DELL EMC

Technical White Paper

Dell EMC SC Series: Distributed Spare

Enable faster recovery from disk failures, increased RAID performance, and extended disk life

Abstract

This document describes the distributed spare feature of Dell EMC™ SC Series storage and details the benefits and performance improvements.

May 2019

Revisions

Date	Description
March 2018	Initial release
April 2019	Updated details in section 3; updated to new template

Acknowledgements

Author: David Glynn

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Table of contents

Revisions....2

Acknowledgements....2

Executive summary....4

1 Introduction....5

2 Understanding distributed spare....6

2.1 Traditional RAID hot spare disk layout....6

2.2 Distributed spare disk layout....7

2.3 Distributed spare rebuild operation....7

2.4 Redistribution of data after failed drive replacement....9

2.5 Avoiding RAID rebuilds....9

3 Enabling distributed spare....11

3.1 Pausing the distributed spare optimizer process....11

4 Reduction in rebuild times....13

5 Additional benefits of distributed spare....14

5.1 Extended longevity of disks....14

5.2 Increased overall RAID performance....14

A Technical support and resources....15

A.1 Related resources....15

Executive summary

Today’s business runs on data, and business applications are built from the ground up with redundancy to maximize the availability of business data and minimize the impact of downtime. This is achieved in many ways, but when it comes to the disk subsystems in arrays, it is accomplished using RAID.

RAID has been used for over thirty years to protect data from disk failures. However, increasing disk sizes have resulted in longer rebuild times when recovering from a failed disk. The RAID distributed spare configuration significantly shortens rebuild operations by spreading the write and read rebuild workload across multiple disks in parallel. This allows full data protection and redundancy to be achieved in a much shorter time frame.

Distributing spare capacity across all disks enables the following benefits:

Shorter RAID rebuild times: When a disk fails on a distributed spare system, the data on the failing disk rebuilds to space allocations on multiple disks. This allows multiple rebuild tasks for a failing disk to run in parallel as the write I/O workload targets multiple disks.

Extended longevity of disks: Portions of each disk are set aside as spare capacity which under-provisions each disk. This under-provisioning provides the disks internal firmware with additional capacity to perform wear-leveling operations.

Increased overall RAID performance: In a hot spare configuration, the dedicated hot spare disk stores no data and does not contribute to the I/O capabilities of the storage system. With distributed sparing, all disks store user data. I/O performance increases because the Dell™ Storage Center OS (SCOS) takes advantage of the additional disks on the back end.

1 Introduction

The primary benchmarks that IT departments are judged by are application availability and application performance. How these are achieved varies greatly depending on where they are implemented in the software or hardware stack. Clustered applications, load balancers, multiple power circuits, and error-correcting memory are just a few example solutions used to help achieve these benchmarks. Ultimately, it comes down to using some level of system redundancy.

Storage arrays are designed for redundancy with redundant controllers, power supplies, fans, back-end SAS chains, and multiple front-end data ports. In addition to these capabilities, SC Series arrays also can provide complete array redundancy with Live Volume Auto Failover, which enables continued data access in the unlikely event of an unplanned failure of an entire site. However, the key factor of redundancy is the array's drives that store critical business data. With any array-based disk subsystem, protecting business data in the event of disk failure comes down to using RAID.

2 Understanding distributed spare

The examples in this section compare traditional hot spare and distributed spare configurations. No particular RAID type is depicted.

SC Series arrays perform RAID at the page level, not at the drive level. Unlike most RAID configurations, physical drives in an SC Series array store data in multiple RAID formats. In the following illustrations, user data and parity or mirror data are co-mingled and are represented by grey blocks, while hot spare or distributed spare data is represented in blue.

Note: For the purposes of explaining distributed sparing, the examples in this document do not fully reflect the sophistication and complexity of the RAID architecture within SC Series arrays. For more detailed information on how RAID is implemented in SC Series storage, see the document, Understanding RAID with Dell SC Series Storage.

2.1 Traditional RAID hot spare disk layout

In all its implementations, excluding striping, RAID provides access to data using either mirroring or parity-based algorithms, in the event of disk failure. This enables data redundancy, increased performance, and both in some cases. Once a disk has failed, it must be replaced before full redundancy can again be achieved. To speed up the disk-replacement process, arrays can include dedicated hot spare disk drives (Figure 1). This allows the rebuild process to begin imminently, rather than requiring time for the drive to be replaced by a data-center technician. This enables full redundancy to be achieved sooner. During this rebuild operation, all rebuild writes are going to a single disk, which limits the rebuild rate to the speed at which this single disk can be written to.

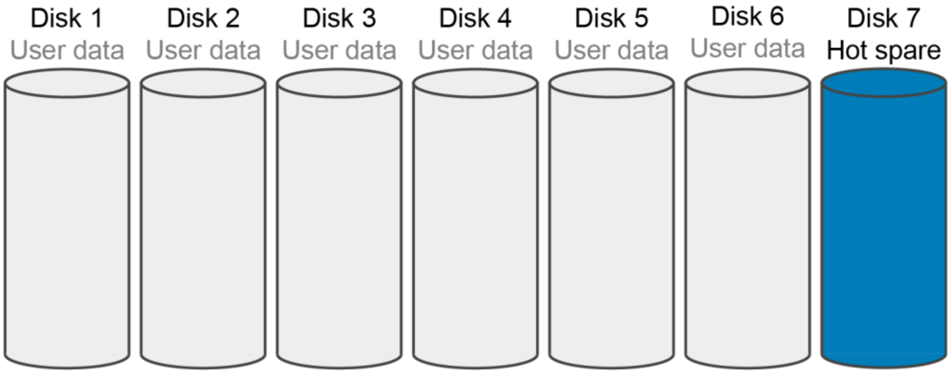


Figure 1 Traditional RAID with dedicated hot spare disk (blue)

Similar to Moore's law of increasing transistor densities, disk drive sizes have also increased rapidly over time. Larger and larger hard drives result in longer and longer rebuild times. Should an additional drive with a rebuild time greater than the failure tolerance become unavailable during the rebuild process, there is potential for data unavailability or data loss to occur, depending on the array and drive configuration.

2.2 Distributed spare disk layout

The primary intent of the distributed spare configuration is to resolve the growing problem of longer rebuild windows caused by larger hard drives. The most significant bottleneck in the failed-disk rebuild operation is the single hot spare disk drive upon which the data is being written. Data can only be written so quickly to a single disk drive. The typical solution to increasing drive write speed is to spread or distribute the workload by writing to more drives. This is addressed in advance of disk failure by changing the way that data is laid out on the disks.

To achieve this goal, two significant changes are made. First, the dedicated hot spare disk is replaced with reserved rebuild space on each of the other disk drives. The total number of drives and the amount of usable space in the array does not change; the only change occurs in the layout of the spare, or reserve space. Rather than placing all sparing capacity on dedicated hot spare disks, the spare capacity is spread across all disks. The second change optimizes how user data, RAID parity data, and mirror data are laid out across the disks. During disk rebuild operations, this optimization allows data to be read from many more disks — no single disk is subject to an excessive read load.

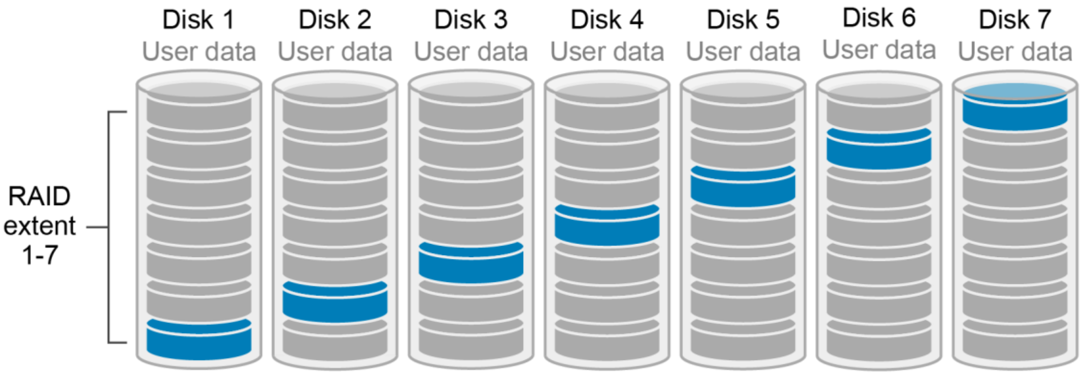


Figure 2 Distributed spare space (blue) reserved on each disk drive

In result, many more disks, (or all the disks in the disk class) are the source of reads and the destination of writes during a disk rebuild. This removes the single-disk write-speed limitation of the traditional RAID hot spare, and allows data to be moved quickly across the entire back end of the SC Series array. To prevent a rebuild operation from saturating the array back end to the point of impacting front-end I/O, front-end I/O is constantly monitored for degradation in latency. Any degradation in latency is countered by adjusting the data rebuild operation. This balances the needs of the rapid data rebuild without sacrificing the performance of business applications.

2.3 Distributed spare rebuild operation

A common IT priority involves designing systems around an inevitable failure so that downtime does not occur. To illustrate how a disk failure is addressed in a distributed spare configuration, Figure 3 shows that disk 3 has failed. No data has been lost or become unavailable due to RAID protections. However, if a single redundant RAID type has been used (such as RAID 10 or RAID 5), the failure of a second disk could put data at risk of becoming inaccessible.

Note: Dual-redundant storage protects against the loss of any two disks. SC Series HDDs larger than 1.9 TB should use dual redundancy, and in some cases, it is mandated. Storage tiers assigned as dual-redundant can contain any of the following types of RAID storage: RAID 10 dual mirror (data is written simultaneously to three separate disks), RAID 6-6 (four data segments with two parity segments for each stripe), and RAID 6-10 (eight data segments with two parity segments for each stripe).

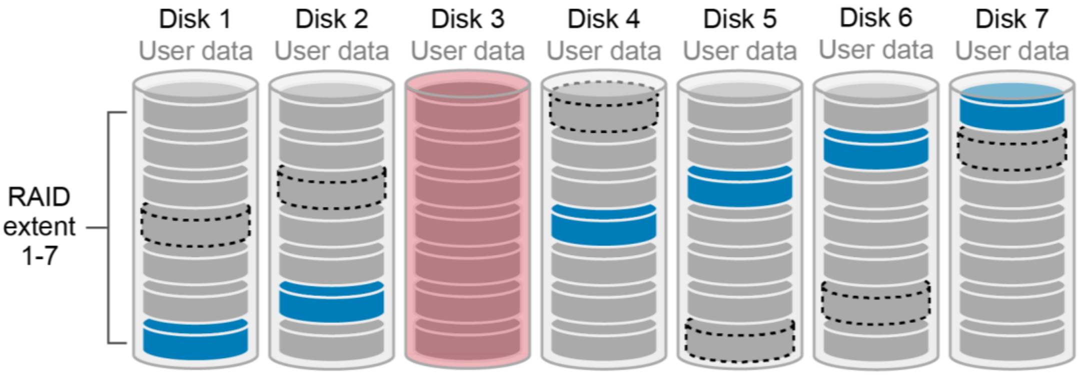


Figure 3 Disk 3 has failed with automatic RAID rebuild pending

Immediately, and without external intervention, the SC Series array will begin a rebuild operation to restore full data redundancy as shown in Figure 4. Since the SC Series array stores data on each physical disk in multiple RAID formats, the rebuild operation consists of mirror copy operations and parity checksum operations to recreate the lost data. The recovered data is written to the RAID extent's respective distributed spare capacity.

In this simplified example showing just mirrored data, RAID extent 1 has lost the data it had on failed disk 3. However, this data was mirrored to extent 1 on disk 4. Therefore, it can be reprotected by remirroring extent 1 on disk 4 to extent 1 on disk 7. The same rebuild process occurs for RAID extents 2 through 7, with each rebuild operation's writes targeting a different physical disk, and each rebuild operation's reads coming from a different physical disk. All rebuild operations occur in parallel. Once data has been rebuilt on the target disk, the data can now be the source of reads, if reading from this disk is faster than reading from the other disks that contain this data. Writes are serviced by these disks as the rebuild continues.

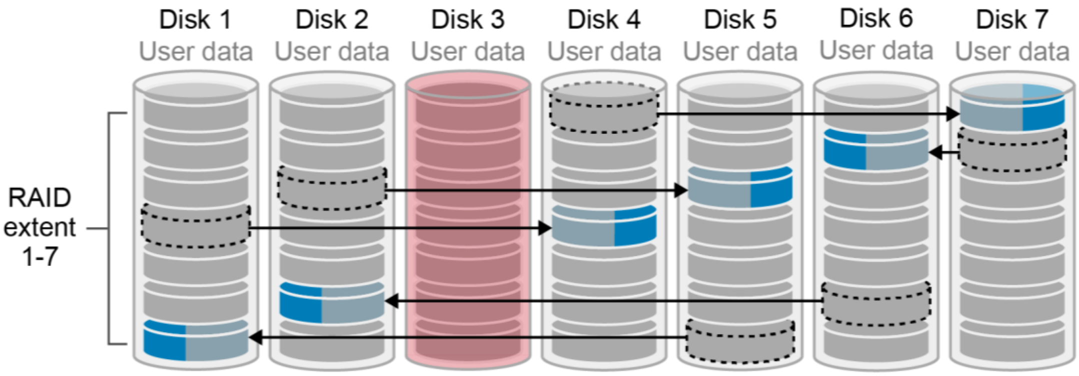


Figure 4 RAID extent rebuild in progress (only RAID 10 shown for clarity)

In most SC Series array configurations, all disks of the same class as the failed disk are involved in the data-rebuilding process. This shifts the rebuild process limitation from the write speed of a single drive to the speed in which the data can be moved across the array's back end.

2.4 Redistribution of data after failed drive replacement

Once the failed drive has been replaced, the array needs to incorporate this new disk and its capacity back into the array and the overall system capacity. Since a complete set of the user data exists, including all parity data and mirror data, no effort needs to be spent on rebuilding data. Rather, the extents that were previously rebuilt are mirrored to the new disk, freeing up capacity on existing disks and redistributing the spare capacity across multiple disks.

During the redistribution of spare capacity, data can be read from multiple disks. In the previous example (Figure 3), when extent 4 of disk 1 lost its mirror due to the failure of disk 3, this data was then mirrored to extent 4 of disk 4. To redistribute the extent 4 data to consume space on disk 3 and free up space for sparing on disk 4, the data being mirrored to extent 4 of disk 3 can be read from either location of that data: disk 1 or disk 4 (Figure 4).

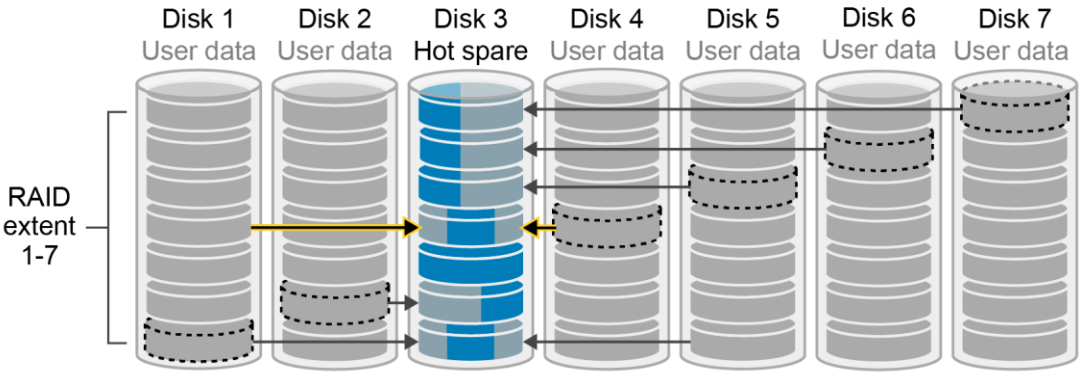


Figure 4 RAID extent optimization in progress

This redistributing of space capacity is completed as a background process since full redundancy has already been achieved. This also minimizes the impact to production I/O. Once complete, the resulting data and spare capacity layout (Figure 5) is similar to the original data and spare capacity layout in Figure 2.

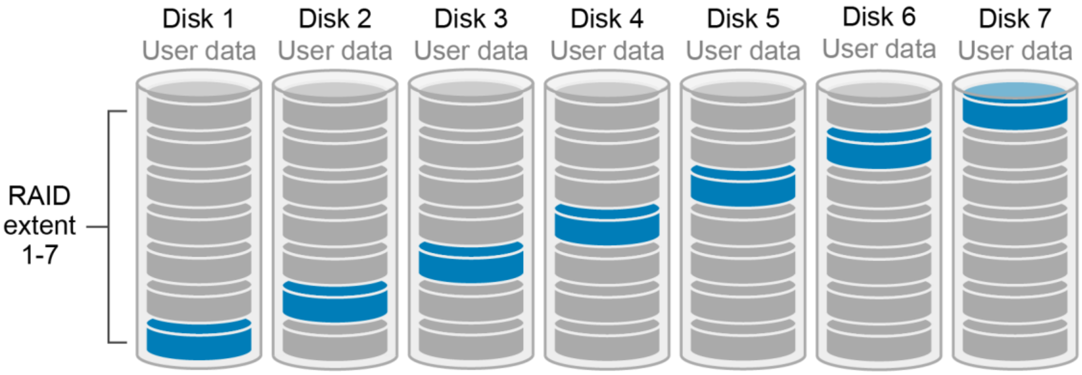


Figure 5 Sample data layout after distributed spare optimization

2.5 Avoiding RAID rebuilds

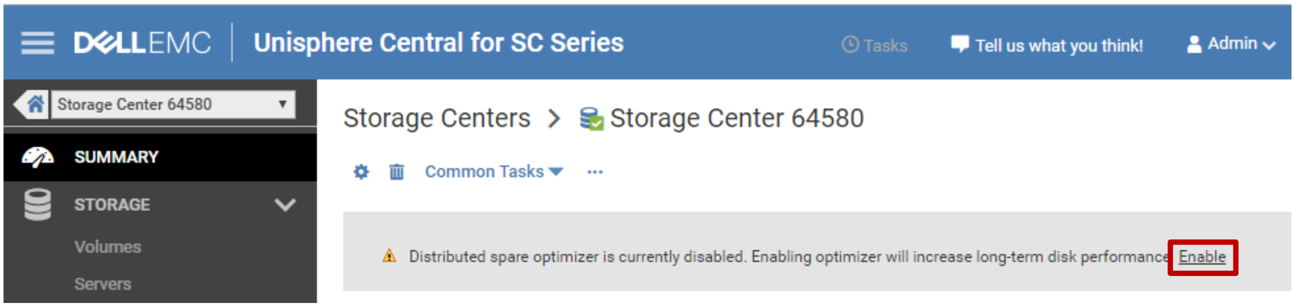
Distributed sparing has significantly increased the speed at which SCOS can return to full data redundancy. However, avoiding data rebuilds altogether is preferable. Hard drive failures can be organized into two categories: the drive has failed or is about to fail and cannot be trusted, or the drive has not failed but has reached the SCOS error recovery threshold and is expected to fail. Drives in the first category are marked as failed and a rebuild occurs.

Drives in the second category are still trusted. Therefore, rather than perform a costly rebuild, a lower-impact copy operation can be performed. When performing this copy operation, data is mirrored from the expected-to-fail drive to spare capacity within the array. While this copying is underway, the drive will continue to accept writes so that data remains in a synchronized state, and may serve reads if the drive will provide them at lower latency then other drives. As each extent finishes mirroring, the now-redundant copy of data on this disk is cleared. Once all extents are mirrored and cleared, the drive is marked as failed and must be replaced, as is typically the case with a failed drive. Should the drive fail during the copy operation, a rebuild operation is initiated for the remaining data on the drive.

3 Enabling distributed spare

The distributed spare feature is a part of the SCOS 7.3 release. A new deployment of an array running SCOS 7.3 will automatically utilize this feature. For existing arrays, enabling distributed sparing is a two-step process.

Step one consists of upgrading the array to SCOS 7.3. This should be completed following standard upgrade best practices in addition to established business guidelines. Step two is to enable the distributed spare optimizer, which becomes available once the upgrade to SCOS 7.3 is completed, as shown in Figure 6.



⚠ Distributed spare optimizer is currently disabled. Enabling optimizer will increase long-term disk performance Enable

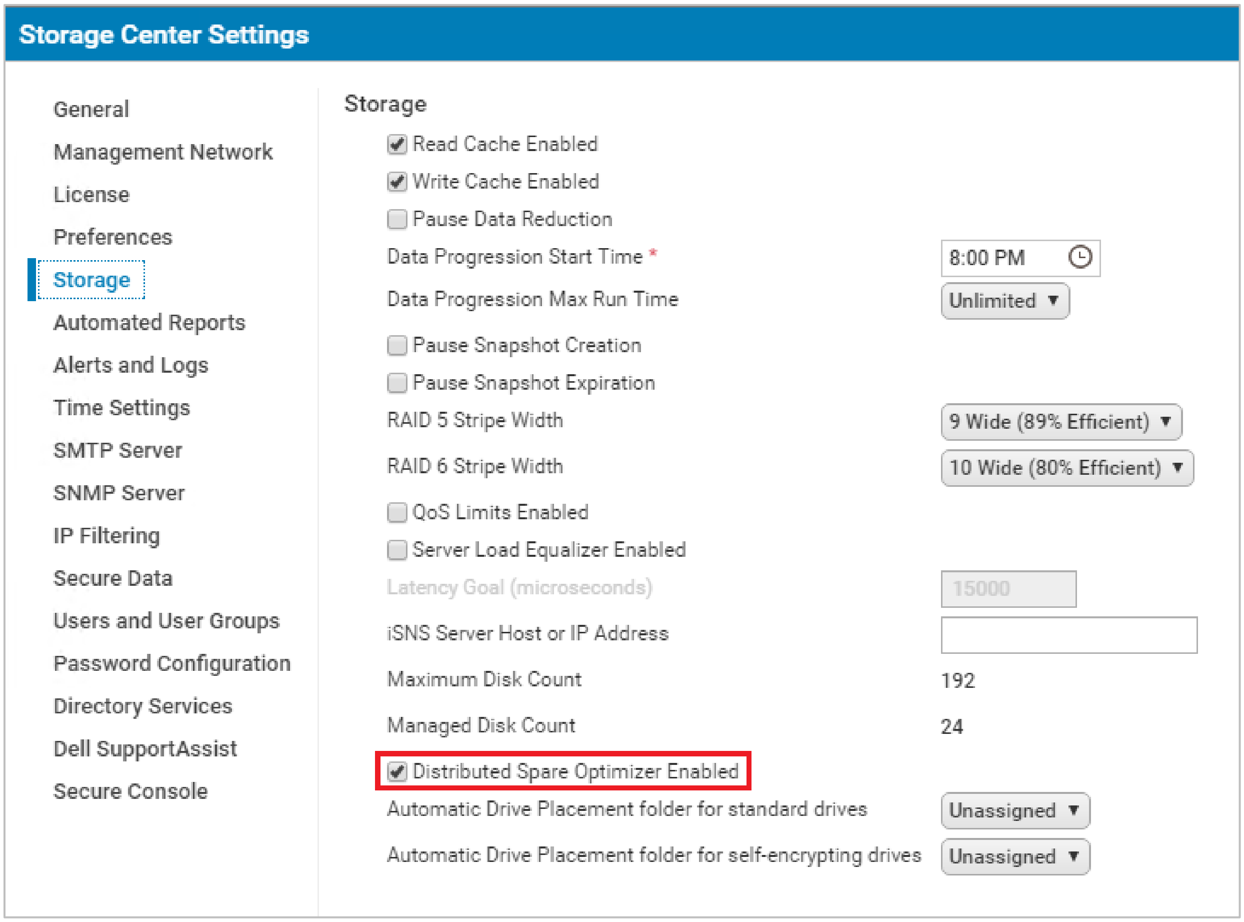
Figure 6 Post 7.3 SCOS upgrade message

Prior to enabling the distributed spare optimizer, SAN administrators should account for the following:

- Once in distributed spare mode the SC Series array cannot be returned to a hot spare RAID configuration.
- Enabling distributed sparing requires redistributing spare capacity from what were previously the hot spare disks to all disks in the array.
- The redistribution of spare capacity is performed as a background operation on the array, which may impact front-end performance under some circumstances. These performance impacts are most frequently seen in arrays operating at close to maximum capabilities.
- For arrays operating at close to maximum capabilities, it is recommended to enable the distributed spare optimizer during a maintenance window or other period of low I/O to minimize impact to front-end I/O and productions systems.

3.1 Pausing the distributed spare optimizer process

In some circumstances, the SAN administrator may prefer to temporarily pause the distributed spare optimizer process. To do this, on the Storage Center Settings screen, uncheck the **Distributed Spare Optimizer Enabled** option as shown in Figure 7. When the optimizer process is paused, any optimizer copy operations in progress are aborted. Any optimizer copy operations that complete while the optimizer was enabled remain in place and are considered progress toward the goal of optimization.



Storage

☑ Read Cache Enabled

☑ Write Cache Enabled

□ Pause Data Reduction

Data Progression Start Time *	8:00 PM
Data Progression Max Run Time	Unlimited ▼

□ Pause Snapshot Creation

□ Pause Snapshot Expiration

RAID 5 Stripe Width	9 Wide (89% Efficient) ▼
RAID 6 Stripe Width	10 Wide (80% Efficient) ▼

□ QoS Limits Enabled

□ Server Load Equalizer Enabled

Latency Goal (microseconds)	15000
iSNS Server Host or IP Address	
Maximum Disk Count	192
Managed Disk Count	24

☑ Distributed Spare Optimizer Enabled

Automatic Drive Placement folder for standard drives	Unassigned ▼
Automatic Drive Placement folder for self-encrypting drives	Unassigned ▼

Figure 7 Uncheck Distributed Spare Optimizer Enabled to pause optimization process

4 Reduction in rebuild times

The primary intent of implementing the distributed spare feature is to increase disk rebuild speeds, thereby reducing the time the array spends on this task. Many factors can impact the speed of a rebuild operation, including:

- Existing production workload
- The read/write ratio of the existing workload
- Number of disks
- Speed of disks
- Amount of RAID mirror data involved
- Amount of RAID parity data involved

Note: Multiple factors and changing variables impact the speed of drive rebuild operations. The results described in this section serve purely as a guideline for the decrease in time required for the rebuild task. While every situation is different, and results will vary, using the distributed spare feature results in less time spent on rebuild operations.

Table 1 Disk rebuild times with distributed sparing

Drive type	Tier 1 size (RAID 10)	Rebuild time
800 GB SSD	Less than 15 SSDs	30–45 minutes
800 GB SSD	More than 15 SSDs	20–30 minutes
7K 8TB HDD	More than 15 HDDs	2–3 hours

The results shown in Table 1 reflect the following improvements over SCOS 7.2 due to implementing the distributed spare feature:

- Approximately a 5x rebuild time performance improvement for RAID 10
- Approximately a 2x rebuild time performance improvement for RAID 5 and 6

5 Additional benefits of distributed spare

The primary benefit of distributed spare is significantly faster RAID rebuild times, but this feature also provides two additional benefits: extended longevity of disks, and increased overall RAID performance.

5.1 Extended longevity of disks

Distributed spare sets aside a portion of each disk as sparing capacity, which results in each disk being under-provisioned. This under-provisioning reduces the workload made to a drive, and provides the hard drive controller additional capacity from which to perform wear-leveling operations. This results in longer lifespans for both spinning HDDs and SSDs.

5.2 Increased overall RAID performance

In a traditional RAID configuration with hot spares, not all disks store user data. The hot spare disks are there to be the target or RAID rebuild operation should a disk fail, and therefore do not contribute to the overall array performance. For example, in a system with 12 disks, of which 1 is a hot spare, only 11 disks contribute to the I/O performance of the system.

With distributed spare, each disk now stores user data. The previous hot spare disk is incorporated into the disk set used for user data, and a small slice of each disk is reserved as the distributed spare capacity. The total number of disks in the array remains the same, as does the total amount of usable capacity, but the total number of disks serving user data I/O increases.

Since each disk now stores user data, I/O performance increases because SCOS takes advantage of all disks on the back end. For example, in a system with 12 disks, of which 1 is a hot spare, that hot spare cannot contribute to user I/O performance since nothing is stored there. In a distributed spare configuration, all 12 disks perform user data I/O. The effective increase can be as much as 8% for random workloads.

A Technical support and resources

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

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

A.1 Related resources

Other documents referenced in this paper include:

- Understanding RAID with Dell SC Series Storage
- The Architectural Advantages of Dell SC Series Automated Tiered Storage