with the failed drive that is not in reconstruction becomes degraded. Degraded aggregates are discussed in the "Degraded Aggregates" section of this document.

NetApp highly recommends the use of RAID-DP in all storage configurations because this affords you a much higher level of storage resiliency than you have when using RAID 4. The purpose of this document is not to persuade you that RAID-DP is better than RAID 4 (and other RAID levels), but rather to reinforce that this is the RAID choice for NetApp storage systems and something this document assumes as a baseline throughout.

As drive capacities increase, the time it takes for those drives to reconstruct becomes longer. The longer it takes drives to reconstruct, the longer your systems are at risk of encountering additional errors that could become unrecoverable. This is a fundamental fact that affects all network storage vendors today. Data ONTAP includes two options that allow a storage administrator to adjust the effect that a RAID reconstruction has on system performance.

The first option is `raid.reconstruct.perf_impact`, which by default is set to medium. The three possible values for this option are low, medium, and high. A setting of low for this option increases the time it takes to complete a RAID reconstruction while minimally affecting foreground I/O performance. Setting this option to high decreases the time it takes to complete a RAID reconstruction while reducing foreground I/O performance.

The second option is `raid.reconstruct.threads`, which by default is set to four. This is a hidden option that can be used to increase or decrease the number of RAID reconstruction threads that can be spawned on a system. This option can be set to from one through eight. Decreasing the number of threads available for RAID reconstruction increases the amount of time it takes to complete a RAID reconstruction. Increasing the number of threads can reduce RAID reconstruction time but at the expense of foreground I/O performance.

Some situations might warrant adjusting these options, but this should be a last resort. NetApp recommends keeping default values for these options.

The two options discussed earlier are clearly related to performance. A third option exists in Data ONTAP that can affect performance and storage resiliency. The option is `raid.reconstruct.max_groups`, which by default is set to two. This option can be set to from one through eight. Given the default value of two, this means that up to four drives from two different RAID groups could be in reconstruction at a given time. Increasing this option's value will affect foreground I/O performance should additional (to the default value) RAID groups enter into RAID reconstruction. Changing this option will also affect the number of hot spares you will want available on the system. NetApp recommends keeping the default value for this option.

Data ONTAP has built-in intelligence that understands that RAID groups that are doubly degraded are at risk of data loss. As a result, the reconstruction processes of at-risk RAID groups are prioritized to be completed more quickly than normal. This is not the same as adjusting the performance-related `raid` options but rather a different approach that Data ONTAP takes to manage reconstruction time. As a result you will notice that a double-drive reconstruction might complete prior to a single-drive reconstruction (assuming similar drives are being reconstructed).

In Data ONTAP 7.3.2 and later, "at-risk" RAID groups—those that are doubly degraded—are prioritized automatically to reconstruct before singly degraded RAID groups. For example, if two RAID groups that are singly degraded are in reconstruction and another RAID group becomes doubly degraded, one of the singly degraded RAID group's reconstruction processes will be paused, and the doubly degraded RAID group will start reconstruction. This behavior is not configurable. A doubly degraded RAID group will never be paused during reconstruction due to prioritization. Prioritization will not occur if insufficient spares are available for reconstruction. Since hot spares are allocated at the start of the reconstruction process, they cannot be redirected for use in other reconstructions if there are insufficient hot spares available. Therefore, it is imperative to consider this when determining the number of hot spares to maintain for your storage configuration.

Once hot spares have been selected and reconstruction initiated, it is not possible to redirect those drives to other reconstructions.

Table 16 can be used as a guideline for RAID reconstruction times on an idle system. Several factors can determine the timing of these types of drive operations such as system configuration, I/O workload, RAID options, and more.

Table 16) Estimated RAID reconstruction time by drive type, speed, and capacity.

RAID Reconstruction Time Estimates			
Capacity	**Type**	**Speed**	**Estimated RAID Reconstruction Time (hrs)**
100GB	SSD	-	0.3
300GB	FC	15k rpm	2.0
450GB			2.3
600GB			2.8
300GB	SAS	15k rpm	2.0
450GB			2.3
600GB			2.8
450GB	SAS	10k rpm	3.8
600GB			4.4
500GB	SATA	7.2k rpm	5.2
1TB			9.3
2TB			18.5
3TB			27.8

Testing has shown that systems under load can increase the base reconstruction times shown in Table 16 by a multiplier of three. For example, one series of testing on a FAS3040 using SIO to simulate a 32KB large sequential workload over NFS with a 70% read-to-write ratio and a RAID group size of 14 returned a reconstruction time of 31 hours for a 1TB drive.

6.4 RAID SCRUBS

Data that is read by Data ONTAP is checked against parity to enable data integrity. This approach is very effective for enabling data integrity of frequently read data but does not benefit data that is not read often. For example, archival data can sit on drive for long periods of time without being accessed. Data in this state is known to be at rest.

RAID scrubs, also known as parity scrubs, are used to enable the integrity of data at rest. This is a process that traverses data at rest and triggers reads on that data. As a result of triggering the read, the data is checked against parity to determine that it is correct. If a block is found to be incorrect, the block is marked as bad and the data recreated from parity, then written to a new block. RAID scrubs minimally affect foreground I/O. Testing suggests that this effect is less than 10% on average. If you want to further limit the effect RAID scrubs could have on foreground I/O, NetApp recommends scheduling RAID scrubs to occur during off-peak hours.

There are several options that can be used to configure RAID scrubs:

- **`raid.scrub.enable.`** By default, RAID scrubs are enabled. RAID scrubs can be disabled by setting the option `raid.scrub.enable` to off. NetApp highly recommends keeping RAID scrubs enabled.
- **`raid.scrub.duration.`** RAID scrubs will run for 6 hours each week as indicated by the default value of 360 (minutes). NetApp recommends using the default value for this option.
- **`raid.scrub.schedule.`** This allows you to schedule RAID scrubs to occur more frequently than once per week. For systems using large-capacity SATA drives, NetApp recommends running RAID scrubs more than once per week. This option can also be used to schedule RAID scrubs to occur during off-peak hours to minimize system performance effect.
- **`raid.scrub.perf_impact.`** RAID scrubs are intended to have low effect on system performance, reflected in the default setting of low for this option. If you want RAID scrubs to be more aggressive, you can set this option to medium or high. NetApp recommends using the default value for this option, but in some storage configurations, for example, in archival environments, it might make sense to change this option.

6.5 BACKGROUND MEDIA SCANS

Background media scans, also known as media scrubs, are a drive diagnostic feature that is used to detect media errors. The frequency of this type of scrub is controlled by RAID. If a media error is detected, Data ONTAP takes corrective action. This includes marking the block as bad and recreating the data from parity. The recreated data is then written to a new block.

Table 17 shows the differences and similarities between background media scans and RAID scrubs.

Table 17) Differences between background media scans versus RAID scrubs.

Background Media Scans Versus RAID Scrubs	
Background Media Scan	**RAID Scrub**
Drive diagnostic feature	RAID-level feature
Checks for media errors that make blocks unreadable, not for correctness of data	Checks the correctness of the data against parity information
Very low effect on system performance (less than 4%) because this is a drive diagnostic feature that does not require system resources to run	Low effect on system performance (less than 10%) because some system resources are needed for the scrub
Scrubs hot spares to detect bad blocks	Does not scrub hot spares
Errors are resolved by recreating data from parity and writing the recreated data to a new block or a spare drive by simply marking the block as bad	Errors are resolved by recreating data from parity and writing the recreated data to a new block

RAID scrubs are not a replacement for background media scans and vice versa. These features should be used in conjunction to enable integrity of data on drive.

There are three options that are used to configure background media scans:

- **`raid.media_scrub.enable.`** This option can be set to on or off. By default, background media scans are enabled. NetApp recommends using background media scans to increase system resiliency.
- **`raid.media_scrub.rate.`** The valid values for this option range from 300 to 3,000. A rate of 300 represents a media scrub of approximately 512MiB per hour, and 3,000 represents approximately 5GiB per hour. The default value for this option is 600, which is a rate of approximately 1GiB per hour. NetApp recommends using the default value of 600.
- **`raid.media_scrub.spares.enable.`** This option can be used to include or exclude background media scans on hot spares. NetApp recommends using the default value of on so that hot spares do not contain bad blocks.

6.6 MAINTENANCE CENTER

Maintenance Center provides configurable in-place drive diagnostics to determine the health of suspect drives. If Data ONTAP drive health monitoring determines that a drive has surpassed an error threshold, rapid RAID recovery is initiated to a hot spare. Afterward, the suspect drive can be placed into Maintenance Center, where it undergoes a series of diagnostic tests. If diagnostic testing shows drive health to be normal and the error condition to be an anomaly, then the drive is returned to the spares pool. If diagnostics do not show normal drive health, or, by default, if the drive is in Maintenance Center for a second time, then Maintenance Center flags the drive as broken, and an RMA process is initiated. Maintenance Center requires that at least two hot spares are present in the system. NetApp recommends using Maintenance Center to provide additional intelligence for drive error discovery and correction.

6.7 LOST WRITE PROTECTION

In rare circumstances, a drive media error can lead to a situation in which a write to drive fails but the drive itself thinks it has succeeded. As a result the drive informs Data ONTAP that the write has been successful. This situation is known as silent data loss. WAFL has a unique feature called write signature, also known as lost write protection, that protects against this kind of silent data loss. It is worth noting that checksums do not protect against silent data loss. The next time the block is read, WAFL will detect the error and automatically recreate the data from parity, then write the data to a new block after marking the original block as bad. This check does not affect system performance and increases system resiliency.

7 STORAGE SHELVES

The introduction of the FAS2000 series systems marked the addition of serial-attached SCSI (SAS) to the NetApp product line. The DS4243 marks the introduction of SAS to NetApp's external storage shelves. There are different considerations to take into account depending on if you use FC-AL–based shelves (for example, DS14) or SAS-based shelves (for example, DS4243) in your storage configuration.

7.1 SAS STORAGE SHELVES

Bandwidth is not a primary concern with SAS-based storage shelves. For example, with a DS4243-based storage configuration using IOM3 shelf modules, each SAS port has four lanes, and each lane runs at 3Gb/sec. That is a total of 12Gb/sec per SAS port. In multipath HA (MPHA) configurations this can be doubled to 24Gb/sec to each shelf in a SAS stack. Future technology implementations will see the 3Gb/sec lanes go to 6Gb/sec or even 12Gb/sec. Given this amount of bandwidth, the limiting factor in systems using SAS-based shelves will more likely be controller resources. The addition of Alternate Control Path (ACP) functionality, which removes management traffic from the data path, further increases the efficiency of the data path bandwidth in SAS stacks.

SAS EXPANDERS

A key advancement with SAS-based shelves is the use of a SAS expander in the shelf module. SAS expanders provide two very key features:

- The ability to establish direct, physical connections to each drive in the shelf. This means that drives are physically separated from each other, and I/O occurs only between the target drive and the shelf module.
- Switching intelligence that allows a SAS expander to identify and control situations that would potentially interrupt I/O on the stack. For example, in a situation in which a drive is causing a broadcast storm, the SAS expander will not forward that traffic within the SAS stack (drives are physically segregated from each other). In an FC-AL loop, a broadcast storm can bring down the entire loop, because the ESH modules don't have this kind of routing logic, and drives are not segregated from each other.

ALTERNATE CONTROL PATH

Although not required, NetApp highly recommends using ACP. ACP provides new capabilities for proactive, nondisruptive recovery of shelf modules. ACP does not add a point of failure to the data path and is not a replacement for management servers (for example, CIM or SMASH). Two key benefits of using ACP are:

- For a single-failure situation in an MPHA configuration, the recovery will occur without I/O interruptions.

- For a multiple-failure situation or single-path configuration, the recovery will incur I/O interruptions, but the alternative of a system panic will be prevented.

The current implementation of ACP supports the reset and power cycle of SAS expanders. Only a single Ethernet port is required per controller, regardless of the number of stacks present in the storage configuration. NetApp highly recommends dedicating a single Ethernet port per controller to enable ACP in order to increase system resiliency.

7.2 FC-AL STORAGE SHELVES

NetApp currently supports FC-AL speeds of 1Gb/sec, 2Gb/sec, and 4Gb/sec. The speed your loop will run at depends on the HBA port, drive, shelf, shelf module, and small-form-factor pluggable (SFP) module being used, because it is possible to mix components that support different speeds within the same configuration. Resulting loop speed will be based on the slowest component present in the configuration.

In FC-AL–based storage configurations, NetApp recommends maximizing the number of loops connected to your storage shelves. This is because loop bandwidth is a concern for system performance due to the following factors:

- FC-AL–based storage shelves use a virtual connectivity architecture. Each drive in the shelf is connected to the other (drives are not physically separate from each other). Increasing the drives present in a loop means there is more overhead to maintain the loop.
- Drive-related activities such as RAID reconstructions, RAID scrubs, drive firmware downloads, and more compete with foreground I/O for loop bandwidth.
- Drives present in highly saturated loops might be underutilized as the loop becomes the bottleneck to system performance.

NetApp recommends monitoring loop bandwidth in order to identify and resolve potential performance bottlenecks.

MIXING FC SHELF MODULES

It is possible to mix certain shelf modules in various configurations. Prior to doing so, it is important to understand the differences in these shelf modules. The following best practices apply to mixing shelf modules together in the same loop:

- Whenever possible, use homogeneous shelf modules.
- If ESH2 and ESH4 shelf modules are mixed together, the resulting loop speed is 2Gb/sec, and the module resiliency features of the ESH4 are bypassed.
- Do not mix ESH shelf modules with ESH2 or ESH4 shelf modules.
- Mixing ESH shelf modules with ESH4 shelf modules is unsupported.
- LRC shelf modules should not be mixed with ESH, ESH2, or ESH4 shelf modules.

When mixing shelf modules in the same loop in supported configurations, determine that the loop speed is set to match the slowest component speed for the shelf, shelf module, SFP, drive, or HBA port.

7.3 NONDISRUPTIVE SHELF REPLACEMENT (NDSR)

NDSR is supported in local SyncMirror configurations for configurations using DS14mk2, DS14mk4, DS4243, and DS2246 storage shelves on systems running Data ONTAP 7.3.2 or later. Given that software has no ability to enforce which shelves are physically removed from a configuration, NetApp recommends approaching NDSR with due diligence and caution to make sure storage is not removed from a system prematurely or inadvertently.

If the storage shelf being replaced contains spare drives, it is important to make sure that enough spares to provide continued resiliency on the system are available in the remaining shelves. If spare drives exist on the shelf to be replaced, Data ONTAP will display "disk missing" messages for them, but this is not critical because these drives do not belong to an aggregate. After the shelf is replaced and power restored, Data ONTAP will show that the drives are "found."

The NDSR procedure depends on a supported and properly configured configuration. Make sure of compliance with the following prerequisites and best practices:

- HA pair controller configuration
- Data ONTAP 7.3.2 or greater

- MPHA cabling, which can be checked by running the `sysconfig` command:

```
fas3040c-svl01> sysconfig

NetApp Release 7.3.2

System ID: 0118045690 (fas3040c-svl01)

System Serial Number: 1082421 (fas3040c-svl01)

System Rev: A1

System Storage Configuration: Multi-Path HA
```

- The storage shelf to be removed contains a fully mirrored aggregate/plex
- Software disk ownership is properly identified and consistent for all drives
- The individuals conducting the replacement are properly grounded to protect against ESD
- All failed drives have been replaced and the system is in a normal operating state
- Confirm that NetApp Global Support (NGS) believes it is safe to proceed with the NDSR

NDSR should not be conducted without first consulting with NGS. The specific procedure for conducting this task falls outside the scope of this document and may be obtained through contacting NGS.

8 CONCLUSION

There are many moving parts to consider when planning your storage configuration. The best practices, recommendations, and guidelines covered by this document are applicable to the largest and smallest of storage configurations. Now that you have a fundamental understanding of many of the factors at play in the storage subsystem, NetApp recommends reading TR-3437, "Storage Resiliency Best Practice Guide," in order to gain the next level of perspective.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

Go further, faster®

www.netapp.com