Introducing the IBM Elastic Storage System 3500

ESS 6.1.4
Spectrum scale 5.1.4.1 efix 13
30 August 2022

Stieg Klein
Spectrum Scale / ESS SME
IBM Advanced Technology Group
Accelerate with IBM Storage Technical Webinar Series

*Advanced Technology Group* – Storage experts cover a variety of technical topics.

**Audience:** Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

**Accelerate with IBM Storage Support Site:**

**ATG – Storage Media Center Channel:**
[https://ibm.biz/BdfEgQ](https://ibm.biz/BdfEgQ)

**2022 Upcoming Webinars:**

**September 13** – Detecting Cyber-Attacks with IBM Spectrum Sentinel
Register [HERE](https://www.ibm.com/support/pages/node/1125513) or [https://ibm.biz/BdPiqj](https://ibm.biz/BdPiqj)

**September 27** – Accelerate Your SAS Analytics with IBM Hardware and Software
Register [HERE](https://www.ibm.com/support/pages/node/1125513) or [https://ibm.biz/BdPjQf](https://ibm.biz/BdPjQf)

**September 29** - Introduction to IBM Cyber Vault for Z
Register [HERE](https://www.ibm.com/support/pages/node/1125513) or [https://ibm.biz/BdPZ6e](https://ibm.biz/BdPZ6e)
ATG-Storage Offerings

CLIENT WORKSHOPS

➢ IBM Cloud Object Storage System Workshop
➢ IBM Spectrum Scale and Elastic Storage System
➢ IBM DS8900F Advanced Functions
➢ Improve your Cyber Resiliency with IBM Storage
➢ IBM FlashSystem 9500 Deep Dive & Advanced Functions
➢ IBM Spectrum Virtualize and Storage Insights
➢ IBM Spectrum Fusion

Please reach out to your IBM Rep or Business Partner for future dates and to be nominated.

TEST DRIVE / DEMO'S

➢ North America ATG-Storage ESS & Spectrum Scale GUI, Functionality, PoT
➢ North America ATG-Storage IBM Spectrum Virtualize Test Drive
➢ North America ATG-Storage IBM DS8900F Storage Management Test Drive
➢ North America ATG-Storage Managing Copy Services on the DS8000 Using IBM Copy Services Manager Test Drive
➢ North America ATG-Storage IBM DS8900F Safeguarded Copy (SGC) Test Drive
➢ North America ATG-Storage IBM Cloud Object Storage Test Drive - (Appliance based)
➢ North America ATG-Storage IBM Cloud Object Storage Test Drive - (VMware based)
➢ North America ATG-Storage Spectrum Protect Live Test Drive
➢ North America ATG-Storage Spectrum Protect Plus Live Test Drive
➢ North America ATG-Storage IBM Spectrum Virtualize Safeguarded Copy with IBM Copy Services Manager Test Drive

Please reach out to your IBM Rep or Business Partner for more information.

*For North America Only*
Please take a moment to share your feedback with our team!

You can access this 6-question survey via Menti.com with code 2628 0279 or

Direct link https://www.menti.com/t4hii66ky8

Or

When the webinar ends, you may see the following:

Clicking on the Continue button will take you to the survey.
Introducing the IBM Elastic Storage System 3500

ESS 6.1.4
Spectrum scale 5.1.4.1 efix 13
30 August 2022

Stieg Klein
Spectrum Scale / ESS SME
IBM Advanced Technology Group
Notices and disclaimers

- © 2022 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.

- U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. This document is distributed “as is” without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity. IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply.”

- Any statements regarding IBM’s future direction, intent or product plans are subject to change or withdrawal without notice.

- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

- It is the customer’s responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer’s business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.
Notices and disclaimers

— Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM’s products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

— The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

— IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)
Agenda

- Spectrum Scale
- What’s an ESS
- ESS Advantages
- IBM Elastic Storage System latest model – 3500
- Spectrum Scale RAID
IBM Spectrum Scale - High Performance Clustered File System

Global name space

IBM Spectrum Scale
Automated data placement and data migration

POSIX  NFS  SMB/CIFS  Object  HDFS Transparency

Flash  Disk  Tape  Transparent Cloud Tiering  Cloud Data Sharing

On/Off Premise

Site A  Site B  Site C

NVMe ESS For high IOP/sec  IBM Elastic Storage Systems w/HDD

IBM Cloud Object Storage

Users and AI applications

Compute farm

Big Data Analytics

Client workstations
Beyond the Usual File System Capabilities

- Use with any block storage usable by the operating system
- Tape and Cloud supported as external storage pool tiers
- Replication capabilities for robust Disaster Recovery
- Built-in encryption, compression and secure erase
What is the “IBM Elastic Storage System”
Designed/Validated/Installed Spectrum Scale building block

Thousands of deployed systems

Latest model: ESS 3500

Spectrum Scale + GNR+ ESS Utilities + ESS aware GUI

Ansible playbook driven installation and updates

ESS mitigates risks and makes it quicker to deploy and grow a Spectrum Scale cluster
What is an ESS solution?

There must be at least one ESS Management Server (EMS) within the Spectrum Scale cluster to manage all the ESS building blocks.

- ESS GUI runs on the EMS
- ESS EMS is POWER9 based (ESS 3500 & 5000’s)

A single ESS building block consists of:
- A pair of recovery group servers (aka I/O servers)
- Storage connected to both I/O servers (twin tailed)

ESS 3500 System includes 2 @ x86-64 integrated I/O servers
NVMe storage + optionally Serial attached SCSI storage enclosures (ESS v6.1.4)

ESS 5000 System include 2 @ POWER9 I/O servers + Serial attached SCSI storage

Multiple ESS building blocks may participate in a single Spectrum Scale cluster.
- File systems may span multiple building blocks.
2U form factor includes 2 servers 12 or 24 NVMe drives
- Drive size: 3.84/7.68/15.36/30.74 TB PCIe Gen4 NVMe Flash Drive
- 2.5-inch Small Form Factor (SFF) hot swappable
- Non-Volatile Memory express (NVMe) drive transport protocol
- ESS 3500 NVMe raw capacity: min 46TB / max 737 TB

ESS3500 with ESS v6.1.4 adds
0-4@4U form factor SAS/SATA storage enclosure 102 drives
- Drive size: 10/14/18/20 TB PCIe HDD
  - 1st shelf ½ or full populated 52 or 102 drives
  - 2.5-inch or 3.5-inch drives hot-swappable
- ESS 3500 HDD raw capacity: min 520 TB / max 8,160TB

- ESS 3500 max raw capacity ~8.9PB
Elastic Storage System - ESS 3500Hx – Models

2 servers with NVMe drives
Optionally 1 - 4 Storage shelves for capacity storage

ESS 3500 / ESS 3500H0
2U/24 NVMe

ESS 3500H1
2U/24 NVMe + 1 shelf - 4U/102 HDD

ESS 3500H2
2U/24 NVMe + 2 shelves - 8U/204 HDD

ESS 3500H3
2U/24 NVMe + 3 shelves - 12U/306 HDD

ESS 3500H4
2U 24 NVMe + 4 shelves - 16U/408 HDD
Elastic Storage System - ESS 3500Cx – Capacity storage (only) Models

2 servers with external storage shelves for capacity storage

ESS 3500C1
2U 4 NVMe - logtip only
1 shelf - 4U/102 HDD

ESS 3500C2
2U 4 NVMe logtip only
2 shelves - 8U/204 HDD

ESS 3500C3
2U 4 NVMe logtip only
3 shelves - 12U/306 HDD

ESS 3500C4
2U 4 NVMe logtip only
4 shelves - 16U/408 HDD
ESS 3500 Hardware High Level Architecture and Topology

- NVMe attached SSD drives
- 4 @ PCIe4 x16 slots PCIe slots (PCIe4 supports 2GB/s per lane)

ESS 3500 Storage Enclosure 5147-102

NVMe - Non-Volatile Memory express
ESS 3500 Front & Rear views

Front view

Two Canisters / Servers
Two Power Supplies

Canister A/1
Canister B/2

Orientation of ports
ESS 3500 - Networking

12Gb/s SAS connected to one to four enclosures

BMC – Baseboard Management Controller
ESS – Elastic Storage System
EMS – ESS Management Server
ESA – Electronic Service Agent
SSR – Support Service Representative
Each 4 port SAS Adapter
Support up to 2 storage shelves
ESS 3500 is a customer setup (CSU) product with a combination of customer-replaceable units (CRUs) and field-replaceable units (FRUs).

IBM Support Service Representative (SSR) deals with FRUs.

### Field Replaceable Unit (FRU) vs. Customer Replaceable Unit (CRU)

<table>
<thead>
<tr>
<th>Field Replaceable Unit (FRU)</th>
<th>Customer Replaceable Unit (CRU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canister</td>
<td>NVMe drive / Bezel</td>
</tr>
<tr>
<td>Memory DIMM</td>
<td>Power supply / Power cord</td>
</tr>
<tr>
<td>Adapter</td>
<td>Cables</td>
</tr>
<tr>
<td>M.2 boot drive</td>
<td></td>
</tr>
</tbody>
</table>
Elastic Storage System - ESS 3500 – Configuration Options

NVMe Capacity – ½ or Fully populated (ESS 3500Hx)
NVMe Drive size – 3.84/7.68/15.36/30.74 TB
Usable* Min/Max ~25.6/427 TB

High Speed Network – Two PCIe slots per canister
1 or 2 port InfiniBand adapter- EDR 100 Gb / HDR 200 Gb
1 or 2 port Ethernet adapter - 100 GbE / 40 GbE

*GNR configuration
- 3% meta data @ 3-way replication, 97% data @ 8+2P
- TRIM and internal data structures plus reserve 10% capacity
**ESS 5000 SL Series**

- Proven 5U92 enclosure – Standard Rack Depth
- Ability to grow (capacity and performance)
- Up to 6 Network Adapters per ESS
  - 100G Ethernet, EDR InfiniBand, HDR InfiniBand
- Five HDD options: 6TB, 10TB, 14TB, 16TB, 18TB

<table>
<thead>
<tr>
<th>Model</th>
<th>Enclosures</th>
<th>U Height</th>
<th>NL-SAS</th>
<th>SSD</th>
<th>Capacity Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model SL1:</td>
<td>2 Enclosures</td>
<td>9U</td>
<td>182 NL-SAS</td>
<td>2 SSD</td>
<td>552TB - 1.47PB (raw)</td>
</tr>
<tr>
<td>Model SL2:</td>
<td>3 Enclosures</td>
<td>19U</td>
<td>274 NL-SAS</td>
<td>2 SSD</td>
<td>1.0PB - 2.9PB (raw)</td>
</tr>
<tr>
<td>Model SL3:</td>
<td>4 Enclosures</td>
<td>24U</td>
<td>366 NL-SAS</td>
<td>2 SSD</td>
<td>1.6PB – 4.3PB (raw)</td>
</tr>
<tr>
<td>Model SL4:</td>
<td>5 Enclosures</td>
<td>29U</td>
<td>458 NL-SAS</td>
<td>2 SSD</td>
<td>2.19PB – 5.8PB (raw)</td>
</tr>
<tr>
<td>Model SL5:</td>
<td>6 Enclosures</td>
<td>34U</td>
<td>550 NL-SAS</td>
<td>2 SSD</td>
<td>2.7PB – 7.3PB (raw)</td>
</tr>
<tr>
<td>Model SL6:</td>
<td>7 Enclosures</td>
<td>42U</td>
<td>642 NL-SAS</td>
<td>2 SSD</td>
<td>3.3PB – 8.8PB (raw)</td>
</tr>
<tr>
<td>Model SL7:</td>
<td>8 Enclosures</td>
<td>34U</td>
<td>642 NL-SAS</td>
<td>2 SSD</td>
<td>3.8PB – 11.5PB (raw)</td>
</tr>
</tbody>
</table>
**ESS 5000 SC Series**

Highest Capacity ESS model family
- Uses the same 4U106 enclosure as “mini CORAL” (enclosure 5147-106)
- Up to 6 Network Adapters per ESS
  - 100G Ethernet, EDR/HDR InfiniBand
- Three HDD options: 10TB, 14TB, 16TB

<table>
<thead>
<tr>
<th>Model</th>
<th>Height</th>
<th>Components</th>
<th>Capacity (raw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1</td>
<td>8U</td>
<td>104 NL-SAS, 2 SSD</td>
<td>1.0 – 1.6PB</td>
</tr>
<tr>
<td>SC2:</td>
<td>12U</td>
<td>210 NL-SAS, 2 SSD</td>
<td>2.1 – 3.3PB</td>
</tr>
<tr>
<td>SC3:</td>
<td>20U</td>
<td>316 NL-SAS, 2 SSD</td>
<td>3.1 – 5 PB</td>
</tr>
<tr>
<td>SC4:</td>
<td>24U</td>
<td>528 NL-SAS, 2 SSD</td>
<td>4.2 – 6.7PB</td>
</tr>
<tr>
<td>SC5:</td>
<td>28U</td>
<td>634 NL-SAS, 2 SSD</td>
<td>5.2 – 8.4PB</td>
</tr>
<tr>
<td>SC6:</td>
<td>32U</td>
<td>740 NL-SAS, 2 SSD</td>
<td>6.3 – 10PB</td>
</tr>
<tr>
<td>SC7:</td>
<td>36U</td>
<td>846 NL-SAS, 2 SSD</td>
<td>6.3 – 11.8PB</td>
</tr>
<tr>
<td>SC8:</td>
<td>36U</td>
<td>952 NL-SAS, 2 SSD</td>
<td>8.4 – 13.5PB</td>
</tr>
<tr>
<td>SC9:</td>
<td>36U</td>
<td>952 NL-SAS, 2 SSD</td>
<td>9.5 – 15.2PB</td>
</tr>
</tbody>
</table>

© Copyright IBM Corporation 2022
ESS 5000 I/O server

POWER9 IO Server

NOTE: Slots C5 & C10
Not Available on 2S2U Server

Flexible Service Processor
[HW MGMT & Control]

3x gen4 x 16 x CX5 Network
2 ports each
5x gen3 X8 HBA
1x gen3 x4 Ethernet
ESS 5000 Protocol Server

NOTE: Slots C5 & C10 Not Available on 2S2U Server

Up to seven network adapters

Flexible Service Processor [HW MGMT & Control]

POWER9 Protocol Server

Slot C2 Network
Slot C3 Network
Slot C4 Network
Slot C6 Network
Slot C8 Network
Slot C9 Network
Slot 11 Quad Ported Gb Ethernet [OS Provisioning]
Slot C12 Network
ESS 5000 Networking (EMS, Protocol nodes)
ESS 5000 Network (P9 + Legacy + 3K)

- **ESS Networking**
- **5105 - P9ZZ Little Endian with Protocol Nodes**
- **100 GbE**
- **100 Gb EDR**
- **External/Campus Network**
- **Clustering/Data Network**
- **ESS 3000**
- **ESS 5K I/O Server – essio1**
- **ESS 5K I/O Server – essio2**
- **ESS I/O Server – gssio1**
- **ESS I/O Server – gssio2**

**Location and Function**
- C11 T1 (IO) – Mgmt
- C11 T2 (FI) – FSP
- C11 T3 (FG) – optional campus
- C11 T4 (FS) – SSR

**Dedicated mgmt. VLAN for ESS 3000**

**MGMT LAN shared with ESS**
- One DHCP server at a time

© Copyright IBM Corporation 2022
ESS high-level software architecture

ESS 6.1.4
Software Stack (ESS5K):
- Operating System: Redhat 8.x
- Spectrum Scale 5.1.4.1
- OFED
- Boot Drive Firmware
- Canister Firmware
- Kernel
- Ansible
- Podman

Elastic Storage System (ESS)
Flexible Service Processor (FSP)
Gems + Enclosure Management Structure (GEMS)
Red Hat Universal Base Images (RH UBI)
Mellanox OpenFabrics Enterprise Distribution (MOFED)
GEMS/PEMS architecture [utilizes SCSI Enclosure Services]

ESS3200 PEMS SCSI/SES Page Request Flow with IPMI/BMC Backend

BMC – Baseboard Management Controller
EMS - Enclosure Management Services
ESS - Elastic Storage System
ESM – Enclosure Services Manager
GEMS – GNR Enclosure Management Services (user space)
IPMI - Intelligent Platform Management Interface
PEMS – Platform Enclosure Management Services (kernel space)
RD – Receive Diagnostic Results / Send Diagnostic reply
SCSI – Small Computer Systems Interface
SES – SCSI Enclosure Services (ANSI SFF-8067 and SFF-8045)
VPD – Vital Product Data
Elastic Storage System models over time

1st generation
4Q 2014
---
Model 5146 / 1818

2nd generation
2Q 2017
---
Model 5148 / 5147-024 / 5147-084

3rd generation
Model 5141-AF8 (NVMe Flash)
---
ESS 3000
4Q 2019

3rd generation
Model 5141-FN1 (NVMe Flash)
---
ESS 5000
3Q 2020

3rd generation
Model 5141-FN1 (NVMe Flash)
---
ESS 3200
2Q 2021

3rd generation
Model 5141-FN2 (NVMe Flash)
---
ESS 3500
2Q 2022

3rd generation
Model 5141-FN2 (NVMe Flash) + 5147-102 SAS/SATA HDD
---
ESS 3500
3Q 2022

Model 5146 / 1818-80E
Spectrum Scale RAID

... the special sauce in the Elastic Storage System

“IBM Spectrum Scale RAID is the most important component of the ESS 5000 software stack.”
Spectrum Scale RAID is a software implementation of “declustered RAID”

- A sophisticated data and spare space disk layout scheme allowing for arbitrarily sized disk arrays while reducing the overhead when recovering from disk failures
- Extremely fast rebuild after a disk failure, with minimal impact on performance
- Very strong data integrity checks
- Additional erasure codes, such as 8+3p
- Error detection codes enable detecting track errors and dropped writes
- Consistent performance from 0 – 99% utilization or 1 to many jobs in parallel

Spectrum Scale RAID is currently available only with Elastic Storage System (IBM’s reference architecture) and Erasure Code Edition (ECE)
Spectrum Scale Native RAID erasure encoding options

- Reed-Solomon Encoding
  - 8 Data Strips + 2 or 3 parity strips
  - Stripe width 10 or 11 strips
  - Storage efficiency 80% or 73% respectively*
- 3-way or 4-way replication
  - Strip size is file system data block size
  - Storage efficiency 33% or 25% respectively

*Excluding user-configurable spare space for rebuilds
Native RAID Layout example from 2014
Declustered RAID Example

3x 1-fault-tolerant mirrored groups (RAID1)

- 3 arrays on 6 disks
- 7 spare strips
- 7 disks

7 stripes per group (2 strips per track/stripe)

- 21 stripes (42 strips)
- 49 strips
Rebuild Overhead Reduction Example

Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs

Rebuild activity spread across many disks, less disruption to user programs

Rebuild overhead reduced by 3.5x
**Declustered RAID6 Example**

14 physical disks / 3 traditional RAID6 arrays / 2 spares

14 physical disks / 1 declustered RAID6 array / 2 spares

Decluster data, parity and spare

Number of faults per stripe

<table>
<thead>
<tr>
<th></th>
<th>Red</th>
<th>Green</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of stripes with 2 faults = 7

Number of faults per stripe

<table>
<thead>
<tr>
<th></th>
<th>Red</th>
<th>Green</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Number of stripes with 2 faults = 1
Benefits of declustering in Spectrum Scale RAID

- Faster Rebuilds
- Integrated spare capability
- More predictable performance
- Only 2% rebuild performance hit

Conventional

- When one disk is down (most common case)
  - Background rebuild with minimal impact to client workload

- When three disks are down (rare case):
  - Fraction of stripes that have three failures ~1%
  - Quickly get back to non-critical (2 failures) state vs. rebuilding all stripes for conventional RAID

De-clustered
Data integrity manager

**Highest priority:** Restore redundancy after disk failure(s)

Rebuild data stripes in order of 3, 2, and 1 erasures

Fraction of stripes affected when 3 disks have failed
(assuming 8+3p, 47 disks):

- 23% of stripes have 1 erasure (= 11/47)
- 5% of stripes have 2 erasures (= 11/47 * 10/46)
- 1% of stripes have 3 erasures (= 11/47 * 10/46 * 9/45)

**Medium priority:** Rebalance spare space after disk install

Restores uniform declustering of data, parity, and spare strips.

**Low priority:** Scrub and repair media faults

Verifies checksum/consistency of data and parity/mirror.
Advantages of ESS Fast Rebuild time

- 1st disk failure
- 2nd disk failure - start of critical rebuild
- Critical rebuild finished, continue normal rebuild

4 Minutes 16 seconds critical rebuild

Rebuild of a Critical Failure in minutes instead of hours and days!
An **mmvdisk** command an integrated command suite and management methodology for Spectrum Scale RAID

```
#mmvdisk -h
Manage GNR configuration.
Usage:
mmvdisk nodeclass    - Manage server node classes
mmvdisk server       - Manage recovery group servers
mmvdisk recoverygroup - Manage recovery groups
mmvdisk vdiskset     - Manage vdisk sets
mmvdisk filesystem   - Manage file systems made from vdisk sets
mmvdisk pdisk        - Manage pdisks
mmvdisk vdisk        - Manage vdisks
```

```
[root@tukeyems ~]# mmvdisk server list

<table>
<thead>
<tr>
<th>node number</th>
<th>server</th>
<th>node class</th>
<th>recovery groups</th>
<th>remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tukeyio1-hs.sdiwsc.ess</td>
<td>ess5k_ppc64le_mmvdisk</td>
<td>ess5k_78A49FA, ess5k_78A4A0A</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>tukeyio2-hs.sdiwsc.ess</td>
<td>ess5k_ppc64le_mmvdisk</td>
<td>ess5k_78A49FA, ess5k_78A4A0A</td>
<td></td>
</tr>
</tbody>
</table>
```
Recovery Groups (RG) in an ESS 5000 are called paired recovery groups and always come in pairs.

Single ESS 5000, two I/O Servers, one storage JBOD
Server 1 is primary for RG1, Server 2 is backup for RG1
Server 2 is primary for RG2, Server 1 is backup for RG2
If Server 1 fails, Server 2 takes over control.
The load on Server 2 increases 100% from one recovery group to two.
paired recovery groups - ESS models: 5000 / GLx / GSx
where a pair of recovery groups are defined on a single common disk enclosure shared by two servers, with each server taking exclusive primary responsibility for one recovery group of the pair

shared recovery group - ESS models 3000 / 3200 / 3500
where a single recovery group is defined on a single common disk enclosure shared by two servers

The ESS 3000 / ESS 3200 / 3500 can have as few as 12 disks
12 disks allows for one equivalent spare and 11-wide 8+3P RAID codes
Spectrum Scale RAID recovery groups – shared recovery group failover

```bash
# mmvdisk recoverygroup list --recovery-group ESS3200RG --server
node
number   server                          active   remarks
-------- ------------------------------- -------- 
  3       canister1.gpfs.net            yes      serving ESS3200RG: LG001, LG003
  4       canister2.gpfs.net            yes      serving ESS3200RG: root, LG002, LG004

# mmvdisk recoverygroup list --recovery-group ESS3000RG --server
node
number   server                          active   remarks
-------- ------------------------------- -------- 
  3       canister1.gpfs.net            yes      serving ESS3200RG: root, LG001, LG002, LG003, LG004
  4       canister2.gpfs.net            no       configured

# mmvdisk rg list --not-ok
recovery group  remarks
--------------------- 
ESS3200RG       server ccanister2.gpfs.net 'down'
```
Spectrum Scale RAID

Checksums
ESS – Data Integrity Enhancements

• End-to-end *checksum provides superior protection to current hardware-based RAID arrays*
  • Checksums maintained on disk and in memory and are transmitted to/from client
  • Eliminates soft/latent read errors
  • Eliminates silent dropped writes
  • Protection against lost writes eliminates additional costs to deploy mirroring alternatives
• Advanced disk diagnostics reduces potential issues and expedites repair actions
End-to-end checksum

- **True end-to-end checksum** from disk surface to client’s Spectrum Scale interface
  - Repairs soft/latent read errors
  - Repairs lost/missing writes.
- **Checksums are maintained on disk and in memory** and are transmitted to/from client.
- **Checksum is stored in a 64-byte trailer of 32-KiB buffers**
  - 8-byte checksum and 56 bytes of ID and version info
  - Sequence number used to detect lost/missing writes.
Read Operations: When Spectrum Scale RAID reads disks to satisfy a client read operation, it compares the disk checksum against the disk data and the disk checksum version number against what is stored in its metadata.

If the checksums and version numbers match, Spectrum Scale RAID sends the data along with a checksum to the NSD client.

If the checksum or version numbers are invalid, Spectrum Scale RAID reconstructs the data using parity or replication and returns the reconstructed data and a newly generated checksum to the client.

Thus, both silent disk read errors and misplaced or skipped disk writes are detected and corrected.
Spectrum Scale RAID

Disk hospital
Comprehensive Disk and Path Diagnostics

Asynchronous disk hospital’s design allows for careful problem determination of disk fault
- While a disk is in the disk hospital, reads are parity reconstructed.
- For writes, strips are marked stale and repaired later when disk leaves.
- I/Os are resumed in under 10 seconds.

Thorough Fault Determination
- Power-cycling drives to reset them
- Neighbor checking
- Supports multi-disk carriers.

Disk Enclosure Management
- Uses SES interface for lights, latch locks, disk power, and so on.

Manages topology and hardware configuration.
Before taking severe actions against a disk, Spectrum Scale RAID checks neighboring disks to decide if some systemic problem may be behind the failure:

- Tests paths using **SCSI Test Unit Ready** commands.
- Power-cycles disks to try to clear certain errors.
- Reads or writes sectors where an I/O occurred in order to test for media errors.
- Works with higher levels to rewrite bad sectors.
- Polls disabled paths.
- ESS allows “commandless disk replacement”
  - `#mmchconfig enableAutomaticDiskReplacement=yes`

**Analysis with predictive actions to support best practice healing**
BER = Errors / Total number of bits

Disk Hospital monitors for BER spikes, in a 4K+ disk region, may indicate impending failure:
- Hospital clears the region to force the device to revector the block(s) to “good” physical space
- VIO layer asynchronously replaces the hospital pattern with reconstructed RAID data

Note: Recovery of a disk region does not by itself mean the disk needs to be replaced.
It merely indicates that data from a failed disk is being rebuilt onto spare space.
Disks will be marked for replacement if the replacement threshold has been met.
Thank you!
Accelerate with IBM Storage Technical Webinar Series - Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via Menti.com with code 2628 0279 or

Direct link https://www.menti.com/t4hii66ky8

Or

QR Code

When the webinar ends, you may see the following:

Clicking on the Continue button will take you to the survey.