Elastic Storage System (ESS)

ESS 3000 v6.0.0.2 - [Released April 2020]
Based on spectrum scale v5.0.4 PTF3 + efix2
https://www.ibm.com/support/knowledgecenter/SSZL24_6.0.0/ess3000_600_welcome.html
[Released April 2020]

Elastic Storage Server (ESS)

ESS GSxS/GLxS/GLxC/GHxy v5.3.5.2 - [Released April 2020]
Based on spectrum scale v5.0.4 PTF3 + efix2
https://www.ibm.com/support/knowledgecenter/SSYSP8_5.3.5/sts535_welcome.html

mmdiag
Current GPFS build: "5.0.4.3 efix2“

Stieg Klein
Spectrum Scale Solution Architect
IBM Washington Systems Center

Accelerate with IBM Storage Webinars

The Free IBM Storage Technical Webinar Series Continues in 2020...

Washington Systems Center – 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.

Located on the Accelerate with IBM Storage Site: https://www.ibm.com/support/pages/node/1125513

Also, check out the WSC YouTube Channel here:
https://www.youtube.com/channel/UCNuks0go01_ZrVVF1jgOD6Q

2020 Upcoming Webinars:

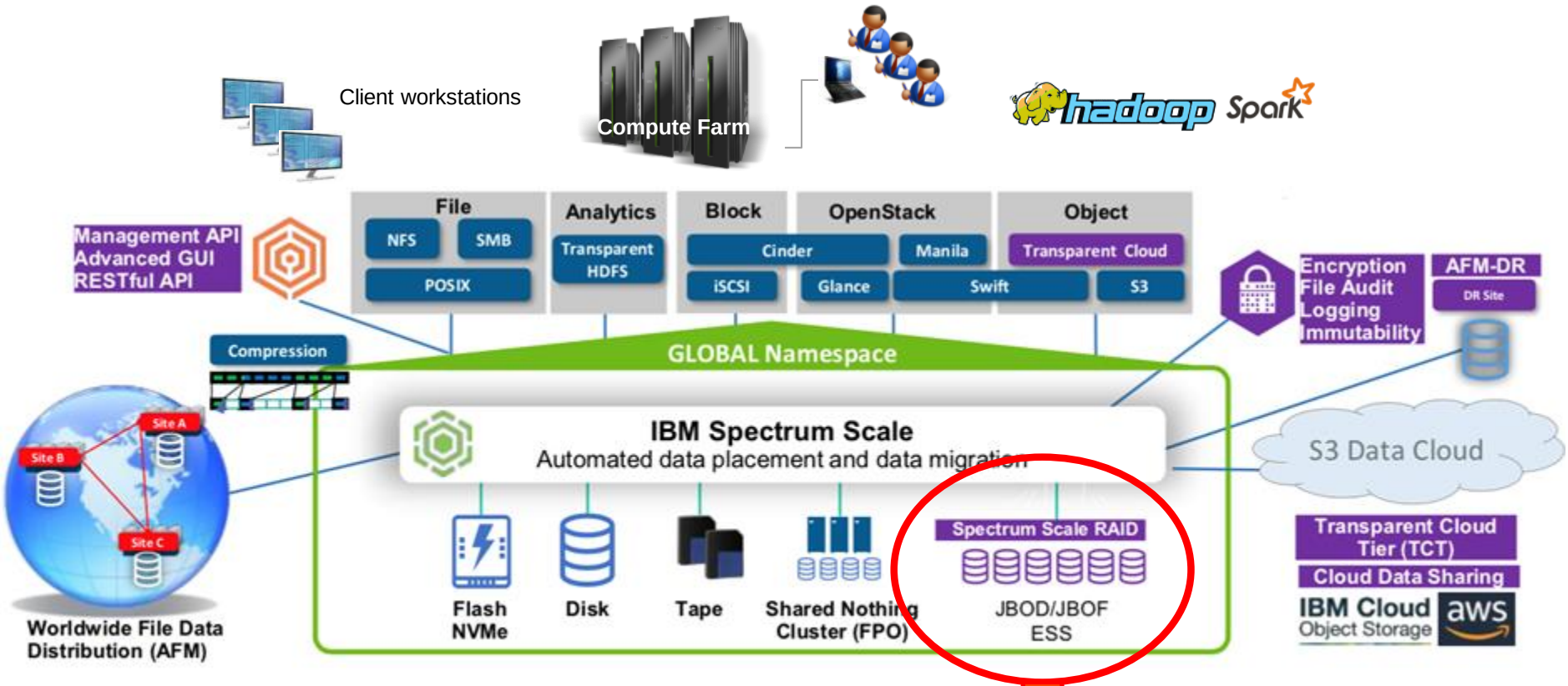
June 4 - TS7700 Systems and zOS - Two Partners Better Together!
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=efdf15a2fcf8a4582d87a6e73d3ac9544

June 9 – Spectrum Discover 2.0.3
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=e26fbf264169a0948ed0bb88685e12ce3

Agenda

- Spectrum Scale
- What's an ESS
- ESS Advantages
- Newest ESS model - ESS 3000
- Additional current ESS models [GSxS / GLxS / GLxC / GHxy]
- ESS storage concepts
- Life with an ESS

IBM Spectrum Scale - High Performance Clustered File System



- AFM Nodes for caching and distribution
- AFM-DR Nodes for non-synchronous DR
- ISKLM for Encryption Key Management
- Protocol Nodes for Object, NFS and SMB access
- Transparent Cloud Tiering (TCT) Nodes
- Hadoop Connector lives in Hadoop Cluster
- Archive via Spectrum Archive
- Native Spectrum Scale File system access

Where ESS & ECE fit in the overall solution

What is the Elastic Storage Server/System?

… Mostly focused on ESS 3000

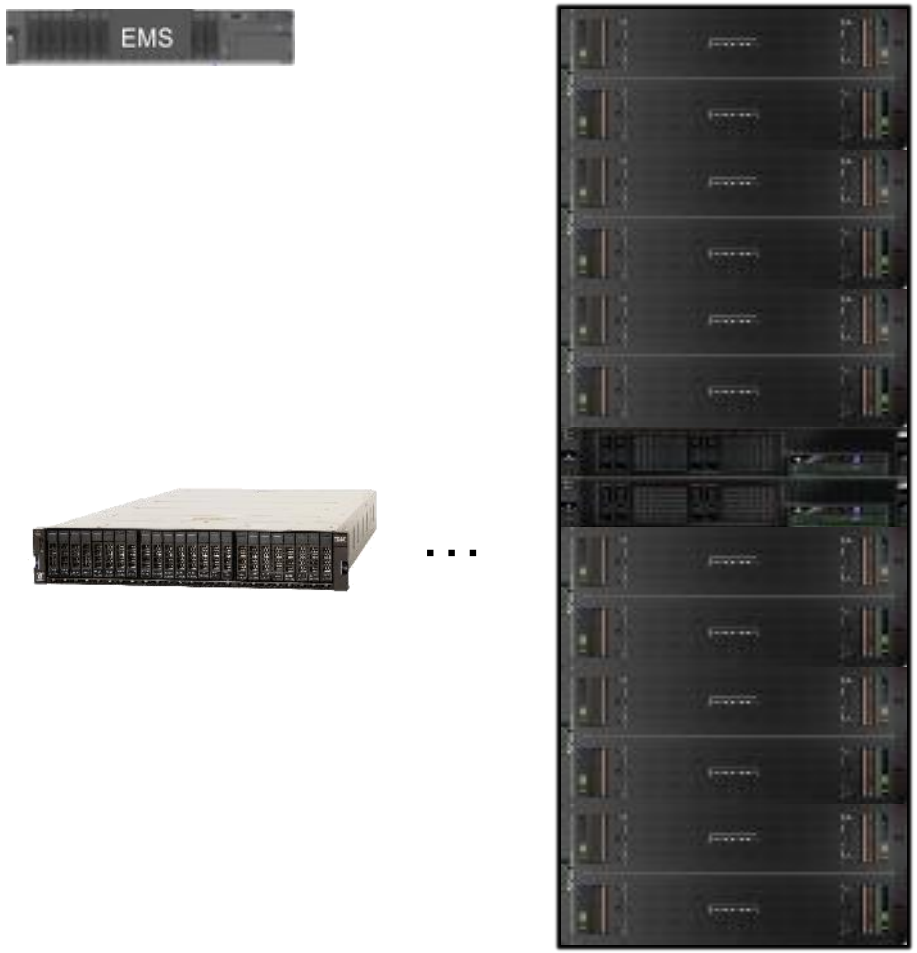
What is the Elastic Storage Server/System (ESS)?

The Elastic Storage Server (ESS) is an integrated & tested IBM provided **NSD-server building block** solution for Spectrum Scale

- Fully validated IBM hardware and software stack
- Pre-assembled, pre-configured and installed
- Spectrum Scale + Scale Native RAID + ESS GUI
- ESS aware performance/monitoring/installation/upgrade

ESS mitigates risks and makes it quicker to deploy and grow a Spectrum Scale cluster

Erasure Code Edition (ECE) is NOT an ESS



What is an ESS solution?

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

- The same EMS can manage all modern ESS models.
- The ESS GUI runs on the EMS supporting a single ESS cluster.

ESS 3000
(integrated storage and I/O nodes)

A single ESS building block consists of:

- Two NSD servers, known as I/O node
- Storage connected to both I/O nodes

ESS 3000 System includes integrated I/O nodes + NVMe storage

2U-24 external storage
(GS*S, GH*S)

Other ESS models include I/O nodes + SAS-attached storage

Pair of S822L I/O node servers
(GS*S, GL*S, GL*C)

5U-84 external storage
(GH*S, GL*S)

Multiple ESS building blocks may participate in a single Spectrum Scale cluster.

- File systems may span multiple building blocks.

4U-106 external storage
(GL*C)

Elastic Storage System - ESS 3000 – NVMe based

Leverages IBM Flashsystem 9150 system design

Peripheral Component Interconnect (PCI)
Non-Volatile Memory Host Controller Interface via PCI Express (NVMe)

2U form factor includes 2 NSD Servers & 12 or 24 NVMe drives

- 1.92/3.84/7.68/15.36TB
- 2.5-inch Small Form Factor (SFF) NVMe drives, hot swappable
- uses the Non-Volatile Memory express (NVMe) drive transport protocol

40 GB/s

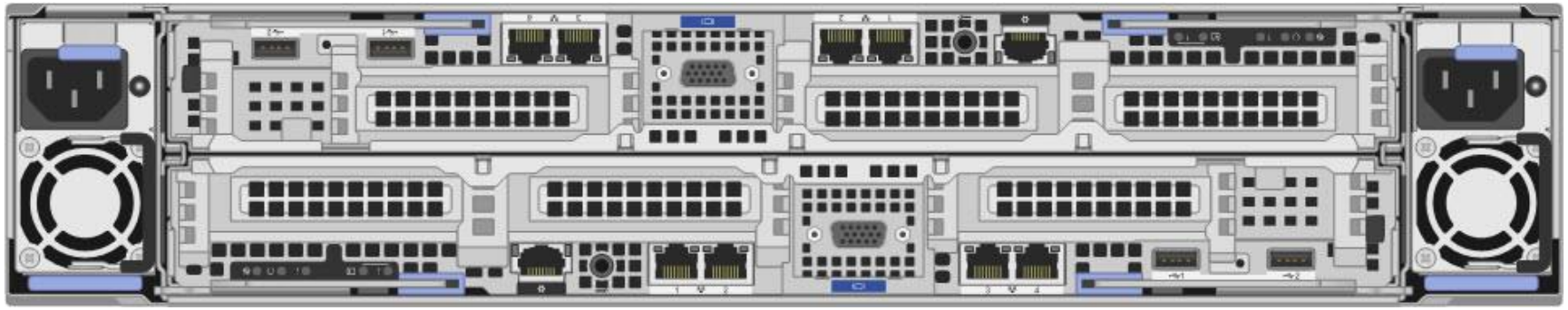
Dual-active, Containerized deployment with mirrored cache

Each NSD server supports up to 3 network adapters
100 GbE or EDR-InfiniBand

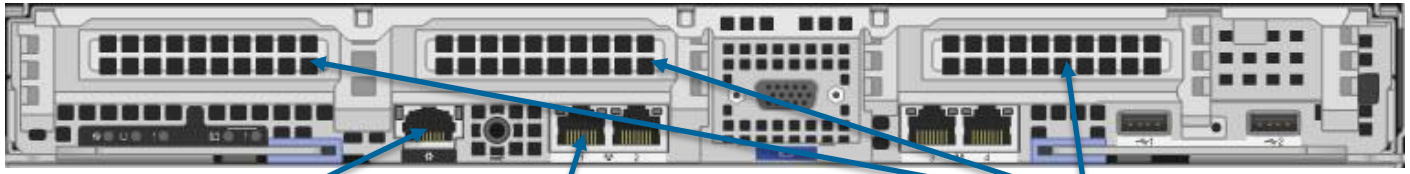
ESS 3000 Common Update location (Scale Software + Embedded RHEL)

ESS 3000 Rear view(s)

"Photo realistic"
Single High Speed Network Adapter



Rear View
Two Canisters / Servers
Two Power Supplies



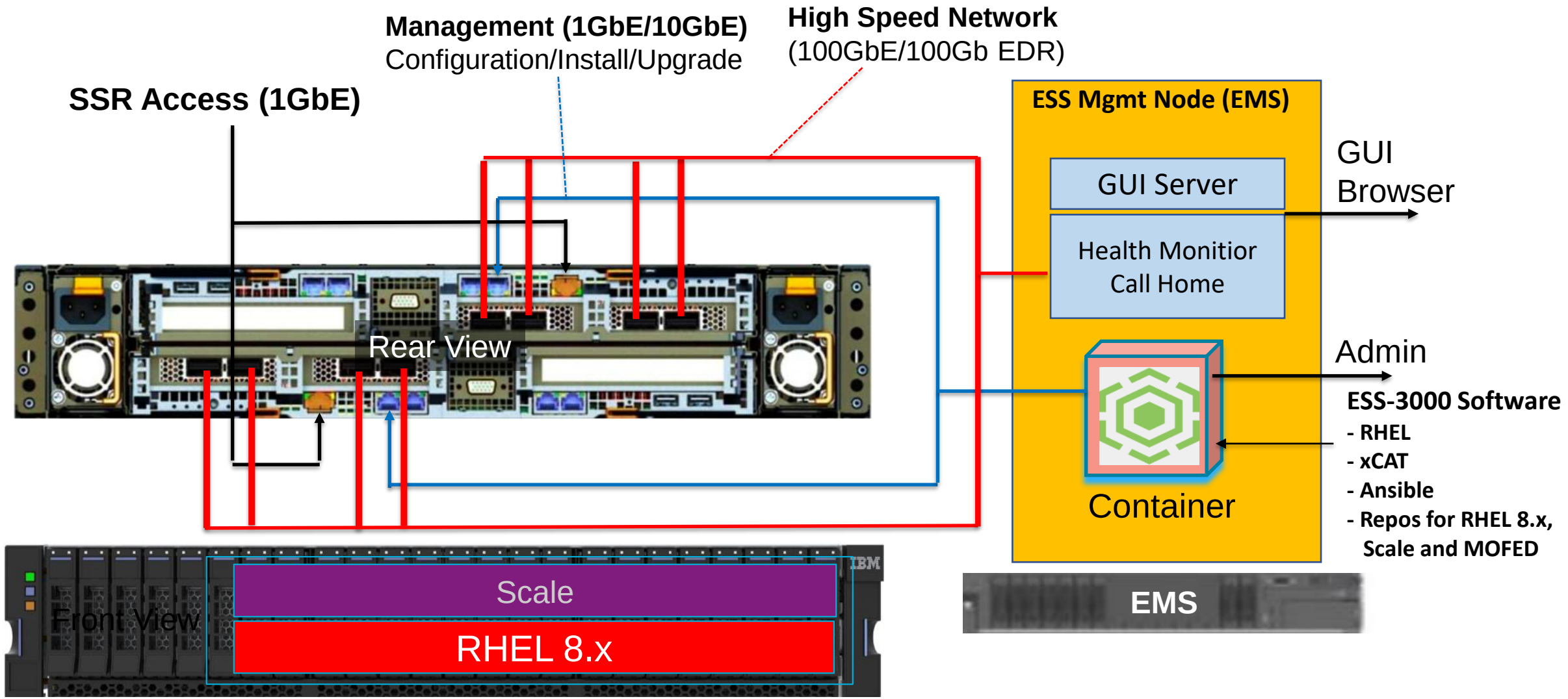
Single Canister / Server

SSR Access (1Gb)
Fixed IP from factory

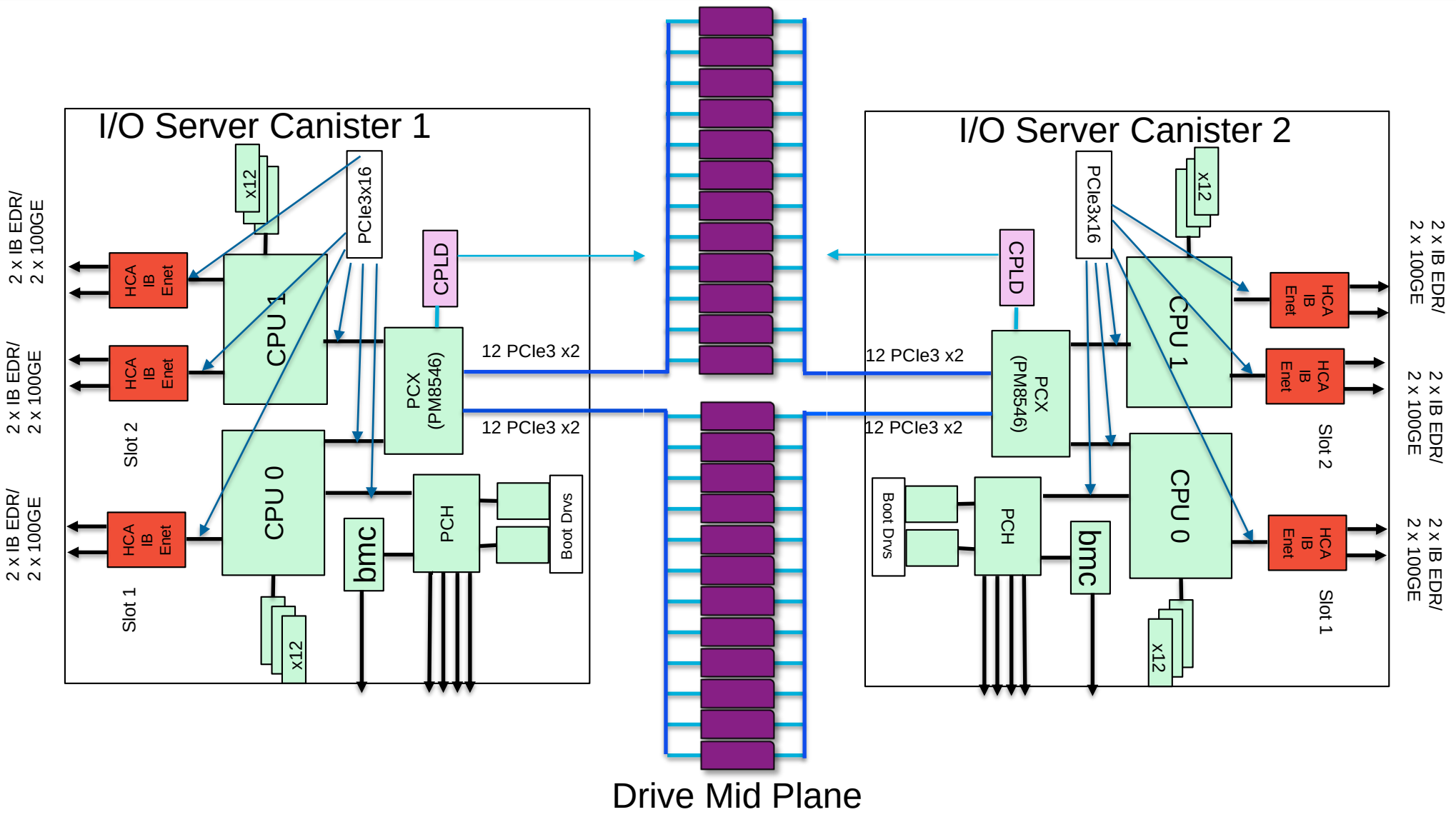
Management (1Gb/10Gb)
Install/Upgrade/Configuration

High Speed Network
(100GbE/100Gb EDR)

ESS 3000 - Networking



ESS 3000 Hardware High Level Architecture and Topology



Elastic Storage System - ESS 3000 – NVMe based performance detail

NVMe is designed specifically for flash technologies. Faster and less complicated storage drive transport protocol than SAS.

The NVMe-attached drives support multiple queues so that each CPU core can communicate directly with the drive. Avoiding latency and overhead of core-to-core communication.

ESS 3000 is a customer setup (CSU) product with a combination of customer-replaceable units (CRUs) and field-replaceable units (FRUs).

Field Replaceable Unit (FRU)	Customer Replaceable Unit (CRU)
Canister	NVMe drive
Memory DIMM	Drive Blank
Adapter	Power supply unit
M.2 boot drive	

40 GB/s

What is the Elastic Storage Server/System?

… and a survey of ESS models

Models built for speed: ESS 3000, GSxS, GHxy

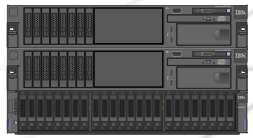
Speed ← → **Hybrid**

IBM Elastic Storage System 3000

2U24 Enclosure
12 or 24 NVMe drives

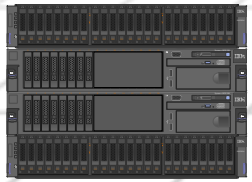
40 GB/s*
23 TB raw/13(8+2),12(8+3)
368 PB/159/236

Model GS1S
24 SSD



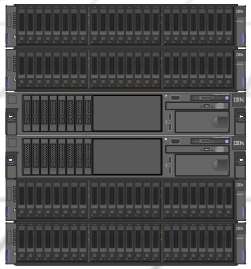
14 GB/s*
92 TB raw/56(8+2),51(8+3)
360 TB/224/205

Model GS2S
48 SSD



26 GB/s*

Model GS4S
96 SSD



40 GB/s*
368 TB raw/51(8+2),46(8+3)
1.14 PB/1/0.94

Model GH12:
1 2U24 Enclosure SSD
2 5U84 Enclosure HDD
166 NL-SAS, 24 SSD

18 GB/s*

Model GH22:
2 2U24 Enclosure SSD
2 5U84 Enclosure HDD
166 NL-SAS, 48 SSD

20 GB/s*

Model GH14:
1 2U24 Enclosure SSD
4 5U84 Enclosure HDD
334 NL-SAS, 24 SSD

38 GB/s*

Model GH24:
2 2U24 Enclosure SSD
4 5U84 Enclosure HDD
334 NL-SAS, 48 SSD

40 GB/s*

* Estimate of performance aggregated across SSD and HDD. All estimates assume EDR Infiniband connections, 100% read performance IOR sequential. Use IBM FOS DE tool to estimate for your network + workload

Models built for high capacity: GLxS

Capacity

Model GL1S:
1 Enclosures, 9U
82 NL-SAS, 2 SSD

6 GB/s*

Model GL2S:
2 Enclosures, 14U
166 NL-SAS, 2 SSD

12 GB/s*

Model GL3S:
3 Enclosures, 19U
250 NL-SAS, 2 SSD

18 GB/s*

Model GL4S:
4 Enclosures, 24U
334 NL-SAS, 2 SSD

24 GB/s*

Model GL5S:
5 Enclosures, 29U
418 NL-SAS, 2 SSD

30 GB/s*

Model GL6S:
6 Enclosures, 34U
502 NL-SAS, 2 SSD

36 GB/s*

* Estimate of performance aggregated across SSD and HDD. All estimates assume EDR InfiniBand connections, 100% read performance. Use IBM FOS DE tool to estimate for your network + workload

Models built for extreme high capacity: GLxC

Capacity

Model GL1C
1 Enclosure, 8U
104 NL-SAS, 2 SSD

1.46 PB raw
1 (8+2P)
0.93 (8+3P)

Model GL2C
2 Enclosures, 12U
210 NL-SAS, 2 SSD

2.9 PB

Model GL3C
2 Enclosures, 12U
316 NL-SAS, 2 SSD

4.4 PB

Model GL4C
4 Enclosures, 16U
432 NL-SAS, 2 SSD

5.9 PB

Model GL5C
5 Enclosures, 28U
528 NL-SAS, 2 SSD

7.3 PB

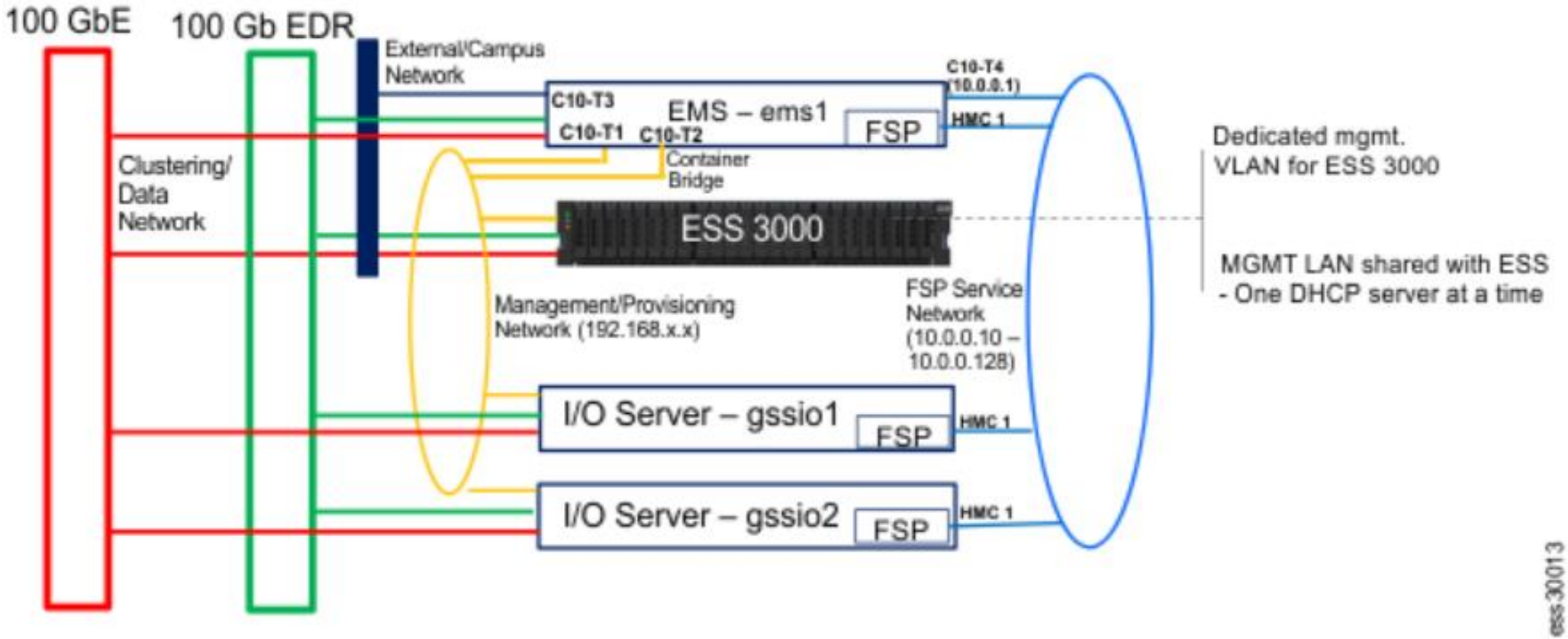
Model GL6C
6 Enclosures, 28U
634 NL-SAS, 2 SSD

8.8 PB

Model GL8C
8 Enclosures, 36U
846 NL-SAS, 2 SSD

11.8 PB raw
8.5 (8+2P)
7.5 (8+3P)

ESS - Networking



IBM GPFS/Spectrum Scale Native RAID Model Timeline

2011 GPFS Native Raid on IBM Power 775 Supercomputer

2013 GPFS Storage Server (GSS) v1.0 on IBM x3650 M4

2014 Elastic Storage Server (ESS) v4.1 on IBM P8 S822L servers
P8 PPC64BE + Hardware Management Console (HMC)
Release GLx Models: GL2/GL4/GL6 + DCS3700 storage
Release GSx Models: GS1/GS2/GS4 + (2U24)EXP24S storage 3.84/15.36TB 2.5” SSDs
Networking: 10/40 GB Ethernet, 40 GB Infiband

2015 New model GS6
MES Upgrade GL2->GL4->GL6 & GS1->GS2->GS4->GS6
Add 100 GB EDR Infiniband

2017 PPC64LE + Advanced System Management Interface (ASMI) in Firmware
Release GSxS Models: GS1S/GS2S/GS4S + (2U24) EXP24S storage
Release GLxS Models: GL2S/GL4S/GL6S + (5U84) storage. 4/8/10TB NL-SAS 3.5” HDDs
Networking: 10/40/**100** GB Ethernet, 56 FDR Infiband/100 EDR GB Infiband

2018 Summit System Operational at Oakridge National Laboratory
Release Mini Coral GL1C/GL2C/GL4C/GL6C (4U106)
Release Hybrid models: GH14/GH24
new Models: GL1S/GL3S & GL4S/GL6S
Upgrade GS1S->GS2S->GS4S & GL1S->GL2S->GL3S->GL4S->GL6S

2019 – ESS 3000 (NVMe based) 1.92/3.84/7.68/15.36 NVME 2.5” flash drives Either 12 or 24
New models GH22/GH24 & GL5S & GL3C/GL5C/GL8C
Upgrade GL1C->GL2C->…->GL5C->GL6C & GL1S->GL2S->…->GL5S->GL6S
(5U84) storage: 4/8/10/**14**TB NL-SAS 3.5”HDDs

2020 – Add PB based licensing Fun fact: TB is 2^{40} bytes PB is 2^{50} bytes

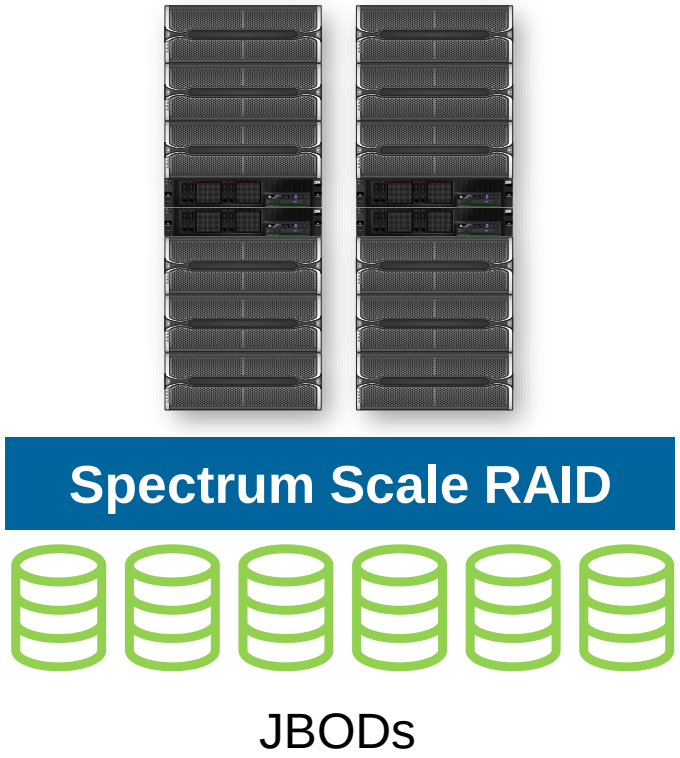
Spectrum Scale RAID

… the special sauce in the Elastic Storage Server

Declustered software RAID

IBM **Spectrum Scale RAID** is a *software* implementation of “declustered” or “**distributed RAID**”:

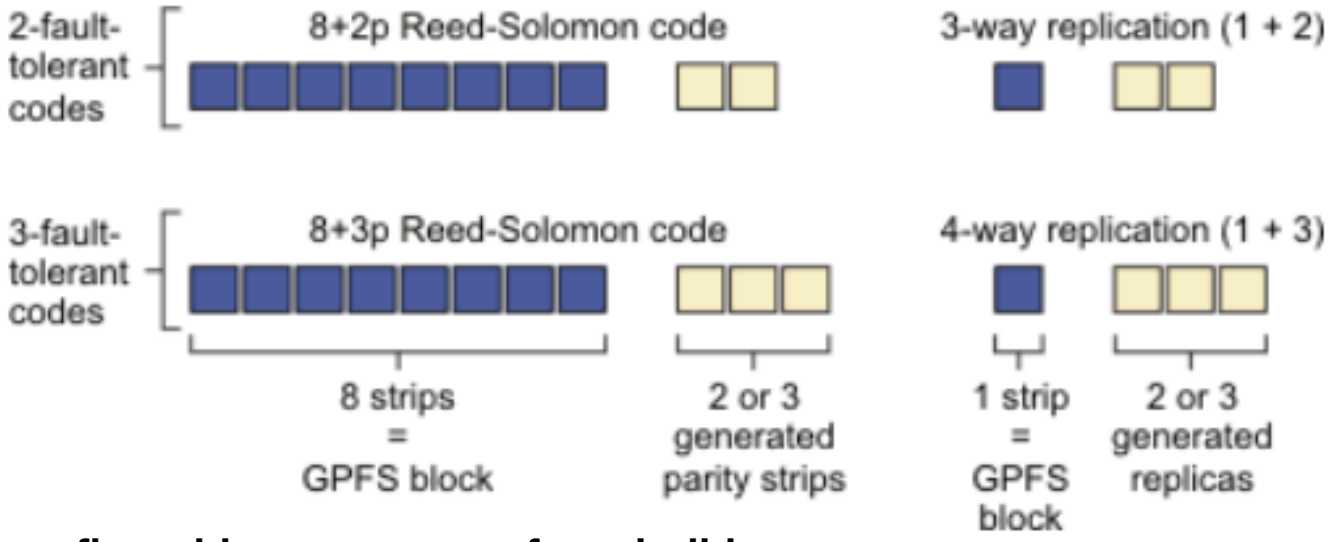
- Extremely fast rebuild after a disk failure, with minimal impact on performance
- Very strong data integrity checks
- Additional erasure codes, such as 8+3p
- Error detection codes enable detecting track errors and dropped writes
- Consistent performance from 0 – 99% utilization or 1 to many jobs in parallel



Spectrum Scale RAID is currently available only with Elastic Storage Server (IBM’s reference architecture) and Erasure Code Edition.

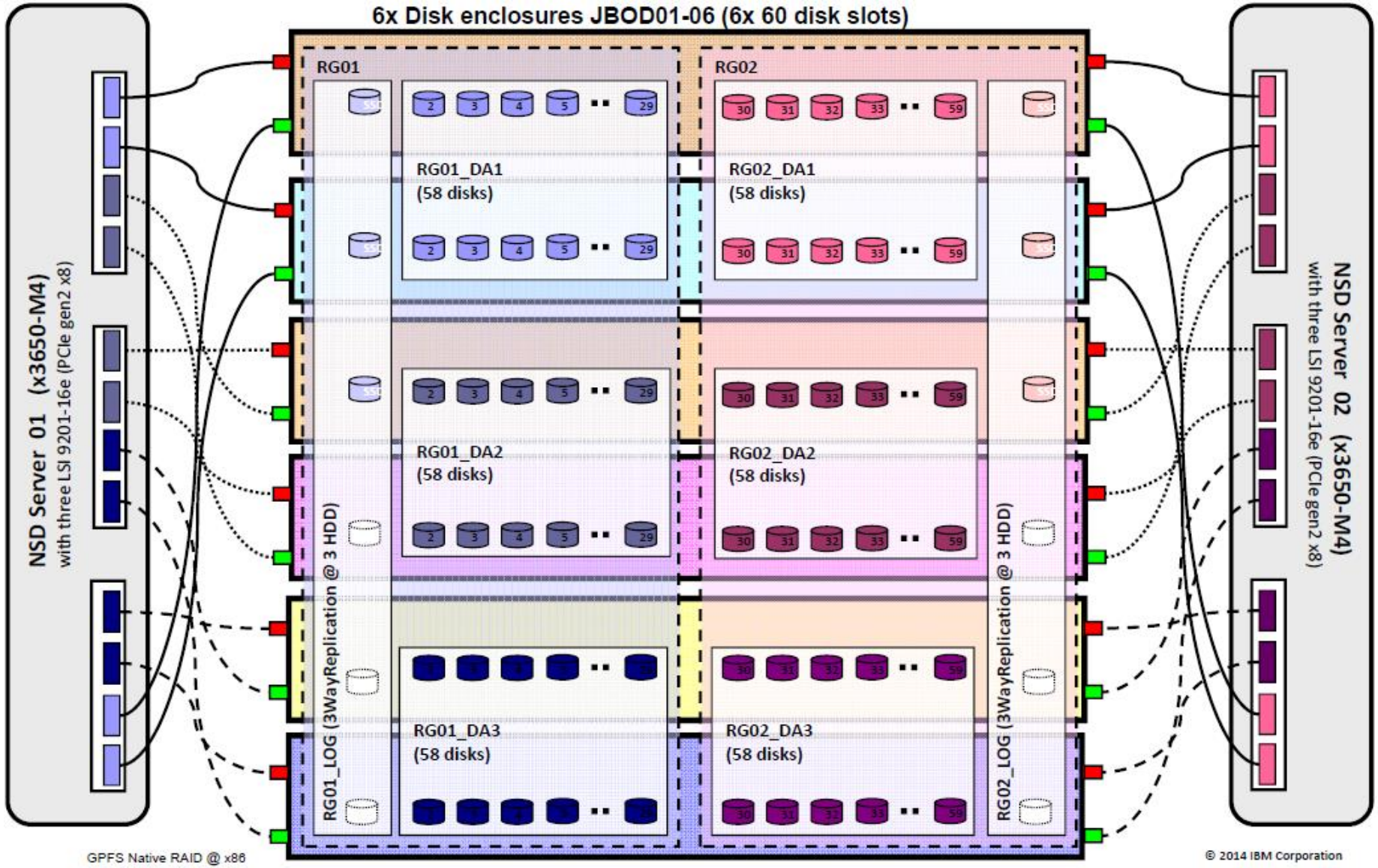
Spectrum Scale RAID erasure codes

- **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**

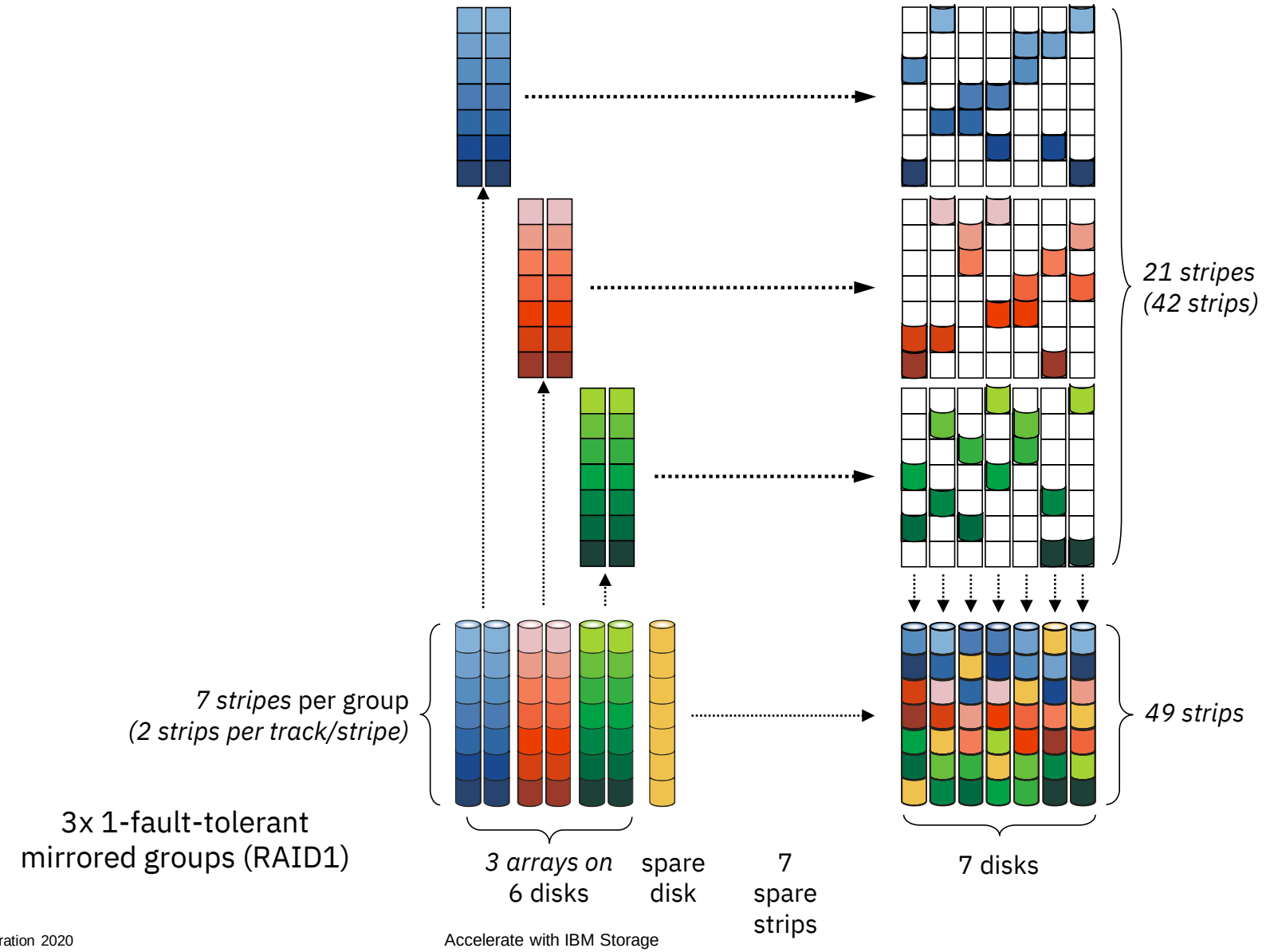


- ***Excluding user-configurable spare space for rebuilds**

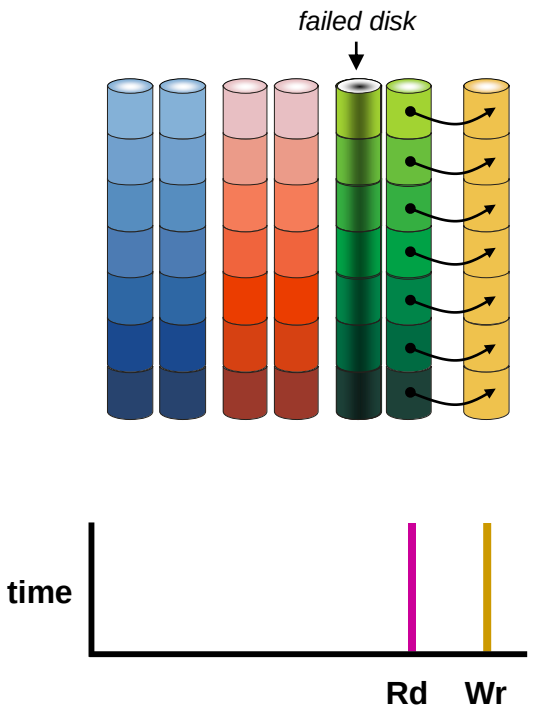
Native RAID Layout example from 2014



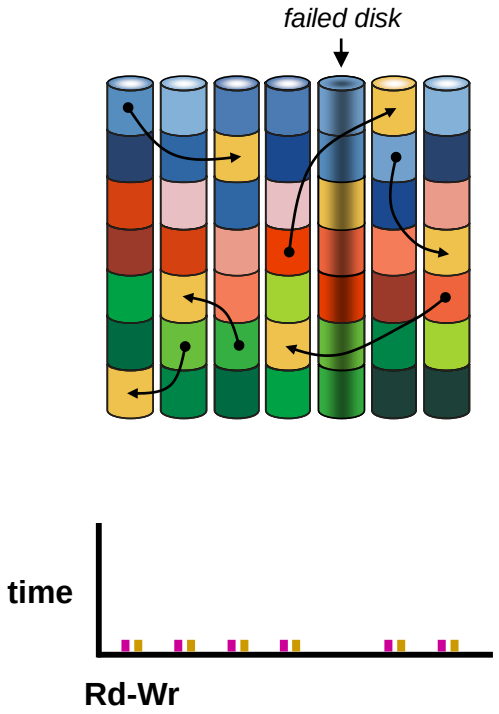
Declustered RAID Example



Rebuild Overhead Reduction Example



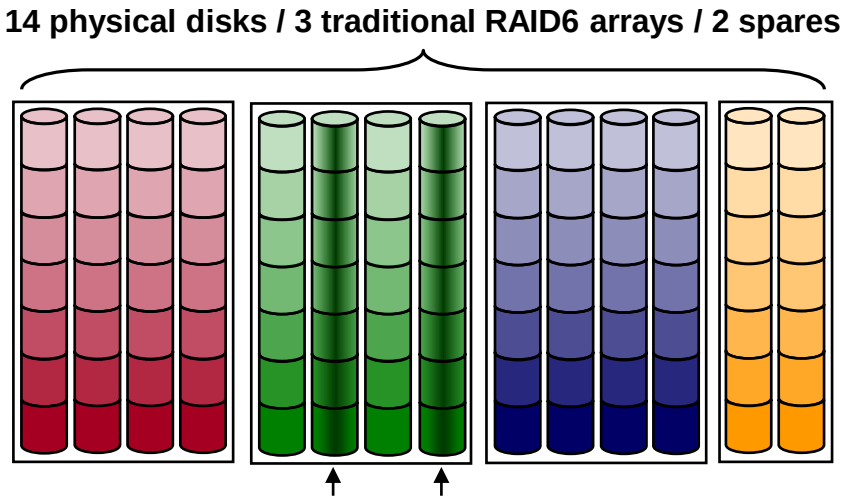
Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs



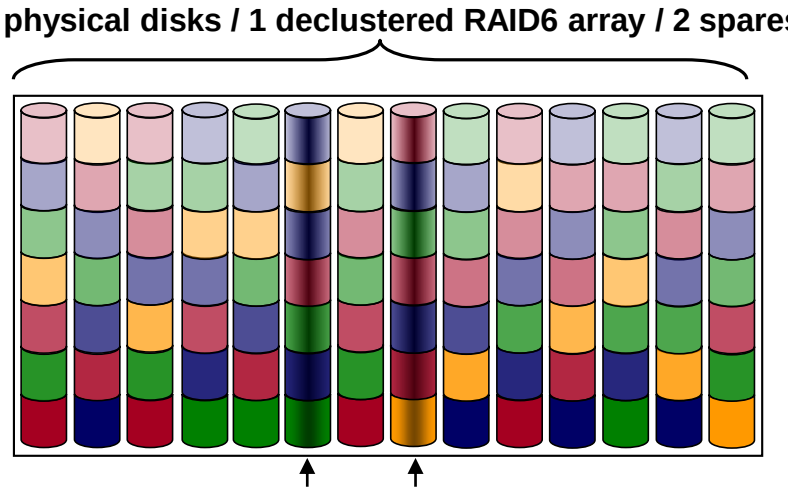
Rebuild activity spread across many disks, less disruption to user programs

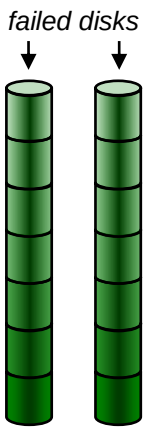
Rebuild overhead reduced by 3.5x

Declustered RAID6 Example



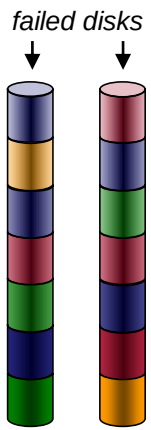
Decluster data, parity and spare





Number of faults per stripe		
Red	Green	Blue
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0

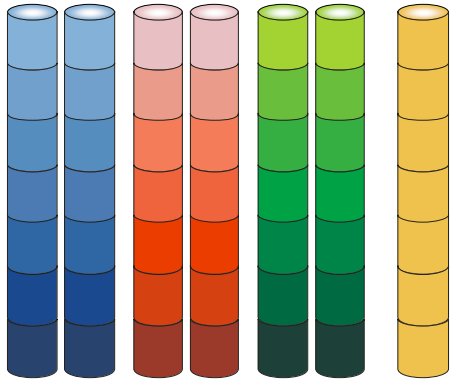
Number of stripes with 2 faults = 7



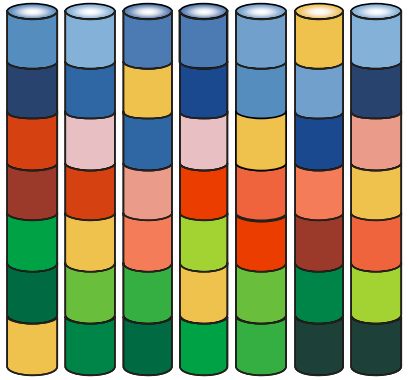
Number of faults per stripe		
Red	Green	Blue
1	0	1
0	0	1
0	1	1
2	0	0
0	1	1
1	0	1
0	1	0

Number of stripes with 2 faults = 1

Benefits of declustering in Spectrum Scale RAID



Conventional



De-clustered

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

- When one disk is down (most common case)
  - – Rebuild slowly 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

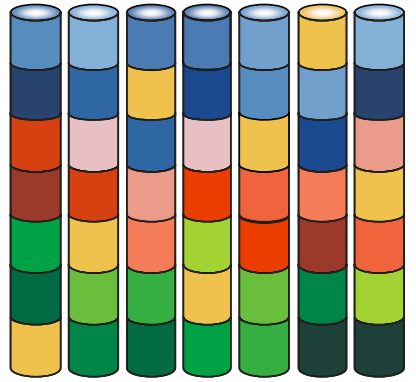
Data integrity manager

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)



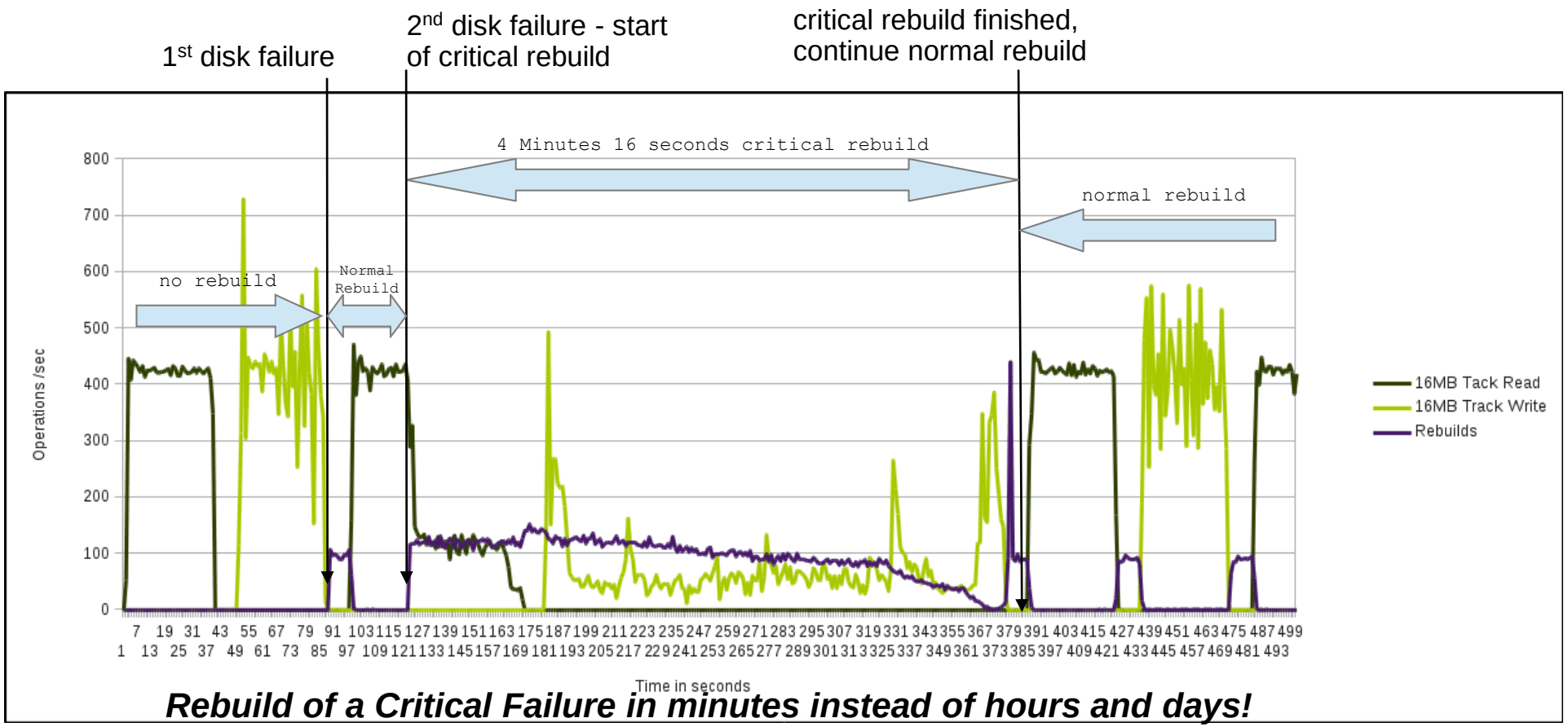
Medium priority: Rebalance spare space after disk install

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



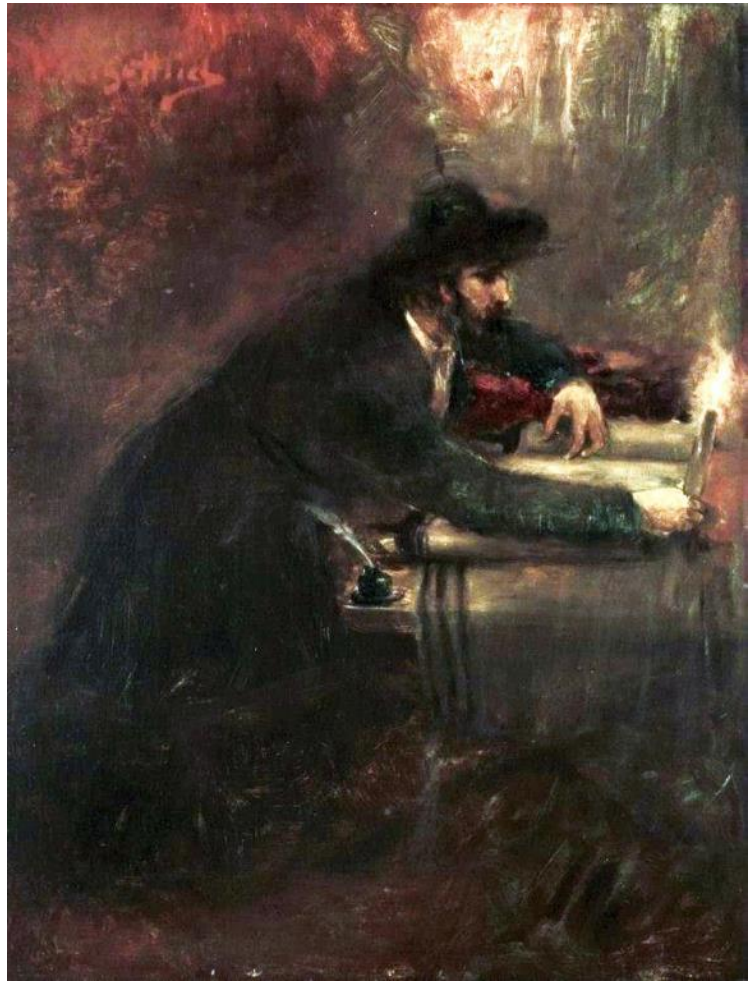
Rebuild of a Critical Failure in minutes instead of hours and days!

Spectrum Scale RAID

Checksums

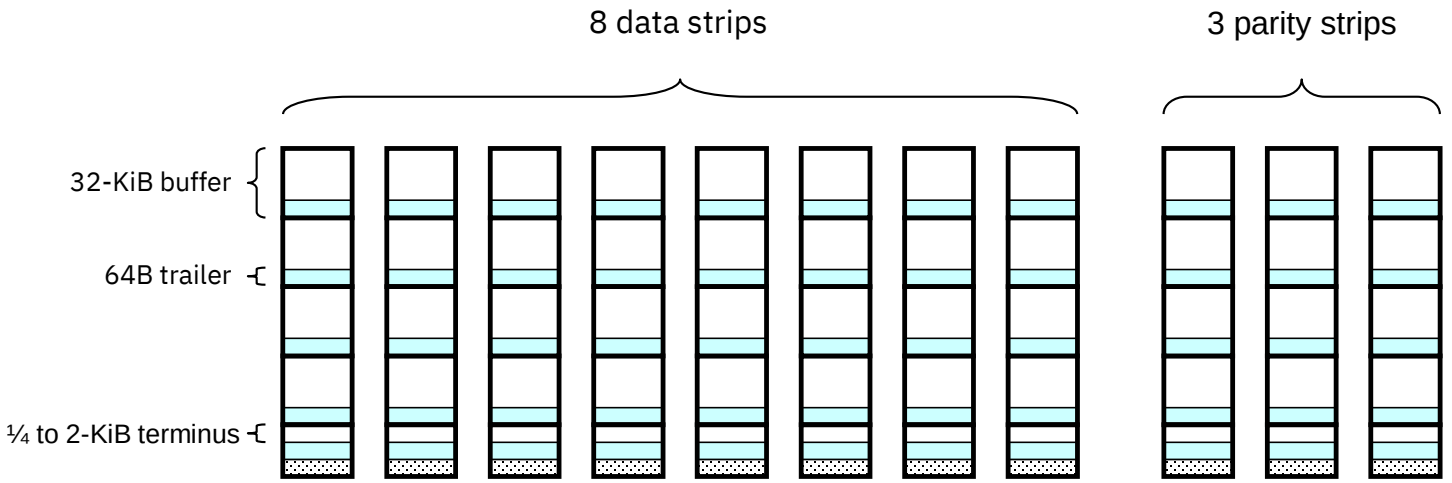
ESS – Data Integrity Enhancements

- 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.



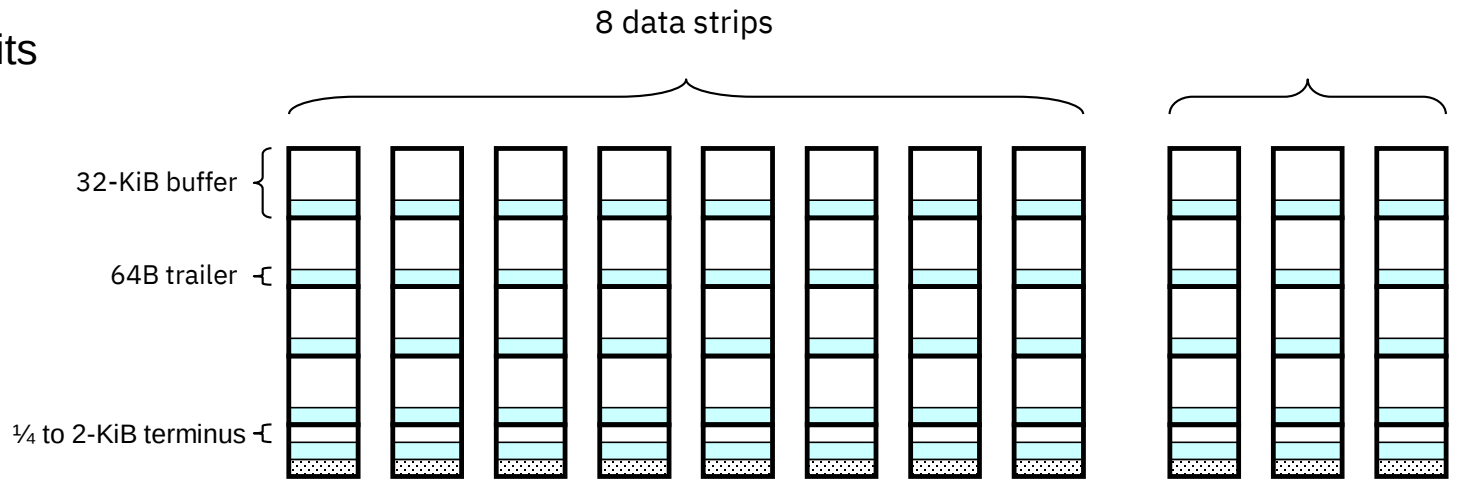
End to End Checksum (Cont)

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

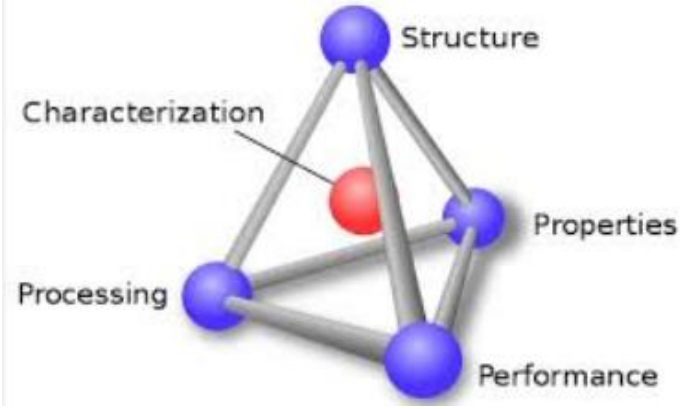
- Uses SES interface for lights, latch locks, disk power, and so on.

Manages topology and hardware configuration.

Disk Hospital Operations

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.



Analysis with predictive actions to support best practice healing (almost like a real hospital)

Thank you!

Accelerate with IBM Storage Survey

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

You can access this 5 question survey via Menti.com with code 78 81 27 or

Direct link https://www.menti.com/mkg7a2x6q8

Or

QR Code