

The Lenovo logo is displayed in white text on a black rectangular background.

DSS-G Declustered RAID Technology and Rebuild Performance

**Provides an overview of the
Spectrum Scale RAID technology**

**Introduces the Spectrum Scale RAID
terminology**

**Explains how to calculate the
volume of critical and non-critical
rebuids**

**Demonstrates the rebuild
performance of the DSS-G
declustered RAID technology**

Michael Hennecke



Abstract

Lenovo Distributed Storage Solution for Spectrum Scale (DSS-G) is a high performance storage solution that combines IBM Spectrum Scale software with Lenovo servers, storage, and networking components. The DSS-G solution is available as hardware building blocks with a selectable number of disk storage enclosures and a variety of disk drive types.

The main distinguishing feature of the DSS-G solution is its sophisticated software RAID technology, which is implemented in the IBM Spectrum Scale RAID layer. This document provides an overview of the Spectrum Scale RAID technology, explains how the RAID theory applies to the various Lenovo DSS-G building blocks, and demonstrates the rebuild performance of the technology under single drive failures and multiple drive failures.

This paper is intended for solution architects and storage administrators who need to understand the technology in order to make informed DSS-G sizing and configuration decisions. The paper will be most useful for technical professionals who have a working knowledge of enterprise storage systems and are familiar with the basic features of IBM Spectrum Scale and DSS-G.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

<http://lenovopress.com>

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Introduction	3
Spectrum Scale RAID overview	3
Understanding Spectrum Scale RAID terminology	9
DSS-G building block layout	16
Determining the volume of critical and non-critical rebuilds	20
Rebuild performance measurements	25
Summary	31
Appendix: Conversion of Decimal and Binary Units	31
Additional resources	32
Author	33
Notices	34
Trademarks	35

Introduction

IBM Spectrum Scale, based on IBM General Parallel File System (GPFS) technology, is a high-performance and highly scalable parallel file system with an extensive suite of enterprise class data management features. Lenovo is a strategic alliance partner of IBM, and combines IBM Spectrum Scale software with Lenovo servers, storage, and networking components into the Lenovo Distributed Storage Solution for Spectrum Scale (DSS-G).

The Lenovo DSS-G solution is available as hardware building blocks with a selectable number of disk storage enclosures and a variety of disk drive types. All industry-standard high-performance interconnects are supported by DSS-G, including Mellanox InfiniBand, Intel OmniPath, and 10/25/40/100 Gb/s Ethernet.

The main distinguishing feature of the DSS-G solution is its sophisticated software RAID technology, which is implemented in the IBM Spectrum Scale RAID layer (also known as GPFS Native RAID or GNR).

This document provides an overview of the Spectrum Scale RAID technology, explains how the GNR theory applies to the various Lenovo DSS-G building blocks, and demonstrates the rebuild performance of the technology under single drive failures and multiple drive failures.

A note about performance data: The performance data presented in this document should not be interpreted as a Lenovo commitment to achieve these numbers, and it is not a guaranteed system property of the respective DSS-G models. Actual performance in customer environments depends on many factors beyond the scope of this document. For performance commitments as part of a Lenovo offer (to business partners or customers), the Lenovo-internal review and approval process needs to be followed.

General IBM Spectrum Scale features and specifications can be found in the IBM Spectrum Scale product documentation. General information about the Lenovo DSS-G building blocks can be found in the Lenovo Press Product Guides for DSS-G. See “Additional resources” on page 32.

Spectrum Scale RAID overview

This section provides a high-level overview of the Spectrum Scale declustered RAID technology and the motivation for its development.

Storage Technology Trends

Figure 1 on page 4 shows the severe predicament that all storage systems for high performance computing environments face: While the computational performance of supercomputers is growing exponentially for decades (as shown by the Top500 graph on the left), the performance metrics of hard disk drives have flattened a long time ago. Only the areal density of disk drives is still increasing, albeit at a much slower pace than before:

Disk Storage Technology Trends

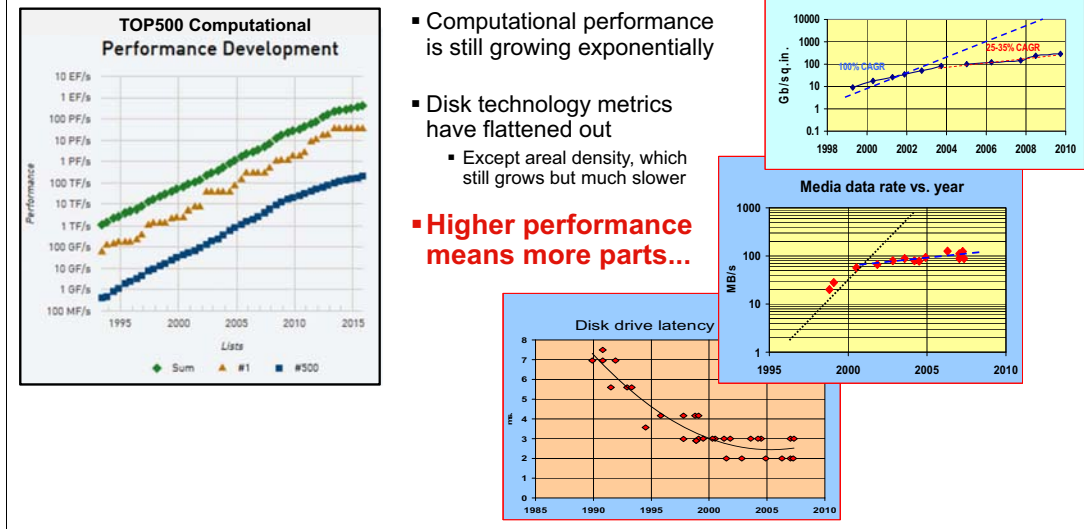


Figure 1 Disk Storage Technology Trends

When individual disk drives do not get faster anymore, then the only option to provide increased storage bandwidth is to add more and more disk drives to the parallel file system.

The continued need for nearline disk drives: Price reductions of SSDs and NVMe storage have made fast storage tiers with SSDs more feasible. But to provide sufficient capacity most HPC storage systems still rely on a very large pool of nearline disk drives, and the staging between multiple storage tiers also needs adequate bandwidth on the nearline tier. This means that the same argument also holds in multi-tiered storage environments.

Parallel file systems like IBM Spectrum Scale, Lustre or BeeGFS provide the software layer to scale out an HPC storage system to thousands of drives. But at larger and larger numbers of disk drives, some hardware properties of the underlying disks become a problem for both the performance and the reliability of the storage subsystem.

These hardware properties include the following:

► Mean Time Between Failure

The first of these properties is the drives' mean time between failure (MTBF). An equivalent metric is the annual failure rate (AFR), which is the inverse of the MTBF for a period of one year. While an MTBF of one million hours seems large, as more and more parts are added the reliability of the solution gets lower and lower. As shown in Figure 2 on page 5, in a system with 10,000 drives there will be a drive failure every four days (on average).

• Unpleasant facts about hard disks #1

MTBF
(Mean Time Between Failures):

- Best specs are ~1 to ~2 million hours
- **Question:**
What does this mean in terms of yearly / monthly failures?
- **Answer:**
Annual Failure Rate (AFR) = 1 / MTBF
MTBF 1.000.000 h → AFR 0.9%
- For 10.000 disks this means:
90 disks/year, or one disk every 4 days

Your Supplier .com

AnyNearlineDisk
2TB and 3TB Capacity-Optimized Enterprise Hard Drive for Bulk-Data Applications

Specification	2TB	3TB	3TB	3TB
Capacity	2TB	3TB	3TB	3TB
MTBF (hours)	1,000,000	1,000,000	1,000,000	1,000,000
Annual Failure Rate (AFR)	0.00009%	0.00009%	0.00009%	0.00009%
MTBF (years)	114	114	114	114
MTBF (months)	9	9	9	9
MTBF (days)	0.37	0.37	0.37	0.37
MTBF (hours)	9,072	9,072	9,072	9,072
MTBF (minutes)	151.2	151.2	151.2	151.2
MTBF (seconds)	9,072,000	9,072,000	9,072,000	9,072,000
MTBF (milliseconds)	9,072,000,000	9,072,000,000	9,072,000,000	9,072,000,000
MTBF (microseconds)	9,072,000,000,000	9,072,000,000,000	9,072,000,000,000	9,072,000,000,000
MTBF (nanoseconds)	9,072,000,000,000,000	9,072,000,000,000,000	9,072,000,000,000,000	9,072,000,000,000,000
MTBF (picoseconds)	9,072,000,000,000,000,000	9,072,000,000,000,000,000	9,072,000,000,000,000,000	9,072,000,000,000,000,000
MTBF (femtoseconds)	9,072,000,000,000,000,000,000	9,072,000,000,000,000,000,000	9,072,000,000,000,000,000,000	9,072,000,000,000,000,000,000
MTBF (attoseconds)	9,072,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

Figure 2 Unpleasant facts about hard drives: MTBF and AFR

This implies that storage systems need to be designed in a way that treats this failure scenario as the norm, rather than as an exceptional state. This is particularly important for bandwidth, as traditional block storage controllers suffer from a severe reduction of sustained bandwidth while RAID rebuilds are in progress. It is a key design goal of Spectrum Scale RAID to enable consistently high bandwidth even in the presence of one or multiple disk drive failures.

▶ Bit Error Rate

The second property of disk drives that becomes problematic at high part counts is the bit error rate (BER). This number indicates how many bits can be read from a disk (on average), before a hard bit error occurs and the affected disk sector cannot be read again. Like MTBF, this is a very large number (for example, one out of 10¹⁵), but it becomes relevant at large scale: With current generation 12 TB drives, a hard (unrecoverable) read error will already happen after the drive is completely read about 10 times, as shown in Figure 3 on page 5.

• Unpleasant facts about hard disks #2

BER
(Bit Error Rate):

- Typical spec is 1 in 10¹⁵ bits read
- **Question:**
How often can you read a 12TB disk before you hit a hard bit error?
- **Answer:**
(10¹⁵ bit) / (12 TeraByte = 12 * 10¹² * 8 bit), so after ~10 complete disk reads

Your Supplier .com

AnyNearlineDisk
2TB and 3TB Capacity-Optimized Enterprise Hard Drive for Bulk-Data Applications

Specification	2TB	3TB	3TB	3TB
Capacity	2TB	3TB	3TB	3TB
MTBF (hours)	1,000,000	1,000,000	1,000,000	1,000,000
Annual Failure Rate (AFR)	0.00009%	0.00009%	0.00009%	0.00009%
MTBF (years)	114	114	114	114
MTBF (months)	9	9	9	9
MTBF (days)	0.37	0.37	0.37	0.37
MTBF (hours)	9,072	9,072	9,072	9,072
MTBF (minutes)	151.2	151.2	151.2	151.2
MTBF (seconds)	9,072,000	9,072,000	9,072,000	9,072,000
MTBF (milliseconds)	9,072,000,000	9,072,000,000	9,072,000,000	9,072,000,000
MTBF (microseconds)	9,072,000,000,000	9,072,000,000,000	9,072,000,000,000	9,072,000,000,000
MTBF (nanoseconds)	9,072,000,000,000,000	9,072,000,000,000,000	9,072,000,000,000,000	9,072,000,000,000,000
MTBF (picoseconds)	9,072,000,000,000,000,000	9,072,000,000,000,000,000	9,072,000,000,000,000,000	9,072,000,000,000,000,000
MTBF (femtoseconds)	9,072,000,000,000,000,000,000	9,072,000,000,000,000,000,000	9,072,000,000,000,000,000,000	9,072,000,000,000,000,000,000
MTBF (attoseconds)	9,072,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000
MTBF (zeptoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000
MTBF (yoctoseconds)	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000	9,072,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

Figure 3 Unpleasant facts about hard drives: BER

Combining the frequency of individual drive failures with the probability to hit an unrecoverable read error highlights an even bigger problem than the reduced performance

while a RAID rebuild is taking place. In a classical “4+P” RAID-5 array, when a drive fails then the four remaining “healthy” drives need to be read from beginning to end in order to reconstruct the data and parity from the failed drive to one of the “hot-spare” drives (or a replacement drive if no hot spares are configured).

However, if a hard bit error occurs on one of the four “healthy” drives during this rebuild, data is lost. Stronger RAID-6 codes like “8+2P” tolerate a second drive failure. But at large drive counts the mean time to data loss (MTTDL) is getting dangerously small even for 2-fault-tolerant RAID-6 codes.

Spectrum Scale RAID addresses this severe threat by a combination of the following:

- ▶ Declustered RAID
- ▶ The ability to distinguish (and prioritize) critical rebuilds from normal rebuilds
- ▶ Enabling 3-fault-tolerant Reed Solomon codes like 8+3P as an optional choice.

These technologies will be explained in the next sections.

Note that the stronger 3-fault-tolerant RAID codes also provide a sufficiently large safety buffer to survive *correlated* failures. The classical calculations of reliability metrics almost always assume *uncorrelated* failures. In reality, there are many cases of correlations between individual component failures, and correlations will cause much smaller MTTDL than what would be expected from “textbook” analysis.

Examples of correlated failures include batches of disk drives which share a common manufacturing problem, environmental factors in the data center that affect a large number of components, and many more.

It should also be noted that failure rates of hard drives exhibit a “bathtub” curve over time: New devices have a higher failure rate that levels off over time, and then towards the end of the device lifetime the failure rate increases significantly. This constitutes both correlated failures and exacerbates the issue of encountering additional failures before the RAID rebuild task can complete.

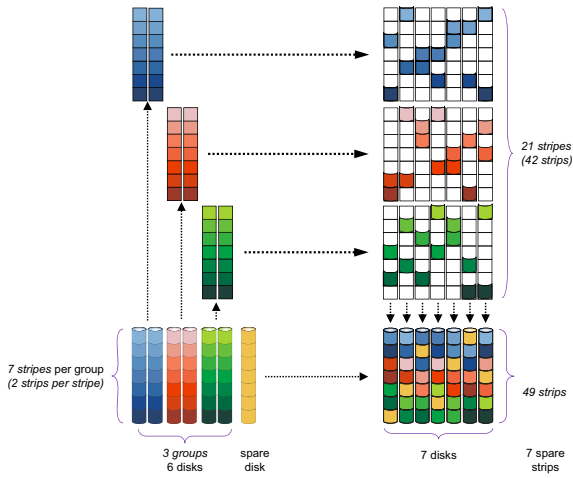
Declustered RAID

One key ingredient to avoid performance degradations during RAID rebuilds is to use a declustered RAID layout instead of traditional RAID arrays.

Figure 4 shows this with an example of RAID-1 arrays (mirroring). On the left, three classical RAID-1 arrays with two drives each are shown, together with one dedicated hot-spare disk. The right side of the figure show how the same three RAID-1 volumes (and the equivalent spare capacity of one drive) would be scattered across all 7 disks drives in a declustered array layout.

How does Declustered RAID work?

- Distributing Data and Parity information as well as Spare Capacity across all disks



Rebuild with Declustered RAID1

- Traditional RAID would have one LUN (logical unit number) fully busy resulting in slow rebuild and high impact overall
- **Declustered RAID** rebuild activity spreads the load across many disks resulting in **faster rebuild** and **less disruption** to user programs
- **Declustered RAID minimizes** critical data exposed to **data loss in case of a second failure.**

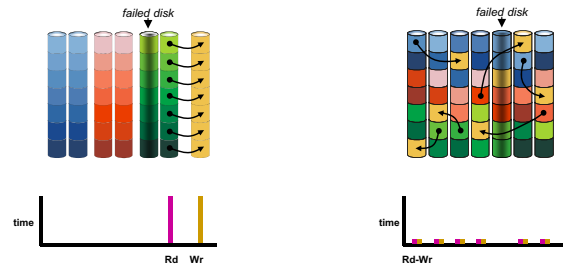


Figure 4 Declustered RAID-1 Example – One Disk Fault

When one drive fails, the behavior of these two layouts is very different:

- ▶ In the classical RAID-1 system, assume that one of the two “green” drives fails. The one surviving green drive now incurs a 100% read load while data is read, and then written to the one spare drive which will have a corresponding 100% write load.

Additional application I/O requests to the green RAID-1 array will be very slow, and in a parallel file system this implies that the whole file system will be slow.

- ▶ In the declustered RAID-1 case, when the same physical drive fails there is an equally large amount of failed RAID-1 stripes, but due to the declustering of the RAID-1 stripes the “surviving” data blocks are now scattered across the 6 remaining healthy drives. The same is true for the spare capacity, which is also scattered across all drives.

This means that the rebuild load can be shared by all of the remaining healthy drives, rather than being concentrated on a single drive being read and another single drive being written.

The advantage of a declustered array obviously increases with the total number of drives: When more disk drives are combined into a declustered array (DA), the relative rebuild load for each individual drive gets lower.

Critical and non-critical rebuild

A second important capability of Spectrum Scale RAID is shown in Figure 5, using the example of two disk faults in a “2+2P” RAID-6 setup.

As in the RAID-1 example (Figure 4 on page 7), the left side shows three classical RAID-6 arrays (and two dedicated hot-spare drives). The right side shows the same three RAID-6 arrays (and the equivalent spare capacity of two drives) scattered across all 14 physical drives in a declustered array.

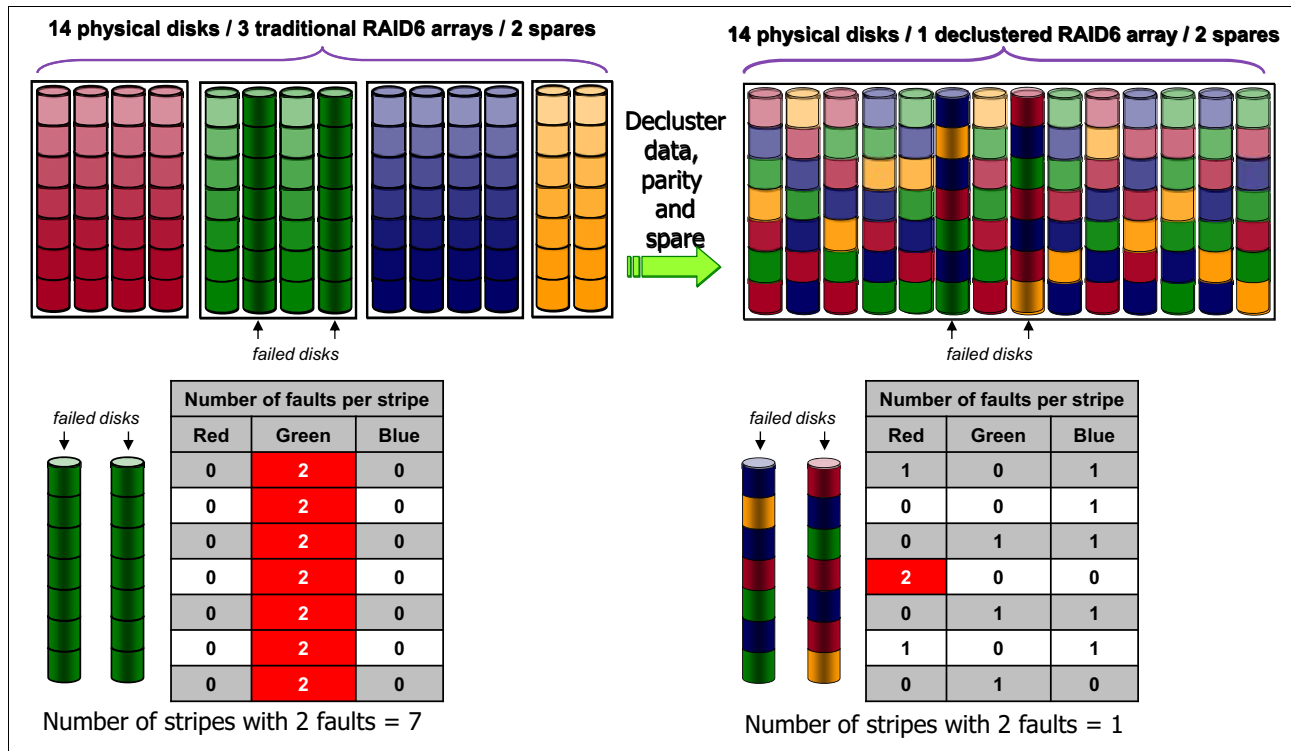


Figure 5 Declustered RAID-6 Example – Two Disk Faults

When the two indicated physical drives fail, the following can be observed:

- ▶ In the classical RAID-6 setup, the “red” and “blue” arrays are completely healthy, while the “green” array has become completely critical.

All the green RAID-6 stripes have lost two strips, and another drive failure in the green array will cause data loss. All green RAID-6 stripes need to be rebuilt to get out of this critical state.

- ▶ In the declustered RAID-6 setup, because all three RAID-6 arrays are scattered across all drives, there are now also failures in the red and blue arrays (not only in the green array).

However, while of all (seven) RAID-6 stripes of the “classical” green array became critical, here only two of the red RAID-6 stripes have become critical (lost two strips). Spectrum Scale RAID can identify those RAID stripes that are in a critical state, and it will rebuild these at a high priority. The RAID stripes that still have at least one strip of parity protection left will only be rebuilt during idle times, when no application I/O is happening. This dramatically speeds up the time that is required to get out of the critical state, and at the same time it further minimizes the impact of the rebuild activity on application I/O performance.

Like in the RAID-1 example, the advantages of the distributed array approach increase with the number of physical disks that are members of the DA. The strength of the RAID code also has a large impact on the fraction of critical RAID stripes. This will be explained in “Determining the volume of critical and non-critical rebuilds” on page 20 for the specific sizes of the Lenovo DSS-G building blocks, and the different protection schemes like 3Way/4WayReplication, 8+2P and 8+3P Reed Solomon coding.

Stronger RAID codes

Most enterprise storage systems provide protection against the failure of a single disk drive (for example 2-way replication or RAID-5) and against the simultaneous failure of two disk drives (for example 3-way replication or RAID-6). One of the original motivations for the design of Spectrum Scale RAID was the fact that these mechanisms do not provide sufficient protection against data loss for the very large HPC storage systems that contain thousands of disk drives within a single parallel file system.

Spectrum Scale RAID therefore provides 4-fault-tolerant protection schemes for both replication (4WayReplicated) and erasure coding (8+3P).

In particular the 8+3P Reed-Solomon coding is very important for storage systems that utilize high-capacity NL-SAS disks. As explained in “Storage Technology Trends” on page 3, its main benefit is the much higher protection against catastrophic data loss, which becomes critical for solutions with thousands of disk drives. 8+3P is also very beneficial for small DSS-G solutions, because it exponentially reduces the amount of data that needs to be rebuilt after multiple disk drive failures. As application I/O performance will be degraded during critical rebuilds, the option to use 8+3P erasure coding allows DSS-G storage administrators to emphasize faster critical rebuilds at the expense of slightly lower usable capacity.

There are several other innovative components within Spectrum Scale RAID, including its end-to-end data integrity features and the disk hospital. Those features have no direct relationship to the rebuild performance of Spectrum Scale’s declustered RAID; they are not discussed in this document.

The IBM Research team has first presented the GNR technology at the USENIX LISA 2011 conference. A replay of that presentation is available on YouTube, and contains much more background information on GNR than what has been covered in this section:

<http://www.youtube.com/watch?v=2g5rx4gP6yU>

Understanding Spectrum Scale RAID terminology

This section explains some important Spectrum Scale RAID terminology that is used in the following sections, as well as the possible states of various Spectrum Scale RAID objects.

A note on Spectrum Scale RAID commands: This document uses the traditional Spectrum Scale RAID commands like `mm1srecoverygroup`, `mm1spdisk` and `mmchpdisk`. Starting with DSS-G version 2.4, Lenovo is also supporting the new `mmvdisk` command set that provides all Spectrum Scale RAID administrative functions under a single command. Please refer to the `mmvdisk` manual page for details. In a building block that is managed with `mmvdisk`, the text of the log entries in the `mmfs.log` can also be slightly different from what is shown in this document.

Spectrum Scale RAID components

The following terms are important to understand the components of a Spectrum Scale RAID solution:

- ▶ **Building Block:** Each Lenovo DSS-G building block with external disk enclosures (JBODs) has two DSS-G servers, and one or more disk enclosures that are each twin-tailed to the two DSS-G servers to provide redundant hardware paths to the physical disks.
- ▶ **Recovery Group and physical disks:** A Spectrum Scale RAID Recovery Group (RG) is the collection of physical disks (pdisks) that are served together by one DSS-G server. There are two RGs per DSS-G building block: One RG contains the pdisks in the left half of the disk enclosure(s), and the other RG contains the pdisks in the right half of the disk enclosure(s).
- ▶ **Primary and Backup RG servers and failover:** The RG is also the unit of failover between the two DSS-G servers: Each RG has a primary RG server and a backup RG server assigned to it, and the active RG server is the one that is currently serving the RG. Failover can be either automatic as the result of a failure, or manually by using the following command:

```
mmchrecoverygroup $RG --active $SERVER
```
- ▶ **Distributed Arrays:** There are multiple Spectrum Scale RAID Distributed Arrays (DA) per RG, which are predefined in the building block's topology and cannot be changed by the administrator. In general there is one DA for each type of storage device in the RG. In a homogeneous DSS-G building block, there are usually three DAs per RG for disk-based building blocks (NVR, SSD, and DA1) and two DAs per RG for SSD-based building blocks (NVR and DA1). In hybrid building blocks, there will also be a DA2 Distributed Array for the second type of user pdisks.
- ▶ **Virtual disks:** Contrary to disk arrays in conventional block storage controllers, a Spectrum Scale RAID Distributed Array does not have a RAID code associated with it. The RAID protection scheme is a property of the virtual disks (vdisks), which correspond to the LUNs of a traditional block storage controller. Multiple vdisks with different size and different RAID codes can be provisioned within the same Distributed Array.
- ▶ Layout of the DSS-G RGs:
 - The NVR and SSD Distributed Arrays are for internal use by the Spectrum Scale RAID code. They do not hold file system data or metadata:
 - logTip is a 2WayReplicated vdisk in the NVR Distributed Array
 - logTipBackup is an Unreplicated vdisk in the SSD Distributed Array.
 - The DA1 (and DA2 for the hybrid models) Distributed Arrays hold the majority of the pdisks; those pdisks are intended for the Spectrum Scale file system data and metadata.
 - The logHome vdisk is 4WayReplicated and is always stored in DA1. Like the logTip, it is used internally by the GNR code.
 - The remaining space in DA1 (and DA2) can be used for vdisks that will hold file system data and metadata. Those vdisks are defined by the administrator.

The number of physical disks (pdisks) in the DA and their capacity (in TB) is important for the critical rebuild discussion below, as well as the strength of the RAID code (2-fault-tolerance or 3-fault tolerance, with replication or erasure coding).

Recovery Group states

A Spectrum Scale RAID Recovery Group (RG) can be online or offline.

The `mm1srecoverygroup` command without any options lists the names of all the RGs that are defined in the cluster, and it should always succeed.

Listing details of a specific RG with the `mm1srecoverygroup $RG` command may fail if the Recovery Group \$RG is currently offline.

There are two typical reasons why an RG may be offline:

- ▶ Either both servers of the DSS-G building block are down
- ▶ The RG is in the middle of an RG failover process

The latter may have been triggered automatically due to a failure of the active RG server, or manually by invoking the `mmchrecoverygroup --active` command.

When the RG is online, the last section of the `mm1srecoverygroup $RG -L` output (see ③ in Example 1) will show the primary and backup RG servers of the RG, as well as the currently active RG server.

Example 1 Listing a Recovery Group's detailed information

```
[root@dssg2201 ~]# mm1srecoverygroup dssg2201 -L --pdisk
```

recovery group	declustered arrays	vdisks	pdisks	current format	current version	allowable format	allowable version	
dssg2201	2	4	26	5.0.0.0		5.0.0.0		①

declustered array	needs service	vdisks	pdisks	spares	replace threshold	free space	scrub duration	background activity task	background activity progress	background activity priority	
NVR	no	1	2	0,0	1	3632 MiB	14 days	scrub	79%	low	②
DA1	no	3	24	2,14	1	2457 GiB	14 days	scrub	71%	low	

pdisk	n. active, total paths	declustered array	free space	state, remarks	
e1s00	2, 4	DA1	133 GiB	ok	
e1s01	2, 4	DA1	133 GiB	ok	
e1s02	2, 4	DA1	133 GiB	ok	
e1s03	2, 4	DA1	133 GiB	ok	
e1s04	2, 4	DA1	133 GiB	ok	
e1s05	2, 4	DA1	133 GiB	ok	
e1s06	2, 4	DA1	133 GiB	ok	
e1s07	2, 4	DA1	133 GiB	ok	
e1s08	2, 4	DA1	133 GiB	ok	
e1s09	2, 4	DA1	133 GiB	ok	
e1s10	2, 4	DA1	133 GiB	ok	
e1s11	2, 4	DA1	133 GiB	ok	
e2s00	2, 4	DA1	133 GiB	ok	
e2s01	2, 4	DA1	133 GiB	ok	
e2s02	2, 4	DA1	133 GiB	ok	
e2s03	2, 4	DA1	133 GiB	ok	
e2s04	2, 4	DA1	133 GiB	ok	

e2s05	2, 4	DA1	133 GiB	ok
e2s06	2, 4	DA1	133 GiB	ok
e2s07	2, 4	DA1	133 GiB	ok
e2s08	2, 4	DA1	133 GiB	ok
e2s09	2, 4	DA1	133 GiB	ok
e2s10	2, 4	DA1	133 GiB	ok
e2s11	2, 4	DA1	133 GiB	ok
n129v001	1, 1	NVR	1816 MiB	ok
n130v001	1, 1	NVR	1816 MiB	ok

vdisk	RAID code	declustered array	vdisk size	block size	checksum granularity	state	remarks	④
-----	-----	-----	-----	-----	-----	-----	-----	
dssg2201_logTip	2WayReplication	NVR	48 MiB	2 MiB	4096	ok	logTip	
dssg2201_logHome	4WayReplication	DA1	40 GiB	2 MiB	4096	ok	log	
dssg2201_d1_1m_3p	8+3p	DA1	3725 GiB	1 MiB	32 KiB	ok		
dssg2201_m1_1m_3w	3WayReplication	DA1	93 GiB	1 MiB	32 KiB	ok		
config data	declustered array	spare space	remarks					⑤
-----	-----	-----	-----					
rebuild space	DA1	18 pdisk						
config data	disk group fault tolerance		remarks					⑥
-----	-----		-----					
rg descriptor	4 pdisk		limiting fault tolerance					
system index	4 pdisk		limited by rg descriptor					
vdisk	disk group fault tolerance		remarks					⑦
-----	-----		-----					
dssg2201_logTip	1 pdisk							
dssg2201_logHome	3 pdisk		limited by rg descriptor					
dssg2201_d1_1m_3p	3 pdisk							
dssg2201_m1_1m_3w	2 pdisk		limited by rg descriptor					
active recovery group server			servers					⑧
-----			-----					
dssg2201			dssg2201,dssg2202					

When the **-Y** option is used to generate colon-separated output, this information can be found in the “server” line, as shown in Example 2.

Example 2 Listing an RG's primary, backup, and active servers in CSV format

```
[root@dssg2201 ~]# mm1srecoverygroup dssg2201 -Y | grep ":server:"
mm1srecoverygroup:server:HEADER:version:reserved:reserved:ActiveRecoveryGroupServer:Servers:
mm1srecoverygroup:server:0:1:::dssg2201:dssg2201,dssg2202:
```

```
[root@dssg2201 ~]# mm1srecoverygroup dssg2202 -Y | grep ":server:"
mm1srecoverygroup:server:HEADER:version:reserved:reserved:ActiveRecoveryGroupServer:Servers:
mm1srecoverygroup:server:0:1:::dssg2202:dssg2202,dssg2201:
```

Pdisk states

The physical disks (pdisks) in a Spectrum Scale RAID Recovery Group can assume one of many different states, which can be viewed in the State field of the pdisk section of the `mm1srecoverygroup $RG -L` output (see ❸ in Example 1 on page 11), or in the pdisk lines of the `mm1srecoverygroup -Y` command as shown in Example 3.

Example 3 Listing an RG's pdisk status in CSV format

```
[root@dssg2201 ~]# mm1srecoverygroup dssg2201 -Y | grep ":pdisk:"
mm1srecoverygroup:pdisk:HEADER:version:reserved:reserved:Pdisk:Paths:DeclusteredArray:Capacity:FreeSpace:State>UserCondition:TotalPaths:
mm1srecoverygroup:pdisk:0:1:::e1s00:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s01:2:DA1:399968829440:143612968960:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s02:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s03:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s04:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s05:2:DA1:399968829440:143612968960:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s06:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s07:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s08:2:DA1:399968829440:143612968960:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s09:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s10:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e1s11:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s00:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s01:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s02:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s03:2:DA1:399968829440:143612968960:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s04:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s05:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s06:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s07:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s08:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s09:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s10:2:DA1:399968829440:143076098048:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::e2s11:2:DA1:399968829440:143344533504:ok:normal:4:
mm1srecoverygroup:pdisk:0:1:::n129v001:1:NVR:2088763392:1904214016:ok:normal:1:
mm1srecoverygroup:pdisk:0:1:::n130v001:1:NVR:2088763392:1904214016:ok:normal:1:
```

The pdisk states are summarized in Table 1:

Table 1 Pdisk states

Pdisk state	Meaning
OK	The disk is available and functioning normally. This is the default state in a healthy system
dead	The disk completely failed.
simulatedDead	The disk is being treated as if it were dead for error injection and testing. Refer to the description of the <code>--simulate-dead</code> parameter of the <code>mmchpdisk</code> command in the Spectrum Scale Command Reference or manual page.
missing	The disk hospital determined that the system cannot connect to the drive.
readonly	The disk has failed; it can still be read but not written


Pdisk state	Meaning
failing	The disk needs to be drained and replaced due to a SMART trip or high uncorrectable error rate.
simulatedFailing	The disk is being treated as if it were failing for error injection (see <code>mmchpdisk --simulate-failing</code> in the Spectrum Scale Command Reference or manual page).
slow	The disk needs to be drained and replaced due to poor performance.
diagnosing	The disk hospital is checking the disk after an error.
PTOW	The disk is temporarily unavailable because of a pending timed-out write.
suspended	The disk is temporarily offline for service (see <code>mmchpdisk</code> and <code>mmchcarrier</code>).
serviceDrain	The disk is being drained of data for service (see <code>mmchpdisk --begin-service-drain</code>).
draining	The data is being drained from the disk and moved to distributed spare space on other disks.
deleting	The disk is being deleted from the system through the <code>mmde1pdisk</code> , <code>mmaddpdisk --replace</code> , or <code>mmchcarrier</code> command.
drained	All of the data was successfully drained from the disk and the disk is replaceable, but the replace threshold was not met.
undrainable	As much of the data as possible was drained from the disk and moved to distributed spare space.
replace	The disk is ready for replacement.

To study the rebuild scenarios in this document, pdisk states will be manually changed back and forth between the OK and SimulatedDead states by using the `mmchpdisk` command.

Distributed Array states

Use either of the following commands to view the state of a Distributed Array (DA):

```
mm1srecoverygroup $RG -L
mm1srecoverygroup $RG -Y
```

Check the Task column of the Declustered Array section (see  in Example 1 on page 11), or in the daSummary lines of the colon-separated output when using the `-Y` option as shown in Example 4.

Example 4 Listing an RG's DA status in CSV format

```
[root@dssg2201 ~]# mm1srecoverygroup dssg2201 -Y | grep "daSummary:"
mm1srecoverygroup:daSummary:HEADER:version:reserved:reserved:DeclusteredArray:NeedsService:Vdisks:Pdisks:Spare:ReplaceThreshold:Capacity:FreeSpace:ScrubDurationInDays:BgTask:BgTaskPctComplete:BgTaskPriority:vcdSpares:
mm1srecoverygroup:daSummary:0:1:::NVR:no:1:2:0:1:4177526784:3808428032:14:scrub:79:low:0:
mm1srecoverygroup:daSummary:0:1:::DA1:no:3:24:2:1:9599251906560:2638452097024:14:scrub:71:low:14:
```

The Distributed Array states are summarized in Table 2.

Table 2 Distributed Array states

Distributed Array state	Meaning
inactive	There are no vdisks defined on the DA, or the declustered array is not currently available. This state should not occur in production systems, except at the very beginning of the configuration of the building block (when no vdisks have been defined yet) or briefly during a recovery group takeover.
scrub	The vdisks are undergoing routine data integrity maintenance. This is the default state of a DA when there are no errors. There is always a low priority background "scrub" task scheduled on GNR Distributed Arrays that hold vdisks.
rebuild-critical	vdisk tracks with no remaining redundancy are being rebuilt. The DA is in this state when any of the vdisks in the DA are in critical state, for example when two pdisks have failed and there is an 8+2P (or a 3WayReplicated) vdisk in the DA, or when three pdisks have failed and there is an 8+3P (or a 4WayReplicated) vdisk in the DA. As soon as the critical stripes of the critical vdisk(s) have been rebuilt, the DA state changes to one of the following states.
rebuild-1r	vdisk tracks with one remaining redundancy are being rebuilt. This could be in an 8+2P (or a 3WayReplicated) vdisk with one pdisk failure, or an 8+3P (or a 4Way-Replicated) vdisk with two pdisk failures.
rebuild-2r	vdisk tracks with two remaining redundancies are being rebuilt. This state occurs for an 8+3P (or a 4WayReplicated) vdisk when one pdisk has failed.

Vdisk states

To view the state of a virtual disk (vdisk), use either of the following commands:

```
mm1srecoverygroup $RG -L
mm1srecoverygroup $RG -Y
```

The vdisk states (which correspond to the above states of the Distributed Array) are reported in the state column of the vdisk section of the `-L` command output (see ❹ in Example 1 on page 11), or in the vdisk lines of the colon-separated output of the `mm1srecoverygroup $RG -Y` command as shown in Example 5.

Example 5 Listing an RG's vdisk status in CSV format

```
[root@dssg2201 ~]# mm1srecoverygroup dssg2201 -Y | grep ":vdisk:"
mm1srecoverygroup:vdisk:HEADER:version:reserved:vdisk:RaidCode:DecIusteredArray:VdiskSizeInGiB:Rem
arks:trackSize:checksumGranularity:state:
mm1srecoverygroup:vdisk:0:1:::dssg2201_logTip:2WayRepIication:NVR:50331648:logTip:2097152:4096:ok:
mm1srecoverygroup:vdisk:0:1:::dssg2201_logHome:4WayRepIication:DA1:42949672960:log:2097152:4096:ok:
mm1srecoverygroup:vdisk:0:1:::dssg2201_d1_1m_3p:8+3p:DA1:4000652984320::1048576:32768:ok:
mm1srecoverygroup:vdisk:0:1:::dssg2201_m1_1m_3w:3WayRepIication:DA1:100169416704::1048576:32768:ok:
```

The vdisk states are summarized in Table 3.

Table 3 vdisk states

vdisk state	Meaning
OK	Means that the vdisk is functioning normally. It is being scrubbed or is waiting to be scrubbed. Only one vdisk in a DA is scrubbed at a time.
1/2-degraded	Means that the vdisk is currently running in degraded mode and that tracks with one fault in a vdisk with two redundancies are being rebuilt.
1/3-degraded	Means that the vdisk is currently running in degraded mode and that tracks with one fault in a vdisk with three redundancies are being rebuilt.
2/3-degraded	Means that the vdisk is currently running in degraded mode and that tracks with two faults in a vdisk with three redundancies are being rebuilt.
critical	Means that the vdisk is currently running in degraded mode and cannot tolerate another pdisk loss. The tracks with no remaining redundancy are currently being rebuilt with high priority.
(need spare)	Means that a rebuild has started, but insufficient spare capacity is available in the DA to complete the rebuild. Rebuild will resume when more spare capacity becomes available. This state applies to both m/n-degraded states and critical states.

DSS-G building block layout

This section presents the layout of the different types of Lenovo DSS-G building blocks. This is useful background information to understand the calculations of the volume of data that needs to be rebuilt after single or multiple drive failures.

DSS-G large form-factor NL-SAS models

The DSS-G2x0 models use two servers and one to six large form-factor D3284 enclosures with NL-SAS drives. Figure 6 shows the G210 and G220 models - in the bottom D3284 enclosure, there are two SSD drives (one in each of the two RGs) that are used for the logTipBackup. This means that in encl1, only 41 NL-SAS drives per RG are available for the user data in DA1. In the G220 (and bigger models), the additional D3284 enclosures do not include these SSDs and all 42 drives per RG are available for user data in DA1.

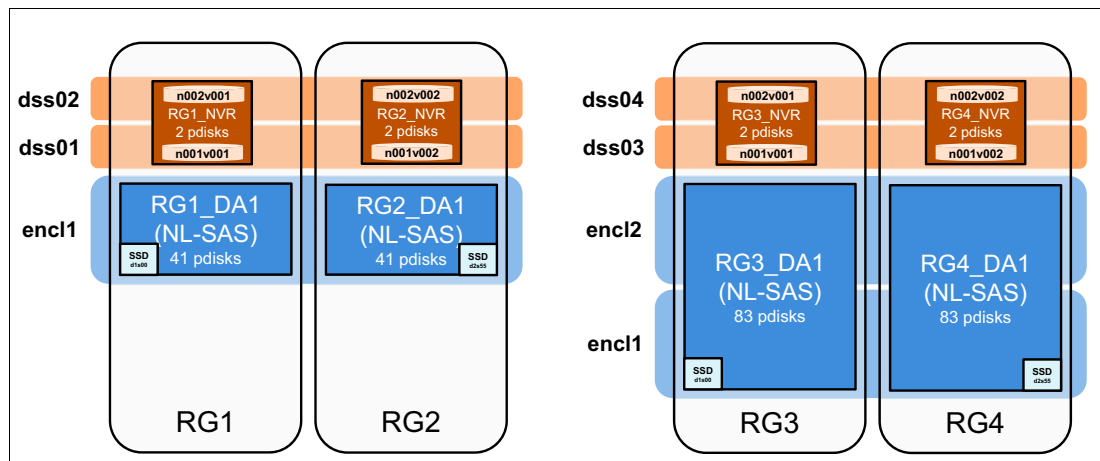


Figure 6 DSS-G210 and DSS-G220 NL-SAS Models

DSS-G small form-factor SSD models

The typical usage of the DSS-G20y models with small form-factor D1224 disk enclosures is to populate them with SSDs. In the SSD case, one to four D1224 enclosures with SSDs are supported. Figure 7 on page 17 shows the G201 and G202 models.

Note that there is no dedicated SSD Distributed Array: Since the main DA1 is already hosted on SSDs, no logTipBackup vdisk is used. In case of a failure of one of the two pdisks of the logTip vdisks (in the “NVR” DA), the code falls back to using just the logHome vdisk that is stored in the SSD-backed DA1.

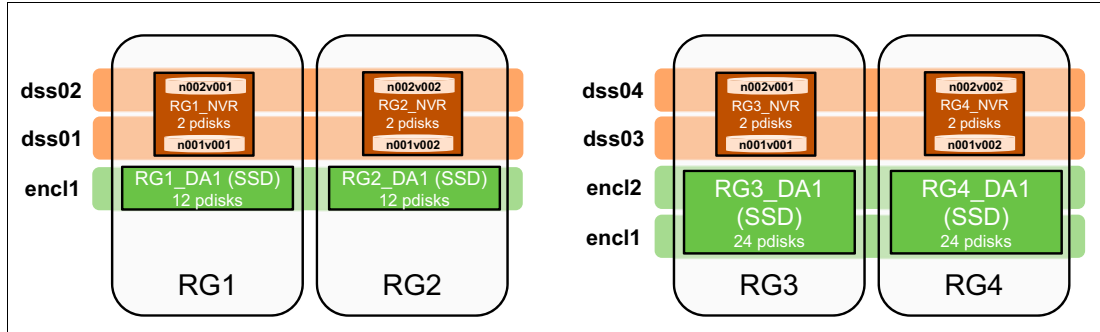


Figure 7 DSS-G201 and DSS-G202 SSD models

DSS-G small form-factor HDD models

Figure 8 shows the G20y models where two, four or six of the small form-factor D1224 enclosures are populated with spinning disks (HDDs). In this case the bottom enclosure will hold two SSD pdisks for the SSD DAs (for the logTipBackup vdisk), just like the NL-SAS building blocks.

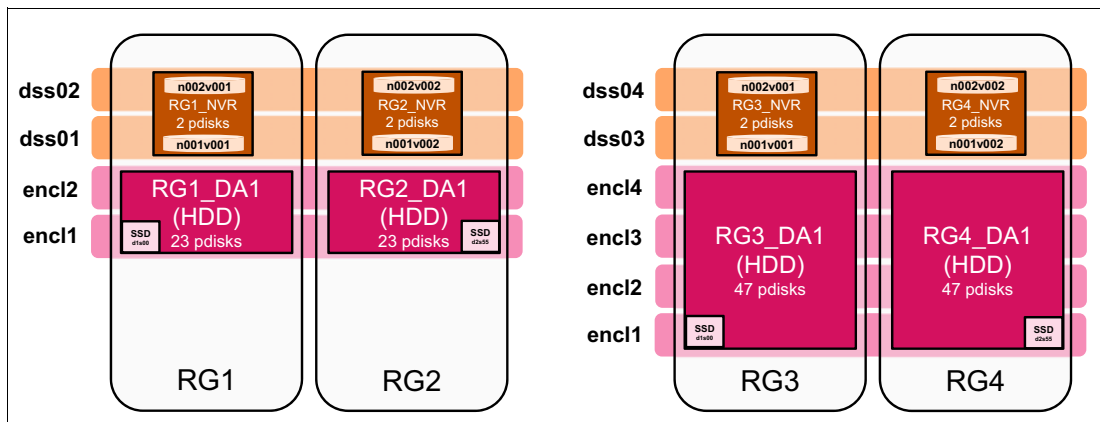


Figure 8 DSS-G202 and DSS-G204 HDD models

DSS-G hybrid models

The hybrid DSS-G2xy models, Figure 9, are identical with a DSS-G2x0 model that uses large form-factor D3284 enclosures with NL-SAS drives, and a number of “y” small form-factor D1224 disk enclosures with SSD drives on top.

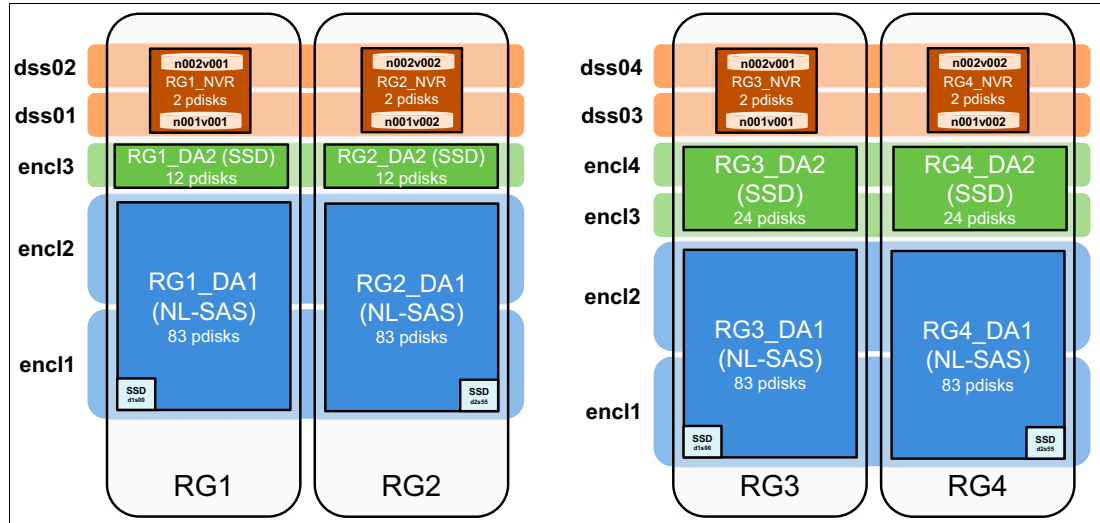


Figure 9 DSS-G221 and DSS-G222 hybrid models

The two different types of disk enclosures are configured as different Distributed Arrays (DA1 and DA2). This implies that for the purpose of analyzing the resiliency and rebuild performance of a hybrid building block, it can be treated as two individual building blocks DSS-G2x0 and DSS-G20y. The remainder of this paper will thus not explicitly cover the hybrid DSS-G models.

Enclosure loss protection and drawer loss protection

This paper mainly focuses on individual pdisk failures and the way Spectrum Scale RAID handles those failures, because this is by far the most common failure scenario. Another related aspect of storage system reliability and availability is the question how resilient it is against the catastrophic loss of a complete storage enclosure (commonly known as *enclosure loss protection*).

Lenovo DSS-G is designed with no single points-of-failure in the storage backend:

- ▶ Redundant direct SAS connections from the two DSS-G servers to the two ESM modules in each JBOD enclosure
- ▶ Redundant SAS paths between the ESMs and the pdisks within an enclosure
- ▶ Redundant fans and power supplies

This means that there should never be a complete failure of a full JBOD enclosure. Nevertheless, it is interesting to understand the level of resiliency that the Spectrum Scale RAID implementation offers in the extremely unlikely event of a complete enclosure loss.

The key concept to understand this scenario is that of a Spectrum Scale RAID failure domain. The Spectrum Scale RAID code is fully aware of the SAS topology of the storage backend, both in terms of the cable topology between the servers and the storage enclosures (as reported by topsummary) and also in terms of the internal SAS topology within the storage enclosures.

Small storage enclosures like the Lenovo D1224 are “simple” in the sense that the two ESM modules of the enclosure directly connect to all 24 pdisks within the enclosure, without other active components. Large storage enclosures like the Lenovo D3284 have a more complex internal structure, with multiple physical drawers and additional active SAS components (specifically, additional SAS switches in the left and right sideplanes of the two drawers).

When determining the pdisks on which to store the strips of a RAID stripe (like the 10 strips of an 8+2P stripe), Spectrum Scale RAID takes this SAS topology into consideration in order to distribute the strips as widely as possible across the building block. It does so by obeying three levels of failure domains:

- ▶ Individual physical disks (pdisks)
- ▶ Individual drawers within a large storage enclosure
- ▶ Individual storage enclosures within a DSS-G building block

The importance of the pdisk level is obvious: RAID protection would be useless if all strips of a RAID stripe would be stored on the same physical disk. If that pdisk fails, the RAID stripe would become inaccessible and data would be lost. All DSS-G building blocks have a large enough number of pdisks to ensure that each strip of a RAID stripe within a vdisk is stored on a different pdisk, thus maximizing the resiliency of the storage solution against individual pdisk failures.

To maximize the resiliency against a catastrophic enclosure loss, Spectrum Scale RAID will pick the pdisks that hold the strips of a RAID stripe from as many different storage enclosures as possible. The ability to do this depends on the size of the DSS-G building block: In the case of a DSS-G210 or DSS-G201 model there is only a single storage enclosure, and there is obviously no way to protect against the loss of that single enclosure. However, for the bigger building blocks with multiple enclosures it is possible to spread out the RAID strips across the available enclosures.

For the D3284 enclosures, the drawer level is also relevant because it is the drawer that holds the individual pdisks, not the enclosure. Spectrum Scale RAID also respects the drawer failure domains within an enclosure when selecting pdisks for the strips of a RAID strips. When two pdisks are selected within a single D3284 enclosure, they will not be selected from the same drawer but will be spread across both drawers.

The following two tables show the resulting fault tolerance of the DSS-G large form-factor NL-SAS building blocks against a complete enclosure or drawer failure: When the fault tolerance of the RAID code is bigger or equal to the average number of strips per enclosure or drawer (rounded up to the next integer), then the building block can survive the complete failure of an enclosure or drawer.

Table 4 Enclosure/drawer loss tolerance for NL-SAS Models with 8+2P Reed-Solomon.

DSS-G NL-SAS Model	Enclosures per BB	Drawers per BB	average 8+2P strips per enclosure	enclosure loss tolerance	average 8+2P strips per drawer	drawer loss tolerance
DSS-G210	1	2	10/1 = 10	0	10/2 = 5	0
DSS-G220	2	4	10/2 = 5	0	10/4 = 2.5	0
DSS-G230	3	6	10/3 = 3.33	0	10/6 = 1.67	1
DSS-G240	4	8	10/4 = 2.5	0	10/8 = 1.25	1
DSS-G250	5	10	10/5 = 2	1	10/10 = 1	2
DSS-G260	6	12	10/6 = 1.67	1	10/12 = 0.83	2

DSS-G NL-SAS Model	Enclosures per BB	Drawers per BB	average 8+2P strips per enclosure	enclosure loss tolerance	average 8+2P strips per drawer	drawer loss tolerance
DSS-G270	7	14	10/7 = 1.43	1	10/14 = 0.71	2
DSS-G280	8	16	10/8 = 1.25	1	10/16 = 0.63	3

Table 5 Enclosure/drawer loss tolerance for NL-SAS models with 8+3P Reed-Solomon

DSS-G NL-SAS Model:	Enclosures per BB	Drawers per BB	average 8+3P strips per enclosure	enclosure loss tolerance	average 8+3P strips per drawer	drawer loss tolerance
DSS-G210	1	2	11/1 = 11	0	11/2 = 5.5	0
DSS-G220	2	4	11/2 = 5.5	0	11/4 = 2.75	1
DSS-G230	3	6	11/3 = 3.67	0	11/6 = 1.83	1
DSS-G240	4	8	11/4 = 2.75	1	11/8 = 1.38	2
DSS-G250	5	10	11/5 = 2.2	1	11/10 = 1.1	2
DSS-G260	6	12	11/6 = 1.83	1	11/12 = 0.92	3
DSS-G270	7	14	11/7 = 1.57	1	11/14 = 0.79	3
DSS-G280	8	16	11/8 = 1.38	2	11/16 = 0.69	4

A word of caution regarding enclosure and drawer loss protection: It is true that, for example, a DSS-G250 with 8+2P erasure coding can tolerate the loss of a complete enclosure, but in this example 100% of the RAID stripes will become critical. It would take a very long time to perform the critical rebuild of all the user data, and any additional pdisk failure during that phase will result in data loss. In addition, in order to enable the critical rebuild of a full enclosure it is necessary to permanently allocate 20% of the building block's capacity as spare space.

If a DSS-G building block is operated in an environment that needs to guard against such extremely unlikely events like a complete enclosure failure, it may be better to use other mechanisms like Spectrum Scale synchronous software replication across two completely separate building blocks or asynchronous DR configurations. This will generally provide much better protection (with or without enclosure loss protection), and it avoids the risk of data loss during a prolonged critical rebuild phase in one of the building blocks by always having a second copy of the data available.

Determining the volume of critical and non-critical rebuilds

This section provides reference information for the amount of user data that needs to be rebuilt when one or multiple physical disks (pdisks) in a Distributed Array (DA) have failed. The general mechanisms of the declustered RAID rebuild have been discussed in "Declassified RAID" on page 6. In this section, we provide the numerical quantities for the specific Lenovo DSS-G building blocks.

The following subsections each contain four tables that show the percentage of *normally degraded* and *critically degraded* stripes for

- ▶ 8+2P Reed-Solomon codes
- ▶ 8+3P Reed-Solomon codes
- ▶ 3WayReplication
- ▶ 4WayReplication

The rows of the tables show the different sizes of the DSS-G models, where an increasing number of enclosures (and corresponding size of the DA) leads to a lower percentage of degraded stripes. The columns show these percentages for the cases where one, two or three pdisks are failing simultaneously. The 2-fault-tolerant codes obviously cannot tolerate a third pdisk failure without data loss, so no information is given for those cases.

Rebuild volume for DSS-G large form-factor NL-SAS models

DSS-G "NL-SAS" Model:	Size [U]	Drives per BB	Drives per DA1	8+2P with <u>one</u> failed drive: % of <u>1/2-deg</u> stripes	8+2P with <u>two</u> failed drives: % of <u>critical</u> stripes	8+2P with <u>three</u> failed drives: % of <u>data loss</u>
DSS-G 2 1 0	9	82	41	(10/41) = 24.39%	(10/41) * (9/40) = 5.488%	
DSS-G 2 2 0	14	166	83	(10/83) = 12.05%	(10/83) * (9/82) = 1.322%	
DSS-G 2 3 0	19	250	125	(10/125) = 8.00%	(10/125) * (9/124) = 0.581%	
DSS-G 2 4 0	24	334	167	(10/167) = 5.99%	(10/167) * (9/166) = 0.325%	
DSS-G 2 5 0	29	418	209	(10/209) = 4.78%	(10/209) * (9/208) = 0.207%	
DSS-G 2 6 0	34	502	251	(10/251) = 3.98%	(10/251) * (9/250) = 0.143%	
DSS-G 2 7 0	39	586	293	(10/293) = 3.41%	(10/293) * (9/292) = 0.105%	
DSS-G 2 8 0	44	670	335	(10/335) = 2.99%	(10/335) * (9/334) = 0.080%	

Table 6 Rebuild volume for NL-SAS models with 8+2P Reed-Solomon

DSS-G "NL-SAS" Model:	Size [U]	Drives per BB	Drives per DA1	8+3P with <u>one</u> failed drive: % of <u>1/3-deg</u> stripes	8+3P with <u>two</u> failed drives: % of <u>2/3-deg</u> stripes	8+3P with <u>three</u> failed drives: % of <u>critical</u> stripes
DSS-G 2 1 0	9	82	41	(11/41) = 26.83%	(11/41) * (10/40) = 6.707%	(11/41) * (10/40) * (9/39) = 1.548%
DSS-G 2 2 0	14	166	83	(11/83) = 13.25%	(11/83) * (10/82) = 1.616%	(11/83) * (10/82) * (9/81) = 0.180%
DSS-G 2 3 0	19	250	125	(11/125) = 8.80%	(11/125) * (10/124) = 0.710%	(11/125) * (10/124) * (9/123) = 0.052%
DSS-G 2 4 0	24	334	167	(11/167) = 6.59%	(11/167) * (10/166) = 0.397%	(11/167) * (10/166) * (9/165) = 0.022%
DSS-G 2 5 0	29	418	209	(11/209) = 5.26%	(11/209) * (10/208) = 0.253%	(11/209) * (10/208) * (9/207) = 0.011%
DSS-G 2 6 0	34	502	251	(11/251) = 4.38%	(11/251) * (10/250) = 0.175%	(11/251) * (10/250) * (9/249) = 0.006%
DSS-G 2 7 0	39	586	293	(11/293) = 3.75%	(11/293) * (10/292) = 0.129%	(11/293) * (10/292) * (9/291) = 0.004%
DSS-G 2 8 0	44	670	335	(11/335) = 3.28%	(11/335) * (10/334) = 0.098%	(11/335) * (10/334) * (9/333) = 0.003%

Table 7 Rebuild volume for NL-SAS models with 8+3P Reed-Solomon

DSS-G "NL-SAS" Model:	Size [U]	Drives per BB	Drives per DA1	3WayReplication with <u>one</u> failed drive: % of <u>1/2-deg</u> stripes	3WayReplication with <u>two</u> failed drives: % of <u>critical</u> stripes	3WayReplication with <u>three</u> failed drives: <u>data loss</u>
DSS-G 2 1 0	9	82	41	(3/41) = 7.32%	(3/41) * (2/40) = 0.3659%	
DSS-G 2 2 0	14	166	83	(3/83) = 3.61%	(3/83) * (2/82) = 0.0882%	
DSS-G 2 3 0	19	250	125	(3/125) = 2.40%	(3/125) * (2/124) = 0.0387%	
DSS-G 2 4 0	24	334	167	(3/167) = 1.80%	(3/167) * (2/166) = 0.0216%	
DSS-G 2 5 0	29	418	209	(3/209) = 1.44%	(3/209) * (2/208) = 0.0138%	
DSS-G 2 6 0	34	502	251	(3/251) = 1.20%	(3/251) * (2/250) = 0.0096%	
DSS-G 2 7 0	39	586	293	(3/293) = 1.02%	(3/293) * (2/292) = 0.0070%	
DSS-G 2 8 0	44	670	335	(3/335) = 0.90%	(3/335) * (2/334) = 0.0054%	

Table 8 Rebuild volume for NL-SAS models with 3WayReplication

DSS-G "NL-SAS" Model:	Size [U]	Drives per BB	Drives per DA1	4WayReplication with <u>one</u> failed drive: % of <u>1/3-deg</u> stripes	4WayReplication with <u>two</u> failed drives: % of <u>2/3-deg</u> stripes	4WayReplication with <u>three</u> failed drives: % of <u>critical</u> stripes
DSS-G 2 1 0	9	82	41	(4/41) = 9.76%	(4/41) * (3/40) = 0.7317%	(4/41) * (3/40) * (2/39) = 0.0375%
DSS-G 2 2 0	14	166	83	(4/83) = 4.82%	(4/83) * (3/82) = 0.1763%	(4/83) * (3/82) * (2/81) = 0.0044%
DSS-G 2 3 0	19	250	125	(4/125) = 3.20%	(4/125) * (3/124) = 0.0774%	(4/125) * (3/124) * (2/123) = 0.0013%
DSS-G 2 4 0	24	334	167	(4/167) = 2.40%	(4/167) * (3/166) = 0.0433%	(4/167) * (3/166) * (2/165) = 0.0005%
DSS-G 2 5 0	29	418	209	(4/209) = 1.91%	(4/209) * (3/208) = 0.0276%	(4/209) * (3/208) * (2/207) = 0.0003%
DSS-G 2 6 0	34	502	251	(4/251) = 1.59%	(4/251) * (3/250) = 0.0191%	(4/251) * (3/250) * (2/249) = 0.0002%
DSS-G 2 7 0	39	586	293	(4/293) = 1.37%	(4/293) * (3/292) = 0.0140%	(4/293) * (3/292) * (2/291) = 0.0001%
DSS-G 2 8 0	44	670	335	(4/335) = 1.19%	(4/335) * (3/334) = 0.0107%	(4/335) * (3/334) * (2/333) = 0.0001%

Table 9 Rebuild volume for NL-SAS models with 4WayReplication

Rebuild volume for DSS-G small form-factor SSD models

DSS-G "SSD" Model:	Size [U]	Drives per BB	Drives per DA1	8+2P with <u>one</u> failed drive: % of <u>1/2-deg</u> stripes	8+2P with <u>two</u> failed drives: % of <u>critical</u> stripes	8+2P with <u>three</u> failed drives: <u>data loss</u>
DSS-G 2 0 1	6	24	12	(10/12) = 83.33%	(10/12) * (9/11) = 68.18%	
DSS-G 2 0 2	8	48	24	(10/24) = 41.67%	(10/24) * (9/23) = 16.30%	
DSS-G 2 0 3	10	72	36	(10/36) = 27.78%	(10/36) * (9/35) = 7.14%	
DSS-G 2 0 4	12	96	48	(10/48) = 20.83%	(10/48) * (9/47) = 3.99%	

Table 10 Rebuild Volume for SSD models with 8+2P Reed-Solomon

DSS-G "SSD" Model:	Size [U]	Drives per BB	Drives per DA1	8+3P with <u>one</u> failed drive: % of <u>1/3-deg</u> stripes	8+3P with <u>two</u> failed drives: % of <u>2/3-deg</u> stripes	8+3P with <u>three</u> failed drives: % of <u>critical</u> stripes
DSS-G 2 0 1	6	24	12	(11/12) = 91.67%	(11/12) * (10/11) = 83.33%	(11/12) * (10/11) * (9/10) = 75.000%
DSS-G 2 0 2	8	48	24	(11/24) = 45.83%	(11/24) * (10/23) = 19.93%	(11/24) * (10/23) * (9/22) = 8.152%
DSS-G 2 0 3	10	72	36	(11/36) = 30.56%	(11/36) * (10/35) = 8.73%	(11/36) * (10/35) * (9/34) = 2.311%
DSS-G 2 0 4	12	96	48	(11/48) = 22.92%	(11/48) * (10/47) = 4.88%	(11/48) * (10/47) * (9/46) = 0.954%

Table 11 Rebuild Volume for SSD models with 8+3P Reed-Solomon

DSS-G "SSD" Model:	Size [U]	Drives per BB	Drives per DA1	3WayReplication with <u>one</u> failed drive: % of <u>1/2-deg</u> stripes	3WayReplication with <u>two</u> failed drives: % of <u>critical</u> stripes	3WayReplication with <u>three</u> failed drives: <u>data loss</u>
DSS-G 2 0 1	6	24	12	(3/12) = 25.00%	(3/12) * (2/11) = 4.55%	
DSS-G 2 0 2	8	48	24	(3/24) = 12.50%	(3/24) * (2/23) = 1.09%	
DSS-G 2 0 3	10	72	36	(3/36) = 8.33%	(3/36) * (2/35) = 0.48%	
DSS-G 2 0 4	12	96	48	(3/48) = 6.25%	(3/48) * (2/47) = 0.27%	

Table 12 Rebuild Volume for SSD models with 3WayReplication

DSS-G "SSD" Model:	Size [U]	Drives per BB	Drives per DA1	4WayReplication with <u>one</u> failed drive: % of <u>1/3-deg</u> stripes	4WayReplication with <u>two</u> failed drives: % of <u>2/3-deg</u> stripes	4WayReplication with <u>three</u> failed drives: % of <u>critical</u> stripes
DSS-G 2 0 1	6	24	12	(4/12) = 33.33%	(4/12) * (3/11) = 9.09%	(4/12) * (3/11) * (2/10) = 1.8182%
DSS-G 2 0 2	8	48	24	(4/24) = 16.67%	(4/24) * (3/23) = 2.17%	(4/24) * (3/23) * (2/22) = 0.1976%
DSS-G 2 0 3	10	72	36	(4/36) = 11.11%	(4/36) * (3/35) = 0.95%	(4/36) * (3/35) * (2/34) = 0.0560%
DSS-G 2 0 4	12	96	48	(4/48) = 8.33%	(4/48) * (3/47) = 0.53%	(4/48) * (3/47) * (2/46) = 0.0231%

Table 13 Rebuild Volume for SSD models with 4WayReplication

Rebuild volume for DSS-G small form-factor HDD models

DSS-G "HDD" Model:	Size [U]	Drives per BB	Drives per DA1	8+2P with <i>one</i> failed drive: % of <i>1/2-deg</i> stripes	8+2P with <i>two</i> failed drives: % of <i>critical</i> stripes	8+2P with <i>three</i> failed drives: <i>data loss</i>
DSS-G 2 0 2	8	46	23	(10/23) = 43.48%	(10/23) * (9/22) = 17.79%	
DSS-G 2 0 4	12	94	47	(10/47) = 21.28%	(10/47) * (9/46) = 4.16%	
DSS-G 2 0 6	16	142	71	(10/71) = 14.08%	(10/71) * (9/70) = 1.81%	

Table 14 Rebuild Volume for HDD Models with 8+2P Reed-Solomon

DSS-G "HDD" Model:	Size [U]	Drives per BB	Drives per DA1	8+2P with <i>one</i> failed drive: % of <i>1/2-deg</i> stripes	8+2P with <i>two</i> failed drives: % of <i>critical</i> stripes	8+2P with <i>three</i> failed drives: <i>data loss</i>
DSS-G 2 0 2	8	46	23	(11/23) = 47.83%	(11/23) * (10/22) = 21.74%	(11/23) * (10/22) * (9/21) = 9.317%
DSS-G 2 0 4	12	94	47	(11/47) = 23.40%	(11/47) * (10/46) = 5.09%	(11/47) * (10/46) * (9/45) = 1.018%
DSS-G 2 0 6	16	142	71	(11/71) = 15.49%	(11/71) * (10/70) = 2.21%	(11/71) * (10/70) * (9/69) = 0.289%

Table 15 Rebuild Volume for HDD Models with 8+3P Reed-Solomon

DSS-G "HDD" Model:	Size [U]	Drives per BB	Drives per DA1	8+2P with <i>one</i> failed drive: % of <i>1/2-deg</i> stripes	8+2P with <i>two</i> failed drives: % of <i>critical</i> stripes	8+2P with <i>three</i> failed drives: <i>data loss</i>
DSS-G 2 0 2	8	46	23	(3/23) = 13.04%	(3/23) * (2/22) = 1.19%	
DSS-G 2 0 4	12	94	47	(3/47) = 6.38%	(3/47) * (2/46) = 0.28%	
DSS-G 2 0 6	16	142	71	(3/71) = 4.23%	(3/71) * (2/70) = 0.12%	

Table 16 Rebuild Volume for HDD Models with 3WayReplication

DSS-G "HDD" Model:	Size [U]	Drives per BB	Drives per DA1	8+2P with <i>one</i> failed drive: % of <i>1/2-deg</i> stripes	8+2P with <i>two</i> failed drives: % of <i>critical</i> stripes	8+2P with <i>three</i> failed drives: <i>data loss</i>
DSS-G 2 0 2	8	46	23	(4/23) = 17.39%	(4/23) * (3/22) = 2.37%	(4/23) * (3/22) * (2/21) = 0.2259%
DSS-G 2 0 4	12	94	47	(4/47) = 8.51%	(4/47) * (3/46) = 0.56%	(4/47) * (3/46) * (2/45) = 0.0247%
DSS-G 2 0 6	16	142	71	(4/71) = 5.63%	(4/71) * (3/70) = 0.24%	(4/71) * (3/70) * (2/69) = 0.0070%

Table 17 Rebuild Volume for HDD Models with 4WayReplication

Rebuild performance measurements

Section “Determining the volume of critical and non-critical rebuilds” on page 20 described the theoretical expectations of the percentage of user data that becomes degraded when one or more pdisks in a Spectrum Scale RAID Distributed Array fail.

This section presents actual measurements of critical and non-critical rebuild times on several DSS-G building blocks with varying degrees of file system utilization. Together with the data from the previous section, these results can be used as general guidelines to estimate the duration of critical and non-critical rebuilds.

As shown in Tables 6 to 17, the percentage of RAID stripes that are impacted by one or multiple pdisk failures in a DA depends on two main factors:

- ▶ The size of the Distributed Array, which is the number of physical disk drives over which the strips of a RAID stripe can be distributed. Larger Distributed Arrays result in a smaller fraction of degraded stripes.
- ▶ The strength of the RAID code of the vdisk: The 2-fault-tolerant codes (3WayReplication and 8+2P) result in a higher fraction of critical stripes than the stronger 3-fault-tolerant codes (4WayReplication and 8+3P).

The rebuild time is directly proportional to the percentage of the data volume that has become degraded due to the pdisk failure(s). The other factors that determine the absolute data volume and the associated rebuild time are the following:

- ▶ The raw capacity of the physical disk drives (for example, 4TB or 10TB). The bigger the pdisk size, the bigger the absolute data volume.
- ▶ The percentage of the DA capacity that is allocated to the vdisk. The administrator may have allocated the full capacity of the DA (to one or multiple vdisks), or may have left a portion of the DA capacity unallocated. Only space allocated to vdisks needs to be rebuild.
- ▶ The percentage of the file system capacity that is actually used by files.

The last bullet highlights a distinguishing feature of Spectrum Scale RAID: Because GNR integrates the software RAID layer with the file system layer, GNR is aware which of the vdisk tracks are actually used by the file system. Even when 100% of a DA’s capacity has been allocated to a vdisk and that vdisk has been added as an NSD to a Spectrum Scale file system, GNR does not have to rebuild the vdisk tracks that have been allocated in the vdisk layer but are not (yet) used by files in the file system.

Other declustered RAID solutions that operate only on the block storage level lack this tight integration, and consequently they have to assume that all space that was allocated on the block level needs to be completely rebuilt when a disk failure happens.

Don’t perform tests on a production system: The tests in this section should not be performed on a production system; they should only be performed in dedicated test environments (or before a production building block is put into user operation).

In any case, it should always be verified that the system is in a completely healthy state before simulating a pdisk failure by running the following command:

```
mm1spdisk a11 --not-ok
```

Neglecting to check the system health and then manually killing a pdisk may increase the number of faulty pdisks beyond the fault tolerance of one or more vdisks. This may cause unexpected (and unrecoverable) data loss.

Simulating single disk failures and multiple disk failures

To study the rebuild behavior of a DSS-G building block, the administrator can simulate the failure of one or more pdisks by using the following command:

```
mmchpdisk --simulate-dead
```

This error injection causes the Spectrum Scale RAID software to initiate the same rebuild activities that a genuinely “dead” pdisk would trigger. After the rebuild testing has completed, those pdisks can be switched back into normal operation by using the `--revive` option:

```
mmchpdisk --revive
```

Like the physical replacement of a genuinely “dead” pdisk, reviving a pdisk that was in SimulatedDead state may trigger some additional activities like the rebalancing of capacity across all the pdisks in the Distributed Array.

The best way to observe the rebuild operations in a Spectrum Scale RAID building block is to open several terminal windows on the DSS-G server that is the active RG server for the Recovery Group into which the simulated pdisk failures will be injected, and perform the following:

- ▶ One terminal should display the GPFS log in `/var/adm/ras/mmfs.log.latest`, for example using `tail -f`. Entries in the GPFS log will signal the beginning and end of all relevant activities, including timestamps, so this is the best way to time the rebuild phases.
- ▶ One terminal can display the Recovery Group status at regular intervals, using the following command:

```
mmisrecoverygroup $RG -L --pdisk
```

This will show the state of the DA and the percentage of completion of the DA rebuild. It will also show a decrease of free capacity of the surviving pdisks over time, as spare capacity is being consumed to store the reconstructed parity strips.

- ▶ A third terminal window can be used to issue the `mmchpdisk` commands that inject the simulated pdisk faults, and to revive those pdisks after the tests.

When simulating multiple pdisk faults, it is best to choose pdisks from the same JBOD drawer in order to simulate the worst-case scenario of pdisk faults. Due to the GNR awareness of SAS failure domains, killing pdisks from different drawers will result in slightly different rebuild activities that do not represent the worst possible scenario.

For the following demonstrations we use a DSS-G210 building block with 10TB drives, and all of the DA capacity in both RGs of (dss21 and dss22) has been allocated to vdisks with 8+2P erasure coding. Those two vdisks have been assigned as NSDs to a Spectrum Scale file system, and the file system has been filled with user data up to a threshold of 20%, as can be seen in the `df` and `mmdf` outputs in Example 6.

Example 6 Verifying the percentage of user data in a Spectrum Scale file system

```
[root@dss21 ~]# df -t gpfs
Filesystem      1K-blocks      Used   Available Use% Mounted on
dss_g210_opa_16m 605229678592 118574350336 486655328256  20% /gpfs/dss_g210_opa_16m
```

```
[root@dss21 ~]# mmdf dss_g210_opa_16m
disk          disk size  failure holds   holds          free in KB          free in KB
name          in KB    group metadata data             in full blocks      in fragments
-----
Disks in storage pool: system (Maximum disk size allowed is 2.21 PB)
dss21_data1_8p2 302614839296      -1 Yes      Yes  243327320064 ( 80%)  4399200 ( 0%)
```

```

dss22_data1_8p2  302614839296      -1 Yes      Yes      243328008192 ( 80%)      4386608 ( 0%)
-----
(pool total)     605229678592                      486655328256 ( 80%)      8785808 ( 0%)
=====
(total)          605229678592                      486655328256 ( 80%)      8785808 ( 0%)

```

Inode Information

```

-----
Number of used inodes:      7205
Number of free inodes:     1000411
Number of allocated inodes: 1007616
Maximum number of inodes:  134217728

```

Simulating a single pdisk fault

A single disk fault is now simulated by killing one pdisk (e1d1s19) in RG dss21, using the following command:

```
[root@dss21 ~]# mmchpdisk dss21 --pdisk e1d1s19 --simulate-dead
```

This command results in entries in the GPFS log, as shown in Example 7.

Example 7 GPFS log output simulating a single pdisk failure

```

2019-04-01_10:41:58.096-0400: [I] Command: tschpdisk --recovery-group dss21
                                --pdisk e1d1s19 --state SimulatedDead
2019-04-01_10:41:58.132-0400: [E] Pdisk e1d1s19 of RG dss21 failed; state
simulatedDead/1000.000 location 'J1256N1-1-19' type '01CX798' WWID 5000c50095306a1b ❶
2019-04-01_10:41:58.132-0400: [E] Replace pdisk e1d1s19 of RG dss21; state
simulatedDead/1000.020 location 'J1256N1-1-19' type '01CX798' WWID 5000c50095306a1b ❶
2019-04-01_10:41:58.232-0400: [I] Start repairing RGD/VCD in RG dss21. ❷
2019-04-01_10:42:00.345-0400: [I] Finished repairing RGD/VCD in RG dss21.
2019-04-01_10:42:00.686-0400: [I] Start normal rebuild on DA DA1 in RG dss21. ❸
2019-04-01_10:42:00.688-0400: [I] Start draining PGs of DA DA1 in RG dss21.
2019-04-01_14:11:44.787-0400: [I] Finish draining PGs of DA DA1 in RG dss21.
2019-04-01_14:11:45.151-0400: [I] Start normal rebuild on DA DA1 in RG dss21. ❹
2019-04-01_14:11:45.161-0400: [I] Start draining PGs of DA DA1 in RG dss21.
2019-04-01_14:15:32.914-0400: [I] Finish draining PGs of DA DA1 in RG dss21.
2019-04-01_14:15:32.914-0400: [I] Finish normal rebuild of DA DA1 in RG dss21. ❺

```

The following can be observed:

- ▶ The command is logged, and two [E] error entries (❶ in Example 7) show the pdisk's state change to SimulatedDead as well as a notification to replace the pdisk. The subsequent rebuild happens in the following phases:
- ▶ Initially there are a few seconds of repairing RGD/VCD (❷), during which the internal data structures of Spectrum Scale RAID are repaired.
- ▶ In this example there is no rebuild-critical phase: The minimum fault tolerance of Spectrum Scale RAID vdisks is two, so a single pdisk fault will not trigger a critical rebuild.
- ▶ The first Start normal rebuild message (❸) in the log signals the beginning of the rebuild-1r phase, in which the missing parity of the 1/2-degraded stripes of the 8+2P user vdisk are reconstructed. This phase ends when the second Start normal rebuild message (❹) is issued. In this example, this phase takes about 3.5 hours.
- ▶ Note that the second normal rebuild phase takes only a few minutes (from the second Start normal rebuild message (❹) to the Finish normal rebuild message (❺)). This phase only appears because there is a 4WayReplicated logHome vdisk in DA1, and this vdisk undergoes a rebuild-2r phase in which its 1/3-degraded stripes are re-replicated.

Simulating two pdisk faults

After reviving the pdisk and waiting for the DA to rebalance, the test can be repeated at the next level by killing two pdisks in RG dss21 at the same time. We again used the simulate-dead parameter to simulate the failure of pdisks e1d1s20 and e1d1s21:

```
[root@dss21 ~]# mmchpdisk dss21 --pdisk e1d1s20 --simulate-dead ;  
mmchpdisk dss21 --pdisk e1d1s21 --simulate-dead
```

This command results in entries in the GPFS log, as shown in Example 8.

Example 8 GPFS log output simulating a two-pdisk failure

```
2019-04-02_02:48:07.638-0400: [I] Command: tschpdisk --recovery-group dss21  
--pdisk e1d1s20 --state SimulatedDead  
2019-04-02_02:48:07.667-0400: [E] Pdisk e1d1s20 of RG dss21 failed; state  
simulatedDead/1000.000 location 'J1256N1-1-20' type '01CX798' WWID 5000c5009531ff77  
2019-04-02_02:48:07.668-0400: [E] Replace pdisk e1d1s20 of RG dss21; state  
simulatedDead/1000.020 location 'J1256N1-1-20' type '01CX798' WWID 5000c5009531ff77  
2019-04-02_02:48:07.768-0400: [I] Start repairing RGD/VCD in RG dss21.  
2019-04-02_02:48:07.818-0400: [I] Command: tschpdisk --recovery-group dss21  
--pdisk e1d1s21 --state SimulatedDead  
2019-04-02_02:48:07.818-0400: [E] Pdisk e1d1s21 of RG dss21 failed; state  
simulatedDead/1000.000 location 'J1256N1-1-21' type '01CX798' WWID 5000c500952e0c9f  
2019-04-02_02:48:07.818-0400: [E] Replace pdisk e1d1s21 of RG dss21; state  
simulatedDead/1000.060 location 'J1256N1-1-21' type '01CX798' WWID 5000c500952e0c9f  
2019-04-02_02:48:07.934-0400: [I] Finished repairing RGD/VCD in RG dss21.  
2019-04-02_02:48:07.934-0400: [I] Start repairing RGD/VCD in RG dss21.  
2019-04-02_02:48:11.906-0400: [I] Finished repairing RGD/VCD in RG dss21.  
2019-04-02_02:48:12.241-0400: [I] Start rebuilding critical PGs of DA DA1 in RG dss21. ❶  
2019-04-02_02:48:12.256-0400: [I] Start draining PGs of DA DA1 in RG dss21.  
2019-04-02_03:19:19.855-0400: [I] Finish draining PGs of DA DA1 in RG dss21.  
2019-04-02_03:19:19.855-0400: [I] Finish rebuilding critical PGs of DA DA1 in RG dss21. ❷  
2019-04-02_03:19:20.194-0400: [I] Start normal rebuild on DA DA1 in RG dss21. ❸  
2019-04-02_03:19:20.195-0400: [I] Start draining PGs of DA DA1 in RG dss21.  
2019-04-02_09:08:05.959-0400: [I] Finish draining PGs of DA DA1 in RG dss21.  
2019-04-02_09:08:06.313-0400: [I] Start normal rebuild on DA DA1 in RG dss21. ❹  
2019-04-02_09:08:06.320-0400: [I] Start draining PGs of DA DA1 in RG dss21.  
2019-04-02_09:11:26.005-0400: [I] Finish draining PGs of DA DA1 in RG dss21.  
2019-04-02_09:11:26.005-0400: [I] Finish normal rebuild of DA DA1 in RG dss21. ❺
```

The following can be observed:

- ▶ In contrast to the single pdisk fault, killing two pdisks now triggers a critical rebuild (❶ in Example 8). In this rebuild-critical phase, the critical stripes of the 8+2P user vdisk are re-created. In this example the critical rebuild takes roughly 30 minutes (❷).
- ▶ The first (rebuild-1r) normal rebuild phase (❸) will reconstruct the 1/2-degraded stripes of the 8+2P user vdisk (and the 2/3-degraded stripes of the logHome vdisk). Because two pdisks failed, the volume of data that needs to be rebuilt in the user vdisk is bigger than in the single drive fault experiment, and consequently the rebuild time is also longer. In this example it takes almost six hours.
- ▶ The second (rebuild-2r) normal rebuild phase (❹) is short again. In this phase the 1/3-degraded stripes of the logHome vdisk are re-replicated within a few minutes (❺).

Simulating three pdisk faults

In order to test the rebuild scenario with three simultaneous drive failures, the vdisks need to be recreated with 8+3P erasure coding, and the file system is filled with user data again up to 20% file system utilization. Now three pdisks (e1d1s11, e1d1s12, e1d1s13) can be killed at the same time, using the following command:

```
[root@dss21 ~]# mmchpdisk dss21 --pdisk e1d1s11 --simulate-dead ;  
mmchpdisk dss21 --pdisk e1d1s12 --simulate-dead ;  
mmchpdisk dss21 --pdisk e1d1s13 --simulate-dead
```

This command results in entries in the GPFS log, as shown in Example 9.

Example 9 GPFS log output simulating a three-pdisk failure

```
2019-04-04_20:29:42.484-0400: [I] Command: tschpdisk --recovery-group dss22 --pdisk e1d1s11  
--state SimulatedDead  
2019-04-04_20:29:42.501-0400: [E] Pdisk e1d1s11 of RG dss22 failed; state  
simulatedDead/1000.000 location 'J1256N1-1-11' type '01CX798' WWID 5000c500952e6e9f  
2019-04-04_20:29:42.501-0400: [E] Replace pdisk e1d1s11 of RG dss22; state  
simulatedDead/1000.060 location 'J1256N1-1-11' type '01CX798' WWID 5000c500952e6e9f  
2019-04-04_20:29:42.602-0400: [I] Start repairing RGD/VCD in RG dss22.  
2019-04-04_20:29:42.619-0400: [I] Command: tschpdisk --recovery-group dss22 --pdisk e1d1s12  
--state SimulatedDead  
2019-04-04_20:29:42.619-0400: [E] Pdisk e1d1s12 of RG dss22 failed; state  
simulatedDead/1000.000 location 'J1256N1-1-12' type '01CX798' WWID 5000c500952dd39f  
2019-04-04_20:29:42.619-0400: [E] Replace pdisk e1d1s12 of RG dss22; state  
simulatedDead/1000.020 location 'J1256N1-1-12' type '01CX798' WWID 5000c500952dd39f  
2019-04-04_20:29:42.739-0400: [I] Finished repairing RGD/VCD in RG dss22.  
2019-04-04_20:29:42.739-0400: [I] Start repairing RGD/VCD in RG dss22.  
2019-04-04_20:29:42.740-0400: [I] Command: tschpdisk --recovery-group dss22 --pdisk e1d1s13  
--state SimulatedDead  
2019-04-04_20:29:42.740-0400: [E] Pdisk e1d1s13 of RG dss22 failed; state  
simulatedDead/1000.000 location 'J1256N1-1-13' type '01CX798' WWID 5000c50095302e3f  
2019-04-04_20:29:42.741-0400: [E] Replace pdisk e1d1s13 of RG dss22; state  
simulatedDead/1000.020 location 'J1256N1-1-13' type '01CX798' WWID 5000c50095302e3f  
2019-04-04_20:29:42.871-0400: [I] Finished repairing RGD/VCD in RG dss22.  
2019-04-04_20:29:42.871-0400: [I] Start repairing RGD/VCD in RG dss22.  
2019-04-04_20:29:48.259-0400: [I] Finished repairing RGD/VCD in RG dss22.  
2019-04-04_20:29:48.621-0400: [I] Start rebuilding critical PGs of DA DA1 in RG dss22. ①  
2019-04-04_20:29:48.627-0400: [I] Start draining PGs of DA DA1 in RG dss22.  
2019-04-04_20:35:11.768-0400: [I] Finish draining PGs of DA DA1 in RG dss22.  
2019-04-04_20:35:11.768-0400: [I] Finish rebuilding critical PGs of DA DA1 in RG dss22. ②  
2019-04-04_20:35:12.197-0400: [I] Start normal rebuild on DA DA1 in RG dss22. ③  
2019-04-04_20:35:12.199-0400: [I] Start draining PGs of DA DA1 in RG dss22.  
2019-04-04_22:38:39.553-0400: [I] Finish draining PGs of DA DA1 in RG dss22.  
2019-04-04_22:38:39.996-0400: [I] Start normal rebuild on DA DA1 in RG dss22. ④  
2019-04-04_22:38:39.998-0400: [I] Start draining PGs of DA DA1 in RG dss22.  
2019-04-05_04:10:57.806-0400: [I] Finish draining PGs of DA DA1 in RG dss22.  
2019-04-05_04:10:57.806-0400: [I] Finish normal rebuild of DA DA1 in RG dss22. ⑤
```

The timing of the rebuild phases is notably different from the 8+2P test:

- ▶ The critical rebuild phase (①) is significantly shorter, as expected for 8+3P. It only took about five minutes (②), compared to half an hour for the 8+2P case.
- ▶ The first (rebuild-1r) normal rebuild phase (③), which reconstructs the 2/3-degraded stripes of the 8+3P user vdisk (and the logHome vdisk), is shorter than in the 8+2P case. This is also expected: The percentage of stripes that have lost two out of three redundancies when three drives fail in an 8+3P vdisk is smaller than the percentage of

stripes that lost one out of two redundancies when two pdisks fail in the 8+2P case. Here it took around two hours to repair those stripes.

- ▶ The second phase of the normal rebuild (Ⓐ) is now much longer than in the 8+2P examples, where only the logHome vdisk needed to be processed in rebuild-2r mode: With 8+3P and three pdisk faults there is a sizeable fraction of the user vdisk that has two redundancies left and needs to have the third redundancy strip rebuilt. This phase took roughly 5.5 hours (Ⓑ).

In the remainder of this section, we report results of running the above tests on several DSS-G building blocks, including tests of 8+2P and 8+3P erasure coding as well as tests with varying levels of file system utilization.

Rebuild performance of DSS-G large form factor NL-SAS models

Table 18 summarizes the timing of the rebuild activity for an 8+2P user vdisk after two pdisks have been set to simulatedDead state.

Building Block	pdisks in DA1	pdisk Capacity [TB]	RAID Level	vdisk filling [% vdisk]	vdisk filling [TB/pdisk]	rebuild-critical [min]	rebuild-1r [min]	rebuild-2r [min] (logHome)
DSS-G210	41	10	8+2P	0	0	0.47	8.37	3.35
DSS-G210	41	10	8+2P	10	1	13.53	162.77	2.68
DSS-G210	41	10	8+2P	20	2	31.12	348.77	3.33
DSS-G210	41	10	8+2P	40	4	47.02	598.57	2.73
DSS-G210	41	10	8+2P	60	6	84.87		
DSS-G220	83	10	8+2P	0	0	0.17	7.03	2.70
DSS-G220	83	10	8+2P	10	1	3.10	101.07	2.47
DSS-G220	83	10	8+2P	20	2	5.78	197.97	2.68
DSS-G220	83	10	8+2P	40	4	10.63	398.05	2.48

Table 18 Rebuild performance for NL-SAS models with 8+2P Reed-Solomon

It can be clearly seen that the rebuild times of the DSS-G220 model with two D3284 enclosures are much shorter than those of the DSS-G210 model with only a single D3284 enclosure. This effect is most pronounced for the critical rebuilds, as those depend quadratically on the number of pdisks.

Note that there is a short rebuild-2r phase. The user vdisk uses 8+2P protection, so there should be no degraded tracks in that vdisk that have two redundancies left and need to be rebuilt. This rebuild is for the internal logHome vdisk in DA1 that is 4WayReplicated. The logHome has a fixed size that only depends on the capacity of the individual pdisks, and not on the total number of pdisks in the DA. For this reason, the duration of the rebuild-2r phase is roughly the same for all building block sizes (assuming the same pdisk capacity), and it also does not depend on the level of file system utilization of the user vdisks.

The timing of the rebuild activity for an 8+3P user vdisk after three pdisks have been set to simulated-dead state is shown in Table 19:

Building Block	pdisks in DA1	pdisk Capacity [TB]	RAID Level	vdisk filling [% vdisk]	vdisk filling [TB/pdisk]	rebuild-critical [min]	rebuild-1r [min]	rebuild-2r [min]
DSS-G210	41	10	8+3P	0	0	0.13	3.15	8.22
DSS-G210	41	10	8+3P	10	1	3.15	60.40	173.05
DSS-G210	41	10	8+3P	20	2	5.38	123.45	332.30
DSS-G210	41	10	8+3P	40	4	10.92	240.45	665.20
DSS-G210	41	10	8+3P	80	8	25.10	489.28	1,288.20
DSS-G220	83	10	8+3P	0	0	0.08	1.08	9.58
DSS-G220	83	10	8+3P	10	1	0.62	14.25	134.23
DSS-G220	83	10	8+3P	20	2	1.58	30.37	265.62
DSS-G220	83	10	8+3P	40	4	1.47	60.50	512.03
DSS-G220	83	10	8+3P	80	8	3.55	122.75	

Table 19 Rebuild performance for NL-SAS models with 8+3P Reed-Solomon

Here the improvement of rebuild times with the size of the building block is even more pronounced than in the 8+2P case. This matches the theoretical expectations.

Summary

This paper has introduced the Spectrum Scale RAID technology, which is one of the distinguishing features of the Lenovo DSS-G high performance storage solution.

The presented calculations and timing measurements of critical and non-critical rebuilds should enable solution architects and storage administrators to select the most suitable sizes of DSS-G models, and the most suitable data protection levels, for their environments.

In particular, the 8+3P protection scheme is also very beneficial to reduce critical rebuild times for the small DSS-G models. This technology is relevant far beyond its original purpose to build exascale storage systems.

Appendix: Conversion of Decimal and Binary Units

When measuring capacity and bandwidth of high-performance storage systems, the numerical differences between base-10 units and base-2 units are significant. For example, 1000 bytes equals one *kilobyte*, with the well-known decimal prefixes of the international SI System. On the other hand, 1024 bytes equals one *kibibyte* with the less well-known binary prefixes (which were first defined in IEC 60027-2). The effect of this difference in scale is compounding with each order of magnitude, and at Petascale it results in a deviation of over 11%: One Petabyte equals only 0.888 Pebibyte.

Table 20 shows the difference between the two measurement units.

Table 20 Storage capacity measured in base-10 units and base-2 units

Base-10 units			Base-2 units			Ratio	Delta
Prefix	Unit	In bytes	Prefix	Unit	Bytes		
Kilo	K	$10^3 = 1,000$	Kibi	Ki	$2^{10} = 1,024$	0.976	2.34%
Mega	M	$10^6 = 1,000,000$	Mebi	Mi	$2^{20} = 1,048,576$	0.953	4.63%
Giga	G	$10^9 = 1,000,000,000$	Gibi	Gi	$2^{30} = 1,073,741,824$	0.931	6.87%
Tera	T	$10^{12} = 1,000,000,000,000$	Tebi	Ti	$2^{40} = 1,099,511,627,776$	0.909	9.05%
Peta	P	$10^{15} = 1,000,000,000,000,000$	Pebi	Pi	$2^{50} = 1,125,899,906,842,620$	0.888	11.18%
Exa	E	$10^{18} = 1,000,000,000,000,000,000$	Exbi	Ei	$2^{60} = 1,152,921,504,606,850,000$	0.867	13.26%

The industry standard is to quote disk capacities in base-10 units and memory capacities in base-2 units. For bandwidth (which is capacity transferred per unit of time), there is no clear standard. Since most high performance I/O applications are transferring data that resides in memory (sized in base-2) to and from disk, it seems more natural to use base-2 numbers to quote bandwidth figures. This convention is followed in this document.

Additional resources

- ▶ DSS-G datasheet
<https://lenovopress.com/ds0026>
- ▶ DSS-G product guide:
<https://lenovopress.com/lp0837>
- ▶ Lenovo Storage D3284 External High Density Drive Expansion Enclosure
<https://lenovopress.com/lp0513>
- ▶ Lenovo Storage D1212 and D1224 Drive Enclosures
<https://lenovopress.com/lp0512>
- ▶ IBM Research GNR presentation at USENIX LISA 2011:
<http://www.youtube.com/watch?v=2g5rx4gP6yU>
- ▶ IBM Spectrum Scale 5.0.2 Product Documentation:
 - Installation Guide
 - Administration Guide
 - Command and Programming Referencehttps://www.ibm.com/support/knowledgecenter/STXKQY_5.0.2/com.ibm.spectrum.scale.v5r02.doc/b11ins_planning_MCS.htm
- ▶ IBM Spectrum Scale RAID FAQ
(focuses on IBM ESS, but most of the contents also applies to Lenovo DSS-G)
<http://www.ibm.com/support/knowledgecenter/SSYSP8/gnrfaq.html>
<http://www.ibm.com/support/knowledgecenter/SSYSP8/gnrfaq.pdf>
- ▶ IBM Spectrum Scale Wiki:
[https://www.ibm.com/developerworks/community/wikis/home/wiki/General%20Parallel%20File%20System%20\(GPFS\)?lang=en](https://www.ibm.com/developerworks/community/wikis/home/wiki/General%20Parallel%20File%20System%20(GPFS)?lang=en)

Author

Michael Hennecke is Lenovo's HPC Chief Technologist and is responsible for Lenovo's HPC storage strategy. He has over 26 years of experience in High Performance Computing. Michael has been working with IBM Spectrum Scale since GPFS version 1.1 and is one of the "fathers" of the GPFS Storage Server (GSS) which Lenovo has evolved into the Distributed Storage Solution for IBM Spectrum Scale RAID (DSS-G). He holds a masters degree in physics from Ruhr-Universität Bochum (Germany), and a "Distinguished IT Specialist" certification from The Open Group.

Thanks to the following people for their contributions to this project:

- ▶ David Watts
- ▶ David Worley
- ▶ Ray Paden

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document was created or updated on September 19, 2019.

Send us your comments via the **Rate & Provide Feedback** form found at <http://lenovopress.com/lp1227>

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at <http://www.lenovo.com/legal/copytrade.html>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Flex System™	ServeRAID™	System x®
Lenovo®	ServerGuide™	vNIC™
Lenovo(logo)®	ServerProven®	

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.