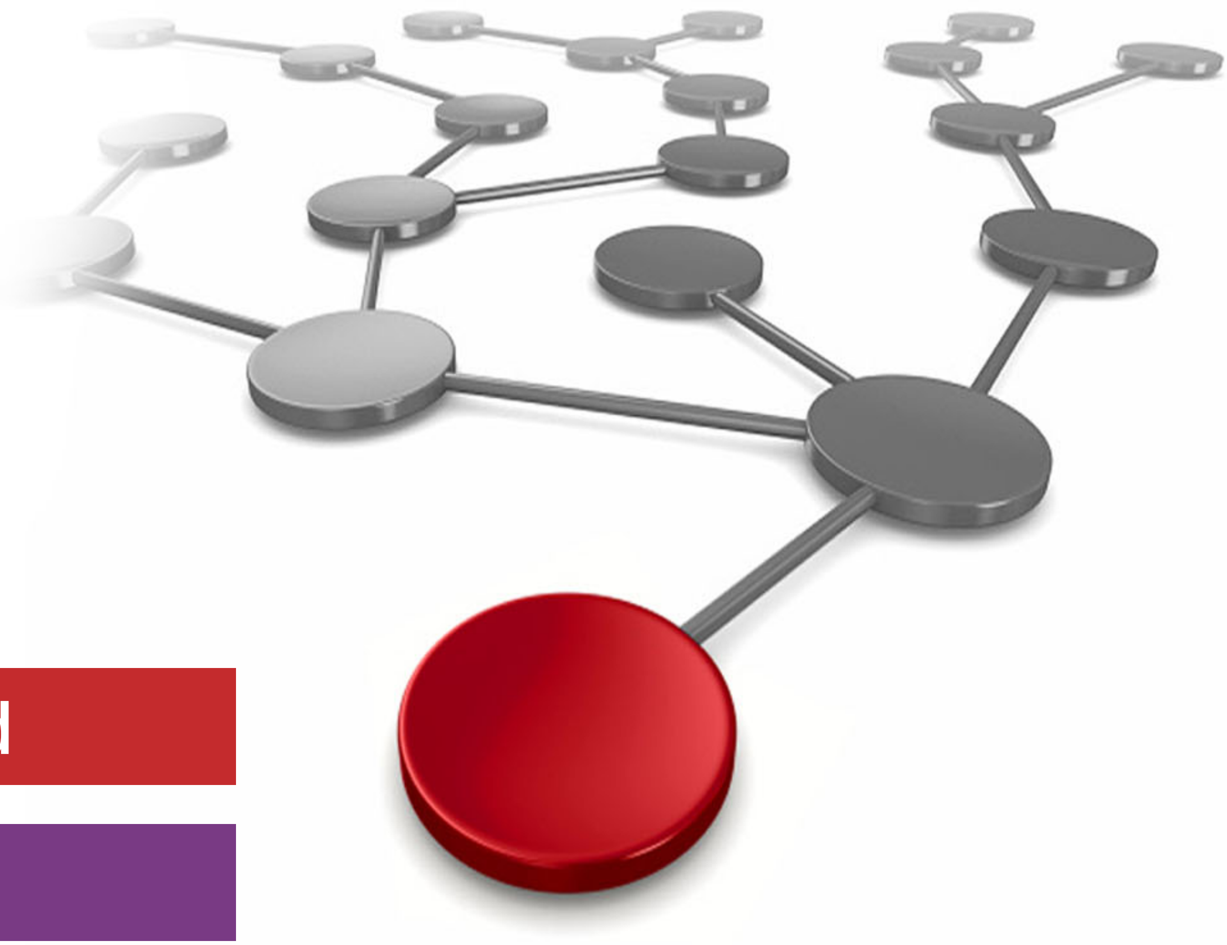
Redbooks
ibm.com/redbooks

Implementation Guide for IBM Elastic Storage System 3200

Christopher Bostic
Farida Yaragatti
Isaiah Eaton
Jay Vaddi
Joe Sanjour
John Lewars
John Sing
Larry Coyne
Leteshia A. Lowe
Luis Bolinches
Meagan Miller
Olaf Weiser
Pidad Gasfar D'Souza
Puneet Chaudhary
Ravindra Sure
Robert Guthrie
Sumit Kumar
Wesley Jones



Cloud

Storage

IBM

Redbooks

IBM Redbooks

Implementation Guide for IBM Elastic Storage System 3200

November 2021

SG24-8516-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (November 2021)

This edition applies to IBM Elastic Storage Server 3200.

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION 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. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

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

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

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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 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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered 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 http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) ®
AIX®
IBM®
IBM Cloud®
IBM Elastic Storage®
IBM FlashSystem®
IBM Spectrum®
POWER®
POWER7®
POWER8®
POWER9™
Redbooks®

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Ansible, Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

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

Preface

This IBM® Redbooks® publication introduces and describes the IBM Elastic Storage® Server 3200 (ESS 3200) as a scalable, high-performance data and file management solution. The solution is built on proven IBM Spectrum® Scale technology, formerly IBM General Parallel File System (IBM GPFS).

IBM Elastic Storage System 3200 is an all-Flash array platform. This storage platform uses NVMe-attached drives in ESS 3200 to provide significant performance improvements as compared to SAS-attached flash drives.

This book provides a technical overview of the ESS 3200 solution and helps you to plan the installation of the environment. We also explain the use cases where we believe it fits best.

Our goal is to position this book as the starting point document for customers that would use the ESS 3200 as part of their IBM Spectrum Scale setups.

This book is targeted toward technical professionals (consultants, technical support staff, IT Architects, and IT Specialists) who are responsible for delivering cost-effective storage solutions with ESS 3200.

Authors

This book was produced by a team of specialists from around the world working at IBM Redbooks, Tucson Center.

Christopher Bostic is a Software Engineer in Austin, Texas. Currently he works with the IBM Spectrum Scale development team on storage enclosures, drive, and host adapter firmware. He joined IBM in 2001 and worked on i, p and z series systems as a firmware developer. He is a device-driver specialist and contributed to development of the open-source Linux kernel.

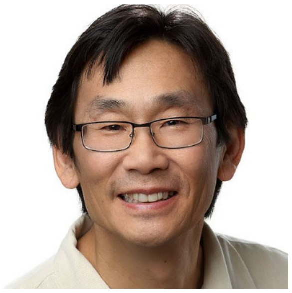
Farida Yaragatti is a Senior Software Engineer at IBM India. She has a BE, Electronics and Communication from Karnataka University, India and has 12 years of experience in the Software Testing field. She is part of manual and automation testing for IBM Spectrum Scale and IBM Elastic Storage System (ESS) deployment as a Senior Tester. Farida worked at IBM for over five years and previously held roles within the IBM Platform Computing and IBM Smart analytics system (ISAS) testing teams. She has strong engineering professional skills in Software deployment testing, including automation using various scripting technologies, such as Python, shell scripting, Robot framework, Ansible, and Linux.

Isaiah Eaton is a Senior Managing Consultant in Denver, Colorado. He works with the IBM Systems Lab Services Storage team on IBM Spectrum Scale and IBM ESS implementations and integration configurations. He joined IBM in 2008 in an advisory role for consulting based on the results of the monitoring and data collection tools for storage optimization. He is a certified Expert Consultant software and architect with extensive expertise in Enterprise systems architectures. He has a lead role as ESS team member for North America. He has worked on IBM Spectrum Scale and ESS storage products since 2013, including IBM Spectrum Archive, Active File Management, and protocol node integration

Jay Vaddi is a Storage Performance Engineer at IBM Tucson, AZ. He has been with IBM and the performance team for over five years. His focus is primarily on performance analysis and evaluations of IBM Spectrum Scale and IBM ESS products.

Joe Sanjour is a Managing Consultant with the IBM Systems Lab Services Storage team. He is currently working on IBM Spectrum Scale and ESS implementations, upgrades, and migrations. Joseph has been with IBM since 2008, performing Storage Infrastructure Optimization workshops, supporting the Federal Storage Sales Team, and is now with Systems Lab Services. With Systems Lab Services, he was the technical lead for both the SONAS and Storwize V7000 Unified products (based on GPFS). He has been working with IBM Spectrum Scale since 2010 and with ESS since 2015. He has performed several migrations from 3rd party NAS systems to IBM Spectrum Scale / ESS systems using Active File Management.

John Lewars is a Senior Technical Staff Member who leads performance engineering work in the IBM Spectrum Scale development team. He has been with IBM for over 20 years, working first on some of IBM's largest high-performance computing systems, and later on the IBM Spectrum Scale (formerly GPFS) development team. John's work on the IBM Spectrum Scale team includes working with large customer deployments and improving network resiliency, along with co-leading development of the team's first public cloud and container-support deliverables.



John Sing is Offering Evangelist for IBM Spectrum Scale / ESS. In his over 25 years with IBM, John has been a world-recognized IBM speaker, author, and strategist in enterprise storage, file and object storage, internet scale workloads and data center design, big data, cloud, it strategy planning, high availability (HA), business continuity, and disaster recovery (DR). He has spoken at over 40 IBM conferences worldwide, and is the author of eight IBM Redbooks publications and nine IBM Redpaper publications.

Larry Coyne is a Project Leader at the International Technical Support Organization, Tucson Arizona center. He has 35+ years of IBM experience with 23 in IBM storage software management. He holds degrees in Software Engineering from the University of Texas at El Paso and Project Management from George Washington University. His areas of expertise include customer-relationship management, quality assurance, development management, and support management for IBM Storage Software.

Leteshia A. Lowe is a Senior Test and Quality Assurance engineer at IBM, where she is responsible for testing of the IBM FlashSystem® product family, an all-flash enterprise storage platform. As a technical test specialist, she provides debug and failure investigation, strategic test management and innovative Reliability, Availability and Serviceability (RAS) testing solutions. Prior to joining the IBM FlashSystem team in Houston, TX, Leteshia worked in Tucson, AZ at IBM as a Tape Storage Development Engineer, programming for the TS7700 Storage Virtualization Engine product and later as a Test Engineer. Leteshia has been with IBM for 17 years and holds a Bachelor of Science degree in Electrical Engineering specializing in Computer Engineering from Tennessee State University and a Master of Education degree in Secondary Education from the University of Arizona.

Luis Bolinches has worked with IBM Power Systems servers for over 15 years, and has been with IBM Spectrum Scale for over 10 years. He works 20% for IBM Systems Lab Services in the Nordic region, and the other 80% as part of the IBM Spectrum Scale development team.

Meagan Miller is a Quality Assurance test lead in Houston, TX for IBM FlashSystem focusing on hardware testing. Meagan has seven years of experience working on both hardware and software quality after starting their first few years as a technical writer. Meagan holds an undergraduate degree in English Literature and Language from Southeastern Louisiana University and a graduate degree in Library and Information Science from Louisiana State University.

Olaf Weiser joined IBM as an experienced professional more than ten years ago and worked in the DACH TSS team delivering Power-based solutions to enterprise and HPC customers. He developed deep skills in IBM Spectrum Scale (previously IBM GPFS) and has a worldwide reputation as the performance-optimization specialist for IBM GPFS. At the IBM European Storage Competence Center (ESCC), Olaf works on Advanced Technical Support (ATS) and Lab Services and Skill Enablement tasks that are required to grow the IBM Spectrum Scale business in EMEA. For the past two years, he worked as a performance engineer for RDMA and RoCE in IBM Research and Development GmbH.

Pidad Gasfar D'Souza is a System Architect who specializes in performance engineering of IBM Spectrum Scale and IBM POWER® systems. He has been with IBM for more than 17 years. He led the system-performance engineering of the GPU-accelerated systems designed for applications in the fields of AI and HPC. He also led development teams in the areas of IBM AIX® base-libraries and JVM. He has presented extensively at several international conferences, and delivered customer workshops and lab sessions.

Puneet Chaudhary is a Technical Solutions Architect who works with the IBM ESS and IBM Spectrum Scale solutions. He has worked with IBM GPFS, now IBM Spectrum Scale, for many years.

Ravindra Sure works for IBM India as a Senior System Software Engineer. He worked on developing workload schedulers for High Performance Computers, Parallel File Systems, Computing Cluster Network Management, and Parallel Programming. He has strong engineering professional skills in distributed systems, parallel computing, C, C++, Python, shell scripting, MPI, and Linux.

Robert Guthrie is a Senior Software Engineer in Austin, Texas. He works with the IBM Spectrum Scale development team on storage enclosures and NVMe. He joined IBM in 1996 and worked on CORBA-based enterprise management software. He is a software and systems-solutions specialist with extensive expertise in networks, firmware, and middleware. He has had lead roles providing superior levels of design, development, test, and system integration functions for multiple projects serving large, international financial, insurance, and government-enterprise clients. He has worked on storage products since 2008, including Information Archive, IBM Spectrum Protect, and IBM ESS.

Sumit Kumar is an Advisory Software Engineer at IBM India. He has a Master in Computer Application degree from IGNOU, New Delhi, India, and has 16 years of experience in the software-development field. He has been part of IBM ESS deployment code development for IBM POWER8® and POWER9™ systems, including x86 servers. He also worked on IBM Spectrum Scale as a deployment developer. Sumit has worked within IBM for over seven years, and previously held roles within the IBM Platform Computing and IBM Systems Director team. He has strong engineering-professional skills in Software deployment and automation, which use various scripting technologies, such as Python, shell scripting, Ansible, and Linux.

Wesley Jones serves as the test-team lead for IBM Spectrum Scale Native RAID. He also serves as one of the principle deployment architects for IBM ESS. His focus areas include IBM Power servers, IBM Spectrum Scale (GPFS), cluster software (xCAT), Red Hat Linux, Networking (especially InfiniBand and Gigabit Ethernet), storage solutions, automation, and Python.

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

- Chiahong Chen, Ricardo Zamora Ruvalcaba
  IBM Storage Systems
- Teena Pareek, Amrita Chatterjee, Monika Thakral
  Altran

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks
- Send your comments in an email to:

  redbooks@us.ibm.com
- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099

2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction

This chapter introduces the IBM Elastic Storage System 3200 (ESS 3200) solution, the software characteristics of IBM Spectrum Scale RAID (Redundant Array of Independent Disks) software that runs on ESS 3200, and provides an overview of the ESS 3200.

This chapter includes the following sections:

- 1.1, “IBM Spectrum Scale RAID ” on page 2
- 1.2, “IBM Elastic Storage System (ESS)” on page 4
- 1.3, “IBM Elastic Storage System 3200” on page 5
- 1.4, “License considerations” on page 7

ESS 3200 is a high-performance, NVMe flash-storage member of the IBM Spectrum Scale and Elastic Storage System family storage solutions for high performance, high-scalability Data and AI applications. For an overview of how ESS 3200 fits into this overall family, see the companion *IBM Redpaper Introduction Guide to the Elastic Storage System*, REDP5253.

1.1 IBM Spectrum Scale RAID

The following section provides a high-level technical overview of the IBM Spectrum Scale RAID that is used in all ESS models including ESS 3200. IBM Spectrum Scale RAID on ESS 3200 provides significant storage cost-reduction while simultaneously providing enterprise-class reliability, performance, and serviceability.

The IBM Spectrum Scale RAID software in ESS 3200 uses local NVMe drives. Because RAID functions are handled by the software, ESS 3200 does not require an external RAID controller or acceleration hardware.

IBM Spectrum Scale RAID in ESS 3200 supports two and three fault-tolerant RAID codes. The two-fault tolerant codes include 8-data plus 2-parity, 4-data plus 2-parity, and 3-way replication. The three-fault tolerant codes include 8-data plus 3-parity, 4-data plus 3-parity, and 4-way replication. Figure 1-1 shows example RAID tracks consisting of data and parity strips.

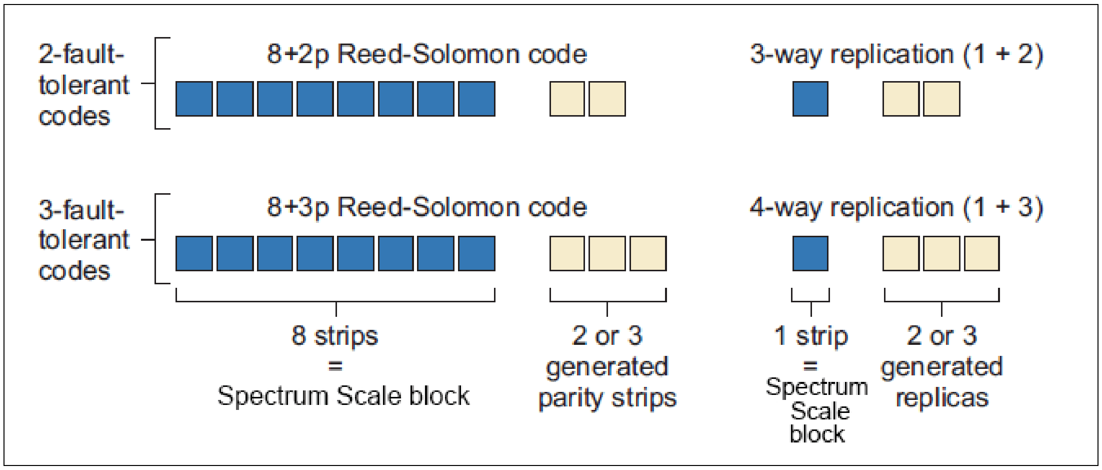

2-fault-tolerant codes
8+2p Reed-Solomon code
3-way replication (1 + 2)
3-fault-tolerant codes
8+3p Reed-Solomon code
4-way replication (1 + 3)
8 strips = Spectrum Scale block
2 or 3 generated parity strips
1 strip = Spectrum Scale block
2 or 3 generated replicas

Figure 1-1 RAID tracks

1.1.1 Product history

In 2003, the Defense Advanced Research Project Agency (DARPA) started their High-Productivity Computing Systems (HPCS) program: Productive, Easy-to-use, Reliable Computing System (PERCS). IBM's proposal for DARPA's HPCS project is what today has become IBM Spectrum Scale RAID.

In 2007, IBM released the first market product based on IBM Spectrum Scale RAID, the P7IH. The system is based on the IBM POWER7® system and SAS disks, delivering tens of gigabytes per second of storage throughput already in 2007.

While P7IH was, and still is, a fantastic engineering machine, in 2012 IBM released the GSS platform that was running what is known today as IBM Spectrum Scale RAID but on commodity hardware.

In 2014, IBM superseded the GSS with the first ESS, based on the IBM POWER8 system but using commercially available servers and disk enclosures while still based on the same IBM Spectrum Scale RAID that was designed in 2003.

The third generation of IBM Elastic Storage Server was announced starting in October 2019 with the ESS 3000 NVMe Flash storage system. This announcement was followed by the announcement of the IBM ESS 5000 HDD storage systems in July 2020 and the ESS 3200 in May 2021. We have a deep and unique understanding of this technology because we have been developing it for the past 17 years.

1.1.2 Distinguishing features

IBM Spectrum Scale RAID distributes data and parity information across node failure domains to tolerate unavailability or failure of all physical disks in a node. It also distributes spare capacity across nodes to maximize parallelism in rebuild operations.

IBM Spectrum Scale RAID implements end-to-end checksums and data versions to detect and correct the data integrity problems of traditional RAID. Data is checked from the PDisk blocks on the ESS 3200 to the memory on the clients that connect over the network. It is the same checksum, not layers or serialized checksums that terminate in between the chain, so it really is an end-to-end checksum.

Figure 1-2 shows a simple example of declustered RAID. The left side shows a traditional RAID layout that consists of three 2-way mirrored RAID volumes and a dedicated spare disk that uses seven drives. The right side shows the equivalent declustered layout, which still uses seven drives. Here, the blocks of the three RAID volumes and the spare capacity are scattered over the seven disks.

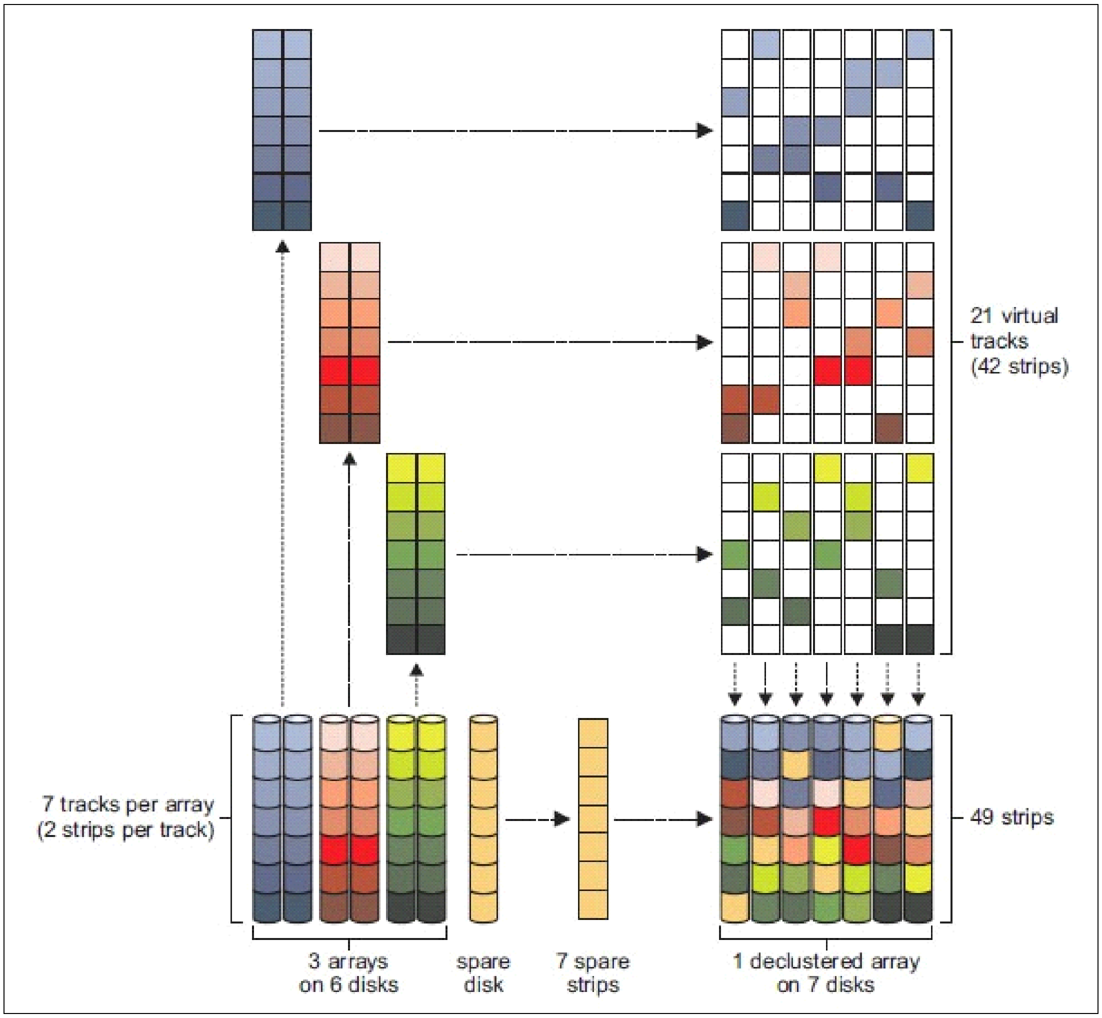


Figure 1-2 Declustered array versus 1+1 array

The declustered RAID layout provides the following advantages over the traditional RAID layout:

- Figure 1-3 shows a significant advantage of the declustered RAID layout over the traditional RAID layout after a drive failure. With the traditional RAID layout on the left side of Figure 1-3, the system must copy the surviving replica of the failed drive to the spare drive, reading only from one drive and writing only to one drive.

  However, with the declustered layout that is shown on the right of Figure 1-3, the affected replicas and the spares are distributed across all six surviving disks. This configuration rebuilds reads from all surviving disks and writes to all surviving disks, which greatly increases rebuild parallelism.

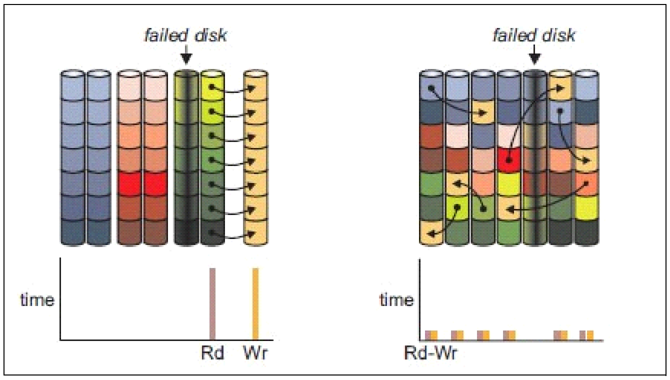


Figure 1-3 Array rebuild operation

- Another advantage of the declustered RAID technology that is used by ESS 3200 (and other IBM systems) is that it minimizes the worst-case number of critical RAID tracks in the presence of multiple disk failures. ESS 3200 can then handle restoring protection to critical RAID tracks as a high priority, while giving lower priority to RAID tracks that are not considered critical.

  For example, consider an 8+3p RAID code on an array of 100 PDisks. In the traditional layout and declustered layout, the probability that a specific RAID track is critical is 11/100 × 10/99 × 9/98 (0.1%). However, when a track is critical in the traditional RAID array, all tracks in the volume are critical, whereas with declustered RAID, only 0.1%, of the tracks are critical. By prioritizing the rebuild of more critical tracks over less critical tracks, ESS 3200 quickly gets out of critical rebuild and then can tolerate another failure.

  ESS 3200 adapts these priorities dynamically; if a *non-critical* RAID track is used and more drives fail, this RAID track's rebuild priority can be escalated to *critical*.

- A third advantage of declustered RAID is that it makes it possible to support any number of drives in the array and to dynamically add and remove drives from the array. Adding a drive in a traditional RAID layout (except in the case of adding a spare) requires significant data reorganization and restriping. However, only targeted data movement is needed to rebalance the array to include the added drive in a declustered array.

1.2 IBM Elastic Storage System (ESS)

ESS is based on IBM Spectrum Scale Native RAID to provide the physical-disk protection and is tightly integrated with IBM Spectrum Scale to provide the file system access over the network to all of the IBM Spectrum Scale clients. Other protocols can be used to access the IBM Spectrum Scale file system.

Because it falls outside of the scope of this publication, for details about the ways to access an IBM Spectrum Scale file system, use the following IBM Documentation link:

https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=planning

You can also see the following IBM Redpaper: Introduction Guide to the IBM Elastic Storage System, REDP-5253.

1.3 IBM Elastic Storage System 3200

IBM Elastic Storage System 3200 (ESS 3200) is designed to meet and beat the challenge of managing data for analytics. Packaged in a compact 2U enclosure, ESS 3200 is a proven data-management solution. It speeds time-to-value for artificial intelligence (AI), deep learning, and high-performance computing workloads thanks to its quick all-NVMe storage and simple, fast containerized software install and upgrade. The no-compromise hardware and software design of ESS 3200 provides you with the industry-leading performance that is required to keep data-hungry processors fully utilized. ESS 3200 is compatible with all IBM Elastic Storage System models.

IBM Elastic Storage System 3200 can contain up to 24 NVMe-attached SSD drives, 12 drives (half-populated), or 24 drives (fully-populated).

For details on the ESS 3200, see the following IBM Documentation link:

https://www.ibm.com/docs/en/ess/6.1.1_cd

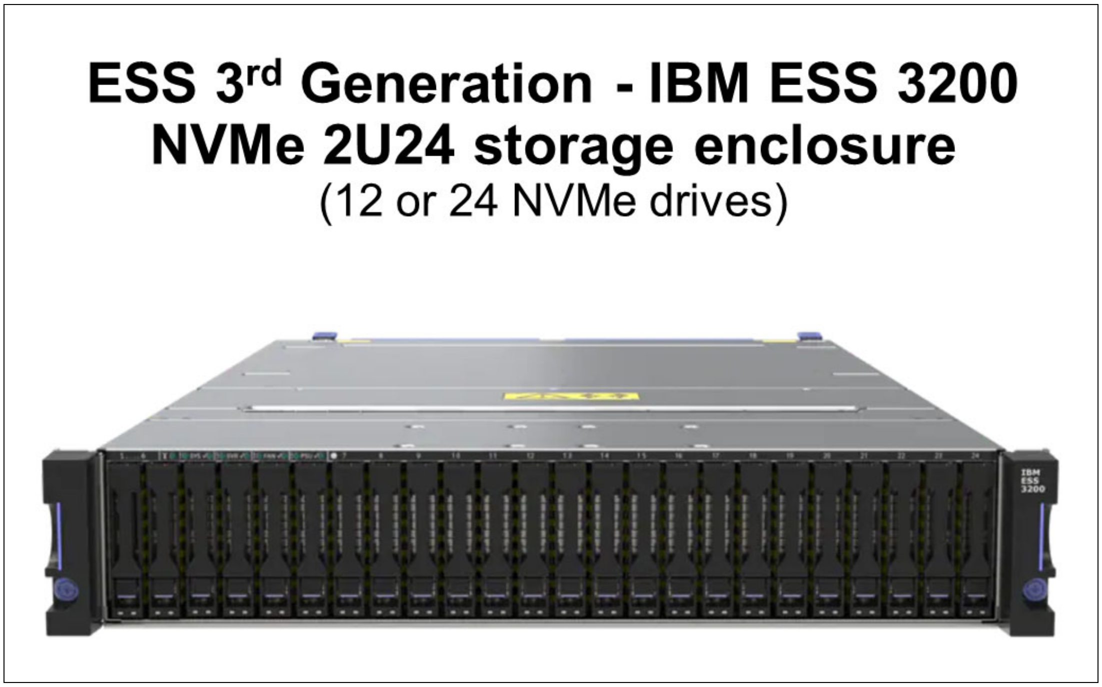
What is new in ESS 3200?

The ESS 3200 is part of the third generation of IBM ESS. ESS 3200 was announced in May 2021 and includes the following features:

- Based on a new 2U24 storage enclosure with PCI Gen4-based x86 server canisters to provide significant improvements in throughput and bandwidth capability.
- Uses standard enterprise class NVMe drives:
  - You can order half-populated 12, or fully-populated 24, 2.5” NVMe drives in capacities of 3.84TB, 7.68TB, and 15.36TB. Using the largest capacity (15.36TB NVMe drives), you can scale to 260TB usable capacity, in a 2U form factor along with associated low-weight and low-power consumption.
  - If ordered half-populated, you can add the remaining 12 NVMe drives non-disruptively (the additional drives must be of the same size as the first 12 drives).
- Designed to provide:
  - High performance: NVMe flash storage with up to 80GB/second read throughput per 2U building block.
  - Designed to provide Edge capability and global data access: This solution can be deployed in either data centers or at the edge, ingesting and processing data that then uses IBM Spectrum Scale Active File Management (AFM) to share the data globally.
  - Simplicity: Containerized software installation and upgrade, plus a powerful management GUI, minimize demands on IT staff time and expertise.
- Deployed by way of containerized Red Hat Ansible playbooks that provide significantly improved ease-of-use and orchestration of complex IBM ESS administration tasks, such as cluster configuration, file system creation, and code update.

The third-generation ESS-3200 addresses the challenges of managing today's data. ESS 3200 delivers a new generation of high-performance software-defined flash storage. It builds on years of experience and couples proven IBM Spectrum Scale software with lightning-fast NVMe storage technology to offer industry-leading file management capabilities. The ESS 3200 builds on and extends a track record of meeting the needs of the smartest, most demanding organizations. The ESS 3200 is up to 100% faster than previous generation of ESS NVMe storage.

Figure 1-4 shows the ESS 3200 NVMe storage building block.



ESS 3rd Generation - IBM ESS 3200
NVMe 2U24 storage enclosure
(12 or 24 NVMe drives)

Figure 1-4 Third-generation ESS 3200 model

1.3.1 Value added

ESS 3200 provides an extreme high-performance tier of IBM Spectrum Scale file storage, for a broad variety of AI, analytics, and Big Data applications. ESS 3200 is designed to keep GPUs in AI workloads running at peak performance. Like all IBM ESSs, ESS 3200 runs the proven IBM Spectrum Scale RAID erasure coding, which provides superior consistent high performance, mitigation of storage hardware failures. It also provides intelligent monitoring, management, and dynamic tuning of all IBM Elastic Storage Systems for IBM Spectrum Scale data.

Note: IBM ESS performance is available upon request from IBM or IBM Business Partner representative. They use the IBM File and Object Solution Design Engine to estimate performance that is based on your workload and network environment.

Optimum IBM ESS performance is derived from unconstrained IOR benchmark for 100% sequential read numbers by using unconstrained InfiniBand networks. Other networks (such as 100 GbE, 40 GbE, and 10 GbE) have more overhead than InfiniBand and typically lower aggregate bandwidth capabilities result.

For more information, contact your IBM or IBM Business Partner representative.

ESS 3200 is designed to be the simplest way, currently, for users to deploy IBM Spectrum Scale. Spectrum Scale is included in a pre-configured system. Installations and updates are delivered by way of containerized software that speeds and simplifies the process.

A storage specialist from IBM System Lab Services implementation is not required if you have an ESS and IBM Spectrum Scale system and you are comfortable with ESS implementations. It is much easier to install, and maintenance can be performed by your IT staff.

If the customer is unfamiliar with IBM Spectrum Scale and ESS, IBM recommends that IBM System Lab Services be used to assure high satisfaction with your initial ESS 3200 installation.

Fast time-to-value

ESS 3200 combines IBM Spectrum Scale file management software with NVMe flash storage for the ultimate in scale-out performance and simplicity, delivering 80GB/s of data throughput per 2U system.

Operational efficiency

The demands on IT staff time and expertise are minimized by the containerized software install and a powerful management GUI. Dense storage within a 2U package means a small data center footprint.

Reliability

The software-defined erasure coding assures data recovery while using less space than data replication. Restores can take minutes, rather than hours or days, and can be run without disrupting operations.

Deployment flexibility

ESS 3200 is available in a wide range of capacities from tens to hundreds of terabytes per 2U. Deploy as a standalone edge system or scale out with additional ESS 3200 systems or with IBM Elastic Storage System.

1.4 License considerations

ESS 3200 follows the same license model as the other ESS products. The two currently available options for ESS are IBM Spectrum Scale for ESS *Data Access Edition* and IBM Spectrum Scale for *ESS Data Management Edition*.

ESS uses capacity-based licensing, which means that a customer can connect as many clients as desired without extra license costs. For other types of configurations, contact IBM or your IBM Business Partner for license details.

For more information about licensing on IBM Spectrum Scale and ESS, see the following IBM Documentation links:

https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=overview-capacity-based-licensing

https://www.ibm.com/docs/en/spectrum-scale?topic=STXKQY/gpfsclustersfaq.html#moreinfo

2

ESS 3200 architecture and overview

This chapter describes the architecture and provides an overview of IBM Elastic Storage System 3200 (ESS 3200). It covers the following topics:

- 2.1, “Platform” on page 10
- 2.2, “GUI overview” on page 12
- 2.3, “Software enhancements” on page 25
- 2.4, “RAS enhancements” on page 31
- 2.5, “Performance” on page 36

2.1 Platform

An IBM Elastic Storage System 3200 (ESS 3200) enclosure contains Non-Volatile Memory Express (NVMe)-attached SSD drives and a pair of server canisters. ESS 3200 is an all-Flash array platform. This storage platform uses NVMe-attached SSD drives to provide significant performance improvements as compared to SAS-attached flash drives. This chapter provides an overview of the ESS 3200 platform.

2.1.1 Canisters and servers

This section describes the CPU, memory, and networking for the ESS 3200 system.

CPU

The ESS 3200 system uses a single socket AMD EPYC Rome processor per I/O canister node for a total of two CPUs per enclosure. Figure 2-1 shows a CPU in a canister.

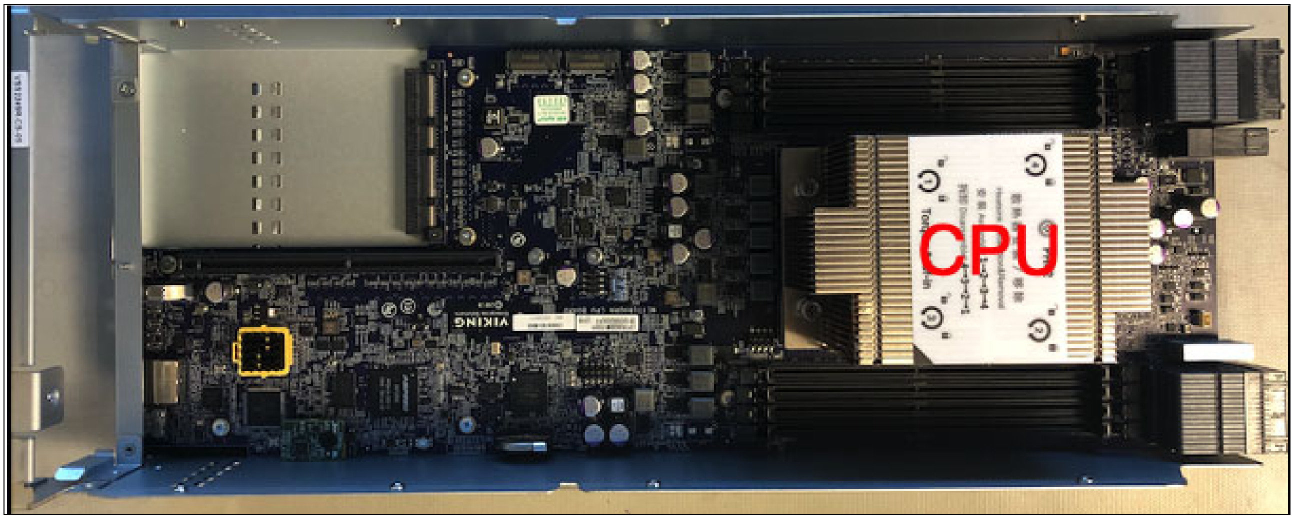


Figure 2-1 Single CPU in a canister

Memory

The memory of each canister and enclosure is depicted in Table 2-1.

Table 2-1 Memory configuration

Number of DIMMs per server canister	Total memory per server canister	Number of DIMMs per server enclosure	Total memory per server enclosure
8 (64 GB DIMM only)	512 GB	16	1 TB

Networking

The ESS 3200 includes two adapters per server canister that consist of the following features:

- InfiniBand - EDR 100 Gb / HDR100 100 Gb / HDR200 200 Gb
- Ethernet - 100 GbE

Figure 2-2 shows the HBA locations.

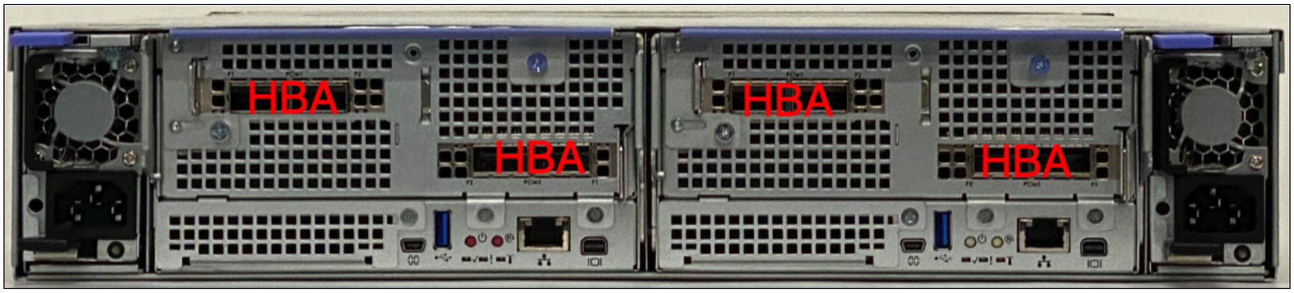


Figure 2-2 Two HBA slots per canister

2.1.2 Peripheral Component Interconnect Express (PCIe)

The following is a breakdown of the PCIe lanes:

- 128 x Gen4 PCIe from each CPU (switchless)
  - NVMe drives - 64 [48 used (24 x x2)]
  - HBA PCIe adapters - 48 (3 x x16)
- Nontransparent bridge-to-peer canister - x8 Gen3
- Boot Drive - 8 (2 x x4 G3)

Figure 2-3 shows the PCIe lane details.

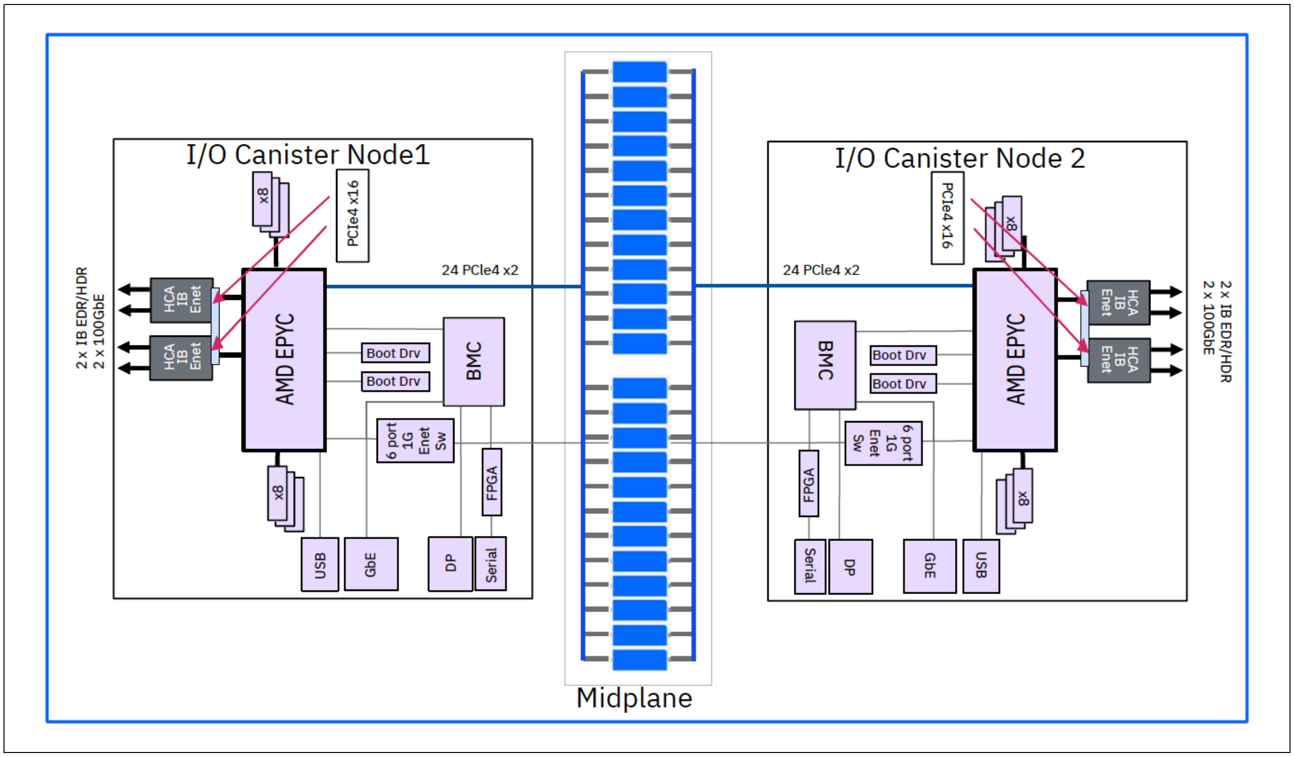


Figure 2-3 PCIe lanes in ESS 3200

2.2 GUI overview

A graphical user interface (GUI) service runs on the enterprise management server (EMS). It can be used to monitor the health of the ESS and to perform management tasks. This chapter provides an overview of the GUI, but is not comprehensive.

The s**ystemctl** command can be run on the EMS to start or stop the GUI. Table 2-2 shows the **systemctl** command options.

Table 2-2 The systemctl command options

Command	Description
Start the GUI service	**systemctl start gpfsgui**
Check the status of the GUI service	**systemctl status gpfsgui**
Stop the GUI service	**systemctl stop gpfsgui**

To access the GUI, enter the IP address or host name of the EMS in a web browser using the secure https mode (`https://`*`<IP or hostname of EMS>`*).

2.2.1 GUI users

GUI users must be created before the GUI can be used. To grant special rights, roles are assigned to the users.

When the GUI is used for the first time, an initial user must be created:

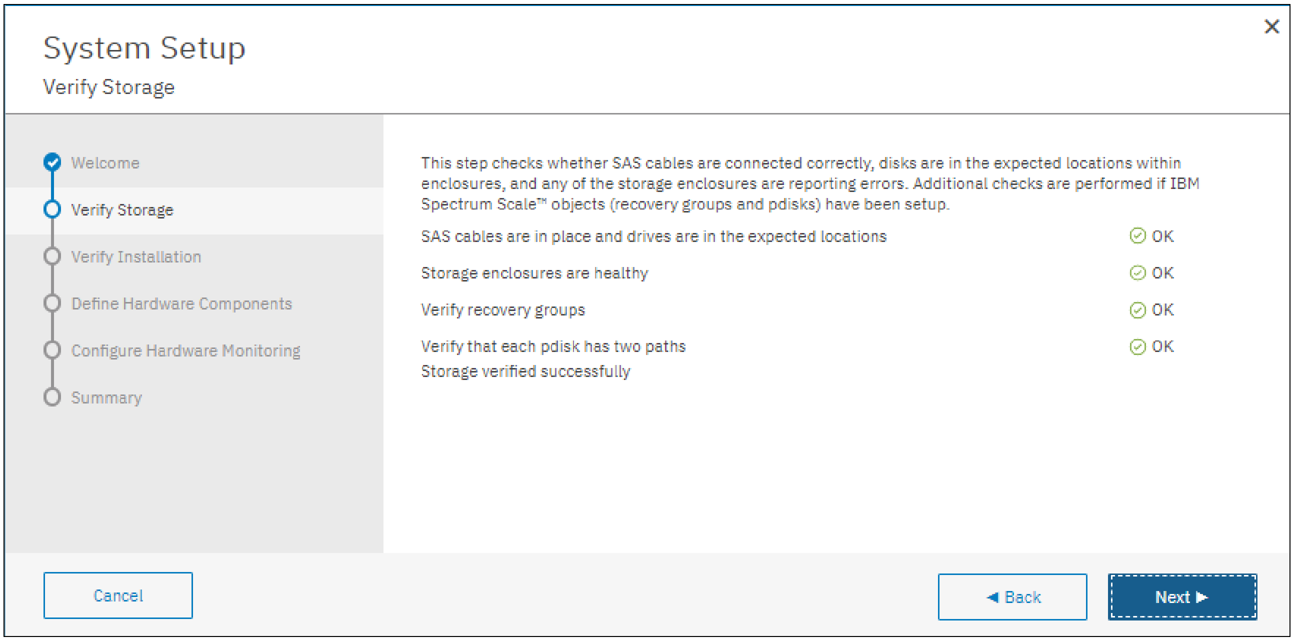
```
/usr/lpp/mmfs/gui/cli/mkuser <username> -g SecurityAdmin
```

Once the initial user is created, a user can log in to the GUI with the newly-created user and create more users on the **Services** → **GUI** → **Users** page. By default, users are stored in an internal user repository. Alternatively, an external user repository can also be utilized. An external user repository can be configured on the **Services** → **GUI** → **External Authentication** page.

2.2.2 System setup wizard

After logging into the GUI for the first time, the system-setup wizard is launched. The following is a high-level overview of what to expect when using the system-setup wizard.

1. After the welcome page, the **Verify Storage** page queries important system information and performs several checks to verify whether the system is ready for use. See Figure 2-4.



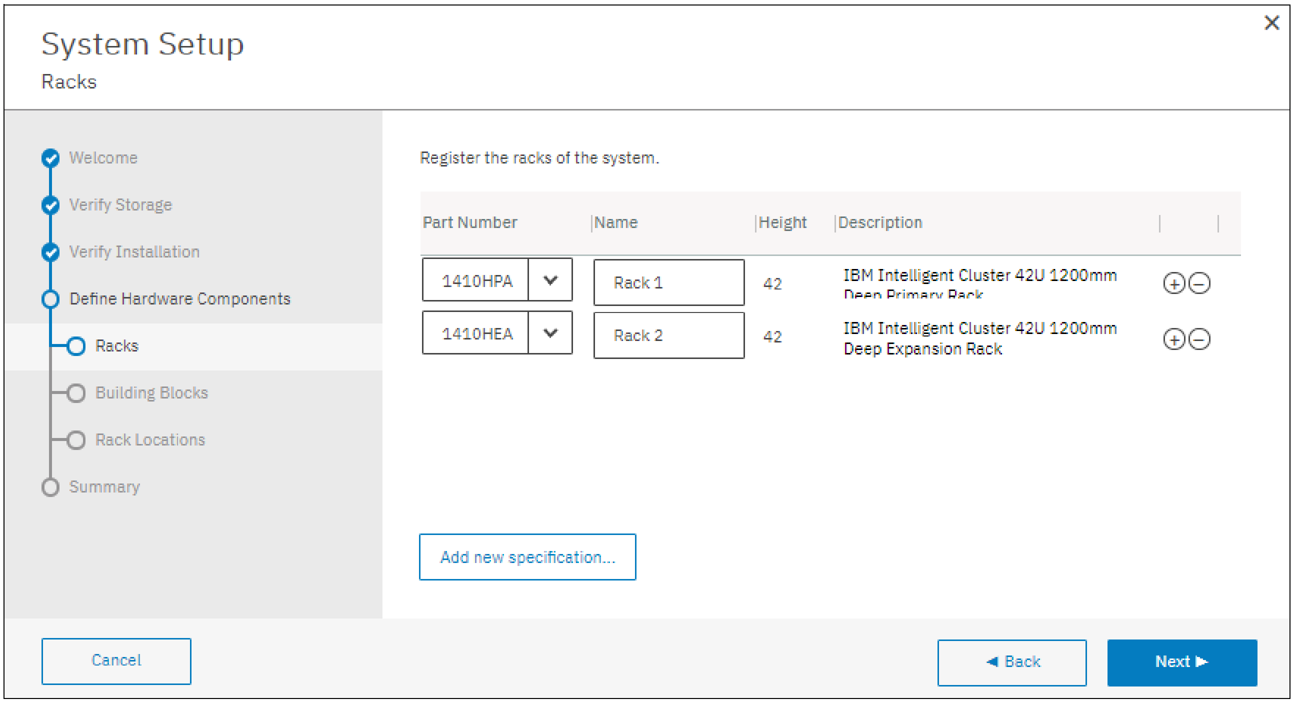
This step checks whether SAS cables are connected correctly, disks are in the expected locations within enclosures, and any of the storage enclosures are reporting errors. Additional checks are performed if IBM Spectrum Scale™ objects (recovery groups and pdisks) have been setup.

SAS cables are in place and drives are in the expected locations	✓ OK
Storage enclosures are healthy	✓ OK
Verify recovery groups	✓ OK
Verify that each pdisk has two paths	✓ OK

Storage verified successfully

Figure 2-4 The System Setup wizard

2. Once all checks pass, the user can define where the ESS 3200 systems are installed on the **Racks** page. The user can either choose a predefined rack type or choose **Add new specification** if none of the rack types match the available rack. It is important that the selected rack-type has the same number of height units. A meaningful name can be specified for the racks to create. See Figure 2-5.

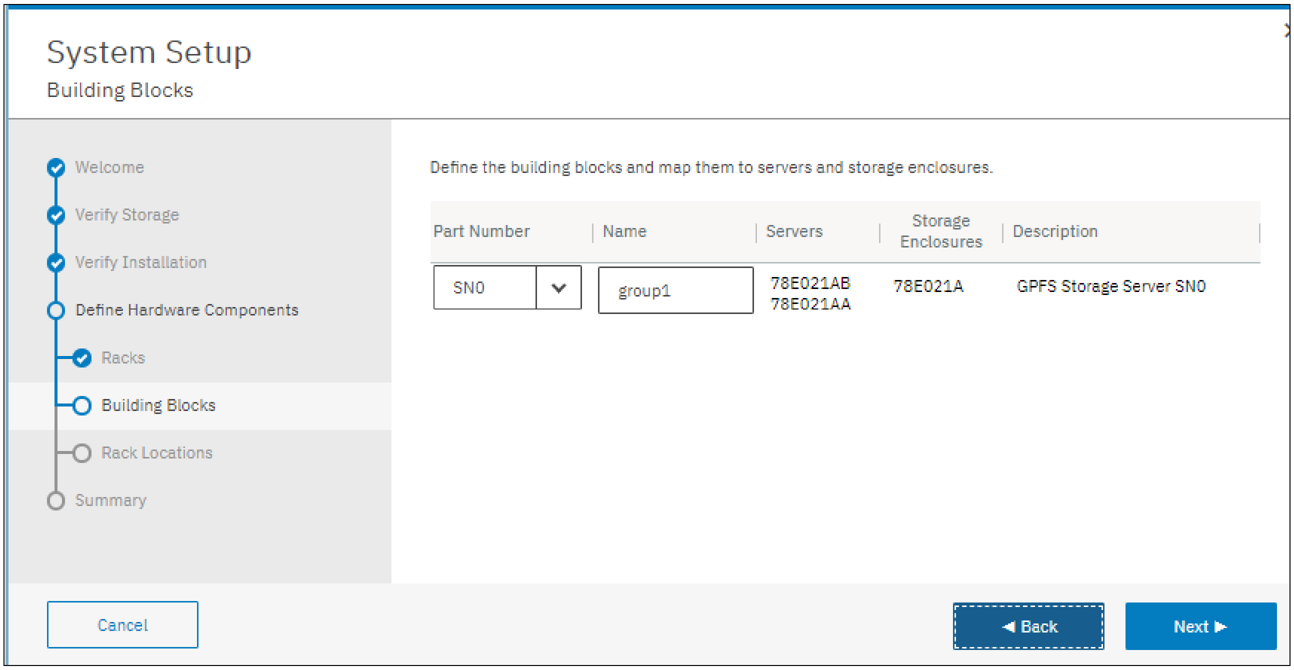


Register the racks of the system.

Part Number	Name	Height	Description
1410HPA	Rack 1	42	IBM Intelligent Cluster 42U 1200mm Deep Primary Rack
1410HEA	Rack 2	42	IBM Intelligent Cluster 42U 1200mm Deep Expansion Rack

Figure 2-5 Specifying the racks

3. The **Building Blocks** page displays one row for each ESS 3200 or other ESS models. The user can assign names to each building block or go with the default. See Figure 2-6.



System Setup

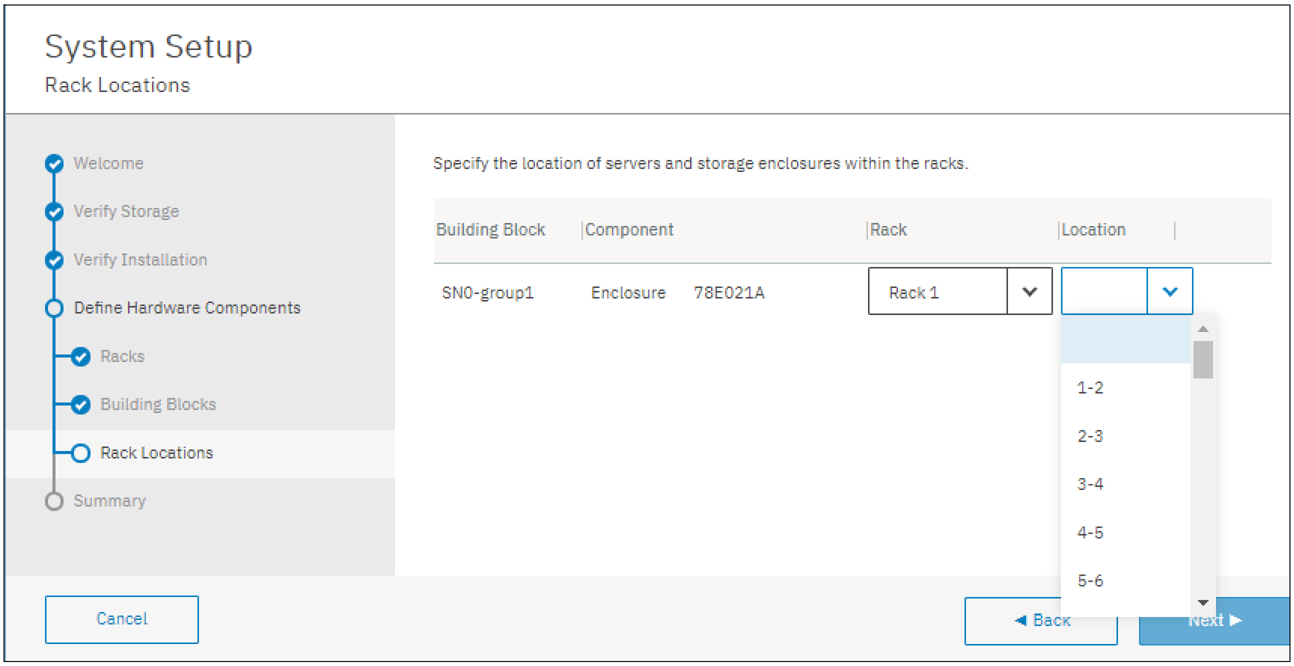
Building Blocks

Define the building blocks and map them to servers and storage enclosures.

Part Number	Name	Servers	Storage Enclosures	Description
SN0	group1	78E021AB 78E021AA	78E021A	GPFS Storage Server SN0

Figure 2-6 Define building blocks

4. The ESS 3200 systems are assigned to the rack locations in which they are mounted. See Figure 2-7.



System Setup

Rack Locations

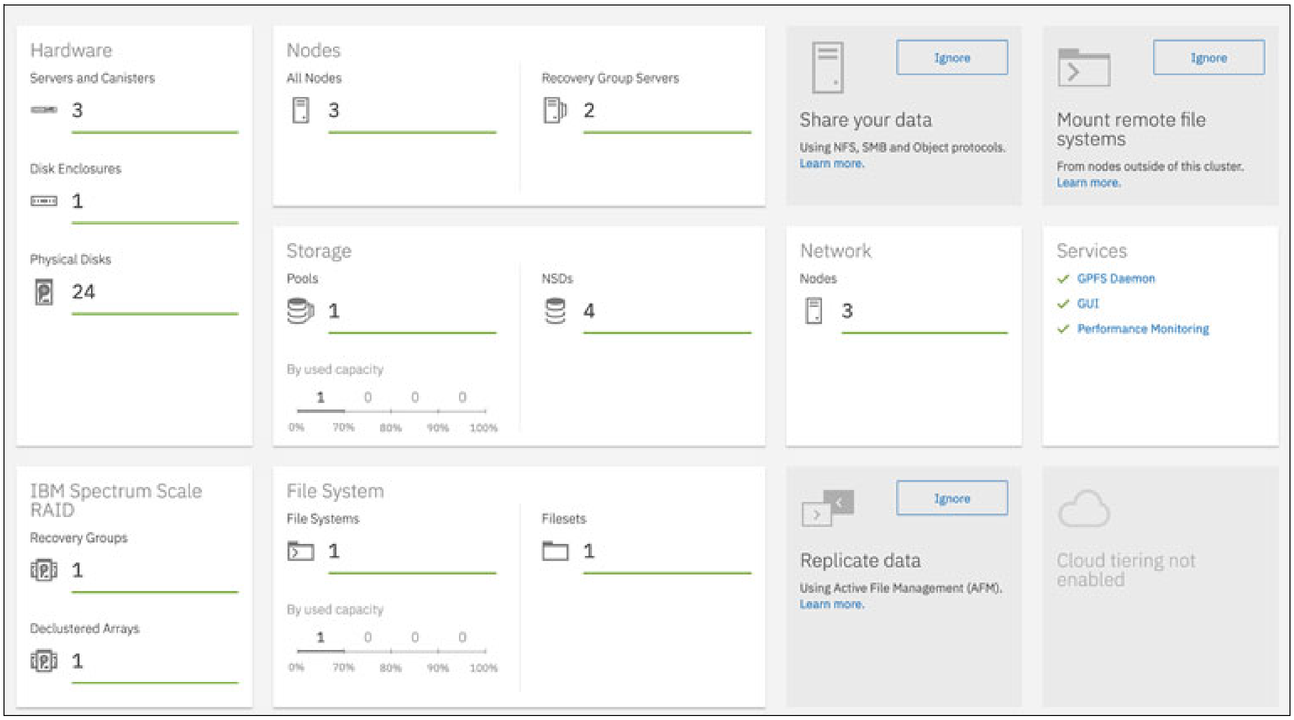
Specify the location of servers and storage enclosures within the racks.

Building Block	Component	Rack	Location
SN0-group1	Enclosure 78E021A	Rack 1	

Figure 2-7 Assign rack locations

2.2.3 Using the GUI

After you log in to the GUI, the **Overview** page is displayed. This page provides a view of all objects in the system and their health states. Clicking the numbers or links displays a more detailed view. See Figure 2-8.



Using NFS, SMB and Object protocols. Learn more.

From nodes outside of this cluster. Learn more.

Using Active File Management (AFM). Learn more.

Figure 2-8 The Overview page

The header area of the GUI provides a quick view of the current health problems and tips for improvement, if applicable. Additionally, links exist to some help resources.

Use the navigation menu on the left side of the GUI page to navigate to other GUI pages (see Figure 2-9). Each GUI page has a unique URL that the user can use to directly access the page, bookmark pages, and start the GUI in-context.

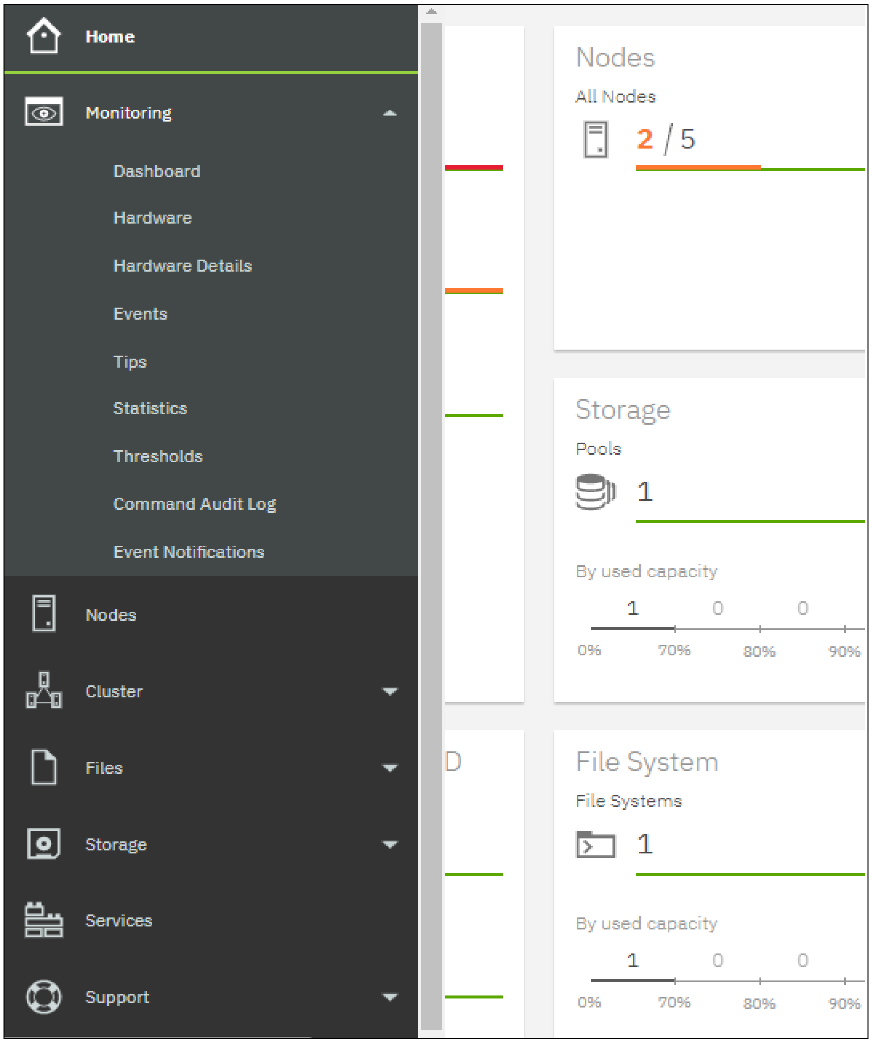
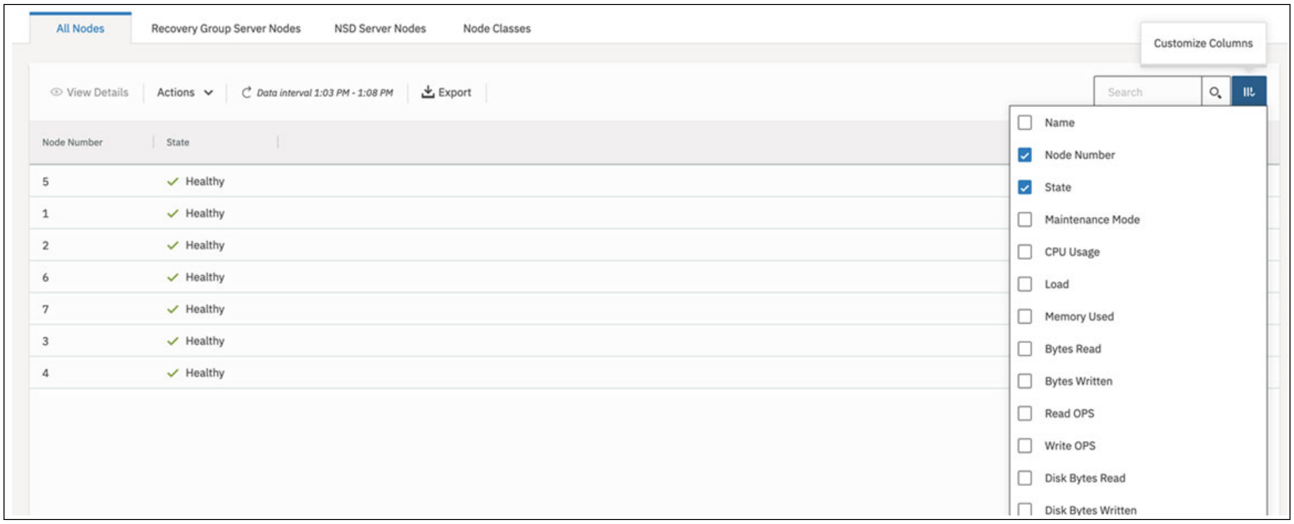


Figure 2-9 Navigation pane of the GUI

Some menus, such as **Protocols**, are only displayed when the related features, such as NFS, SMB, or AFM are enabled.

Most tables that are shown in the GUI have columns that are hidden by default. Right-click the table header and select the columns to display the columns. See Figure 2-10.



Node Number	State
5	✓ Healthy
1	✓ Healthy
2	✓ Healthy
6	✓ Healthy
7	✓ Healthy
3	✓ Healthy
4	✓ Healthy

- □ Name
- ☑ Node Number
- ☑ State
- □ Maintenance Mode
- □ CPU Usage
- □ Load
- □ Memory Used
- □ Bytes Read
- □ Bytes Written
- □ Read OPS
- □ Write OPS
- □ Disk Bytes Read
- □ Disk Bytes Written

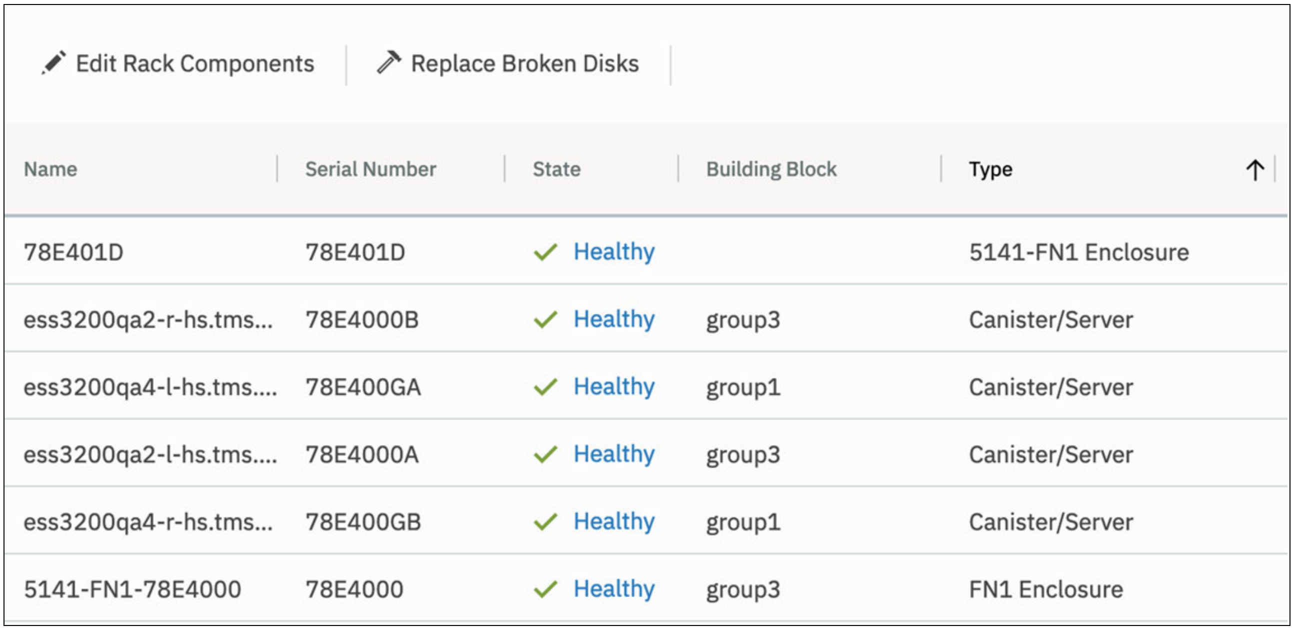
Figure 2-10 Show and hide table columns

The table values can be sorted by clicking one of the column headers. A little arrow in the table header indicates the sorting.

Double-click a table row to open a more detailed view of the selected item.

2.2.4 Monitoring of ESS 3200 hardware

The **Monitoring** → **Hardware** page displays the ESS 3200 enclosures within the racks. A table lists all enclosures and the related canisters. See Figure 2-11.



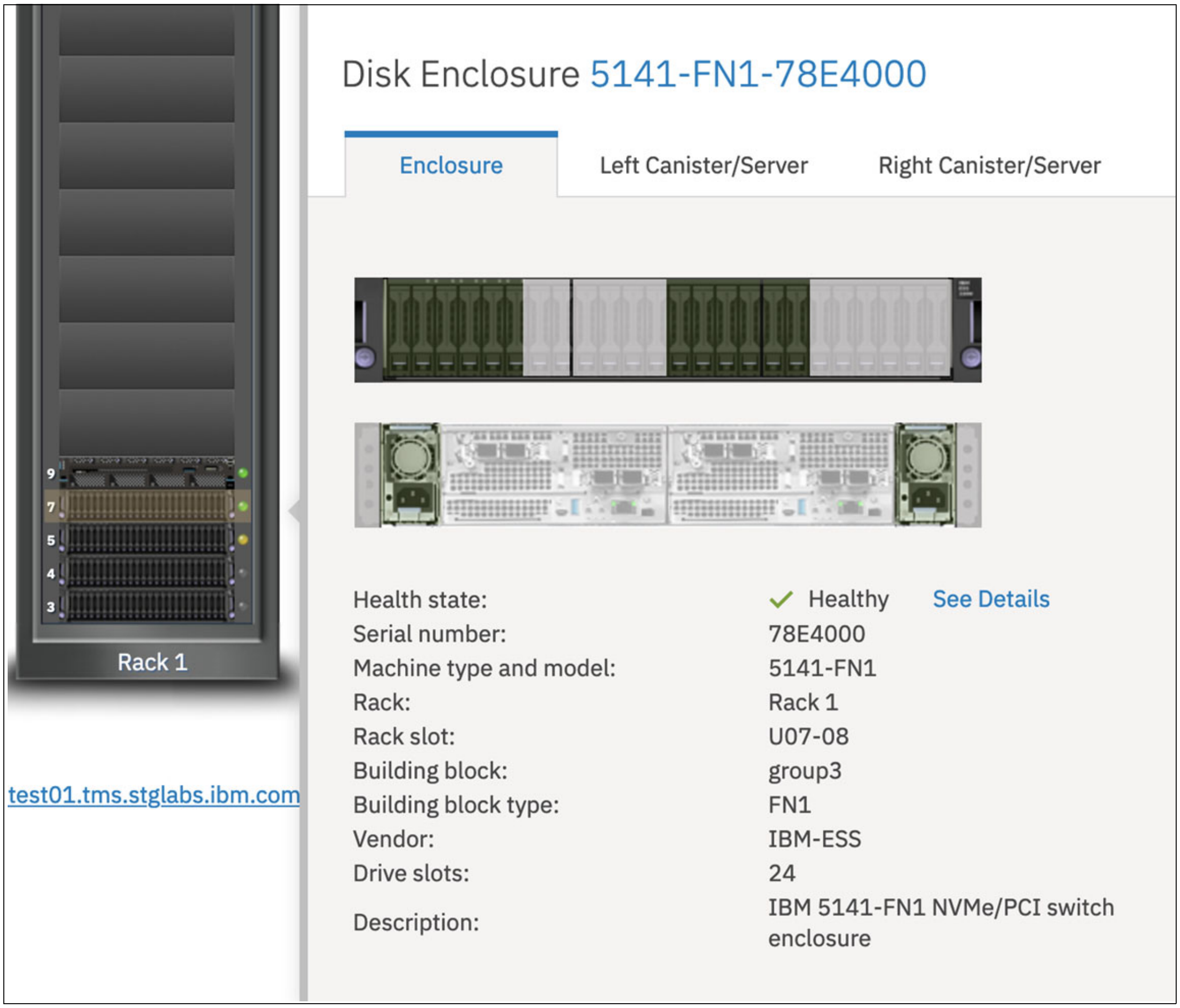
Name	Serial Number	State	Building Block	Type ↑
78E401D	78E401D	✓ Healthy		5141-FN1 Enclosure
ess3200qa2-r-hs.tms...	78E4000B	✓ Healthy	group3	Canister/Server
ess3200qa4-l-hs.tms....	78E400GA	✓ Healthy	group1	Canister/Server
ess3200qa2-l-hs.tms....	78E4000A	✓ Healthy	group3	Canister/Server
ess3200qa4-r-hs.tms...	78E400GB	✓ Healthy	group1	Canister/Server
5141-FN1-78E4000	78E4000	✓ Healthy	group3	FN1 Enclosure

Figure 2-11 Hardware page with two ESS 3200 systems

Use **Edit Rack Components** when ESS enclosures or servers are added or removed, or if their rack location changes.

The **Replace Broken Disks** action launches a guided procedure to replace broken disks if there are any.

Click the ESS 3200 in the virtual rack to see more information about the ESS 3200, including the physical disks and the two canisters (Figure 2-12). Move the mouse over the components, such as drives and power supplies, to see more information. Clicking components moves to a page with more detailed information. Broken disks are indicated with the color red, and a context menu (right-click) enables the user to replace the selected broken disk.



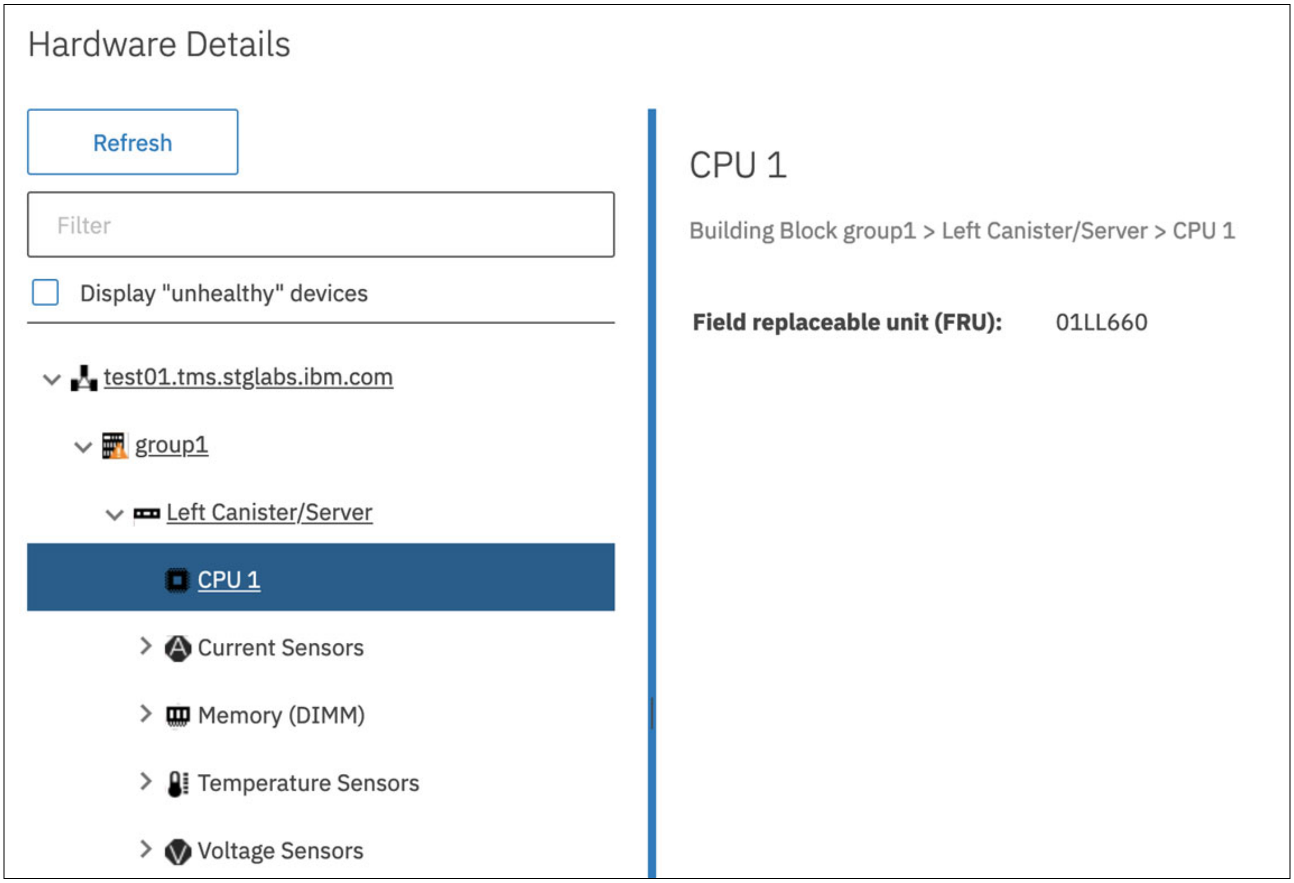
Disk Enclosure 5141-FN1-78E4000

Health state:	✓ Healthy
Serial number:	78E4000
Machine type and model:	5141-FN1
Rack:	Rack 1
Rack slot:	U07-08
Building block:	group3
Building block type:	FN1
Vendor:	IBM-ESS
Drive slots:	24
Description:	IBM 5141-FN1 NVMe/PCI switch enclosure

Figure 2-12 ESS 3200 details in the Monitoring → Hardware page

If there is more than one rack, click the arrows that are displayed on the left and the right side of the rack to switch to another rack.

The **Monitoring** → **Hardware Details** page displays more detailed information and the health states of the ESS 3200 and its internal components. See Figure 2-13.


Hardware Details

CPU 1

Building Block group1 > Left Canister/Server > CPU 1

Field replaceable unit (FRU): 01LL660

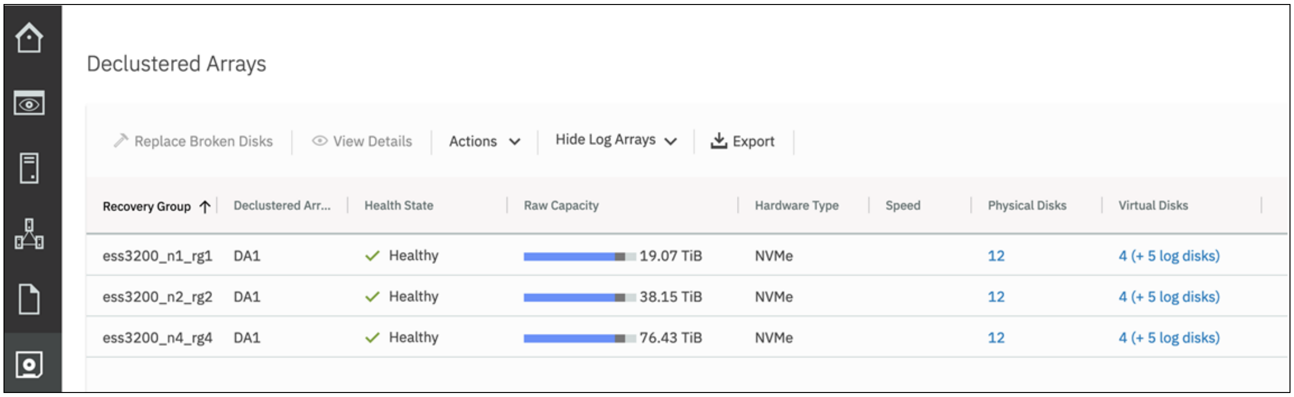
Figure 2-13 The Hardware Details page

This page allows the user to search for components by text and filter the results to display only unhealthy hardware.

Click the **>** icon on the tree nodes to display subsequent children. For example, the user can click to display all Current Sensors of the canister in Figure 2-13.

2.2.5 Storage

The **Storage** menu provides various views into the storage, such as the physical disks, declustered arrays (Figure 2-14), recovery groups, virtual disks, network shared disks (NSDs), and storage pools.



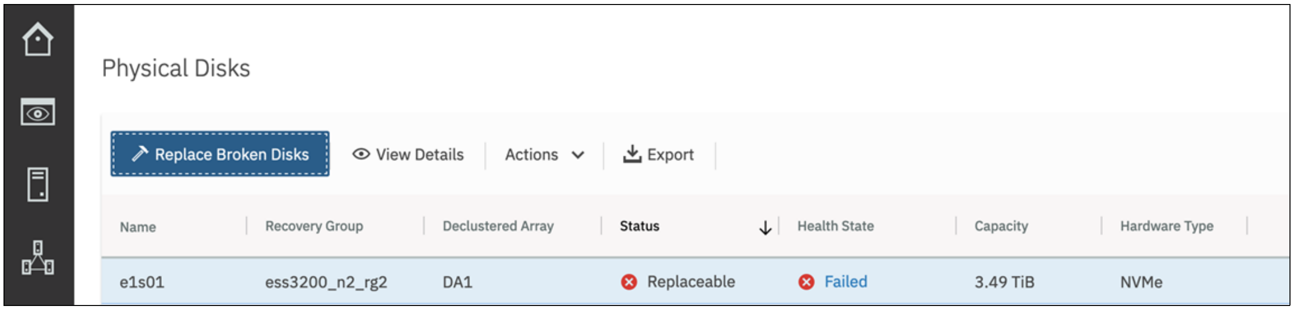
Recovery Group	Declustered Arr...	Health State	Raw Capacity	Hardware Type	Speed	Physical Disks	Virtual Disks
ess3200_n1_rg1	DA1	✓ Healthy	19.07 TiB	NVMe		12	4 (+ 5 log disks)
ess3200_n2_rg2	DA1	✓ Healthy	38.15 TiB	NVMe		12	4 (+ 5 log disks)
ess3200_n4_rg4	DA1	✓ Healthy	76.43 TiB	NVMe		12	4 (+ 5 log disks)

Figure 2-14 Storage → Declustered Arrays

2.2.6 Replace broken disks

The GUI provides a guided procedure that can be used to replace broken disks. Make sure that the replacement disks have the same field-replaceable units (FRUs) as the disks that are going to be replaced.

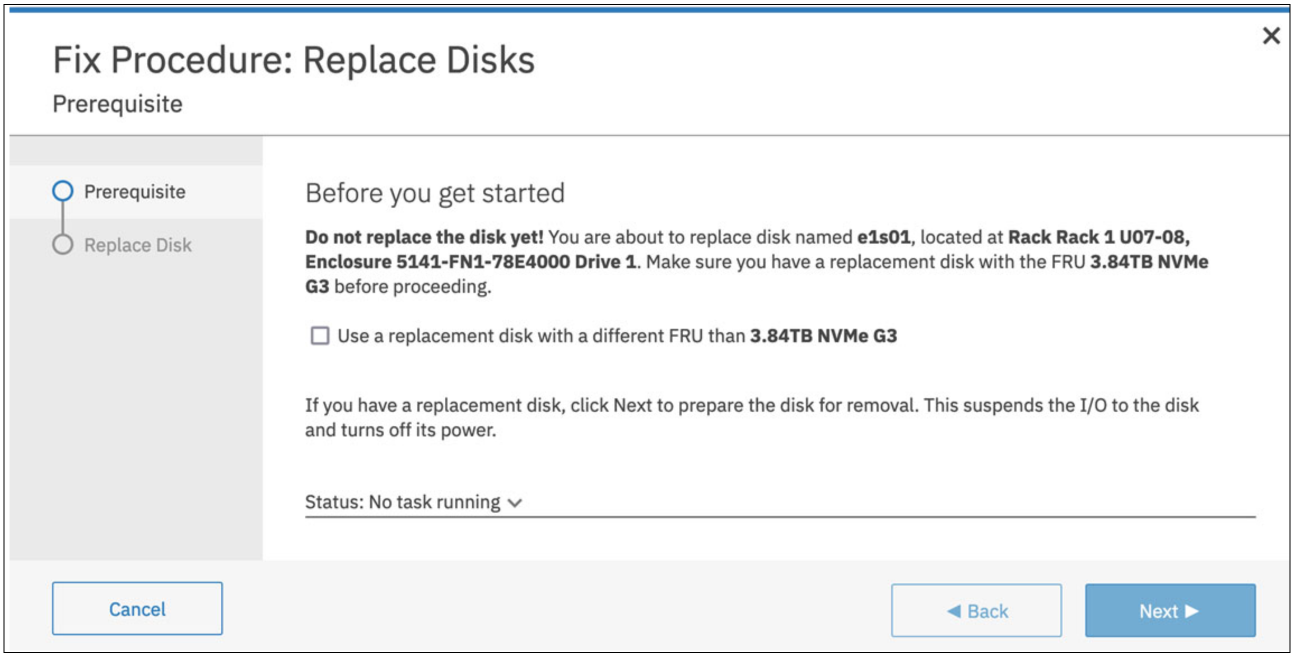
The procedure can be launched from different places. A good place to look for broken disks is the **Storage** → **Physical Disks** page that is shown in Figure 2-15.



Name	Recovery Group	Declustered Array	Status	Health State	Capacity	Hardware Type
e1s01	ess3200_n2_rg2	DA1	⊗ Replaceable	⊗ Failed	3.49 TiB	NVMe

Figure 2-15 Physical Disks page

Choose the **Replace Broken Disks** action to get a list of all broken disks and choose some to replace. Optionally, select an individual disk from the table and choose the **Replace Disk** action to replace the selected disk. In both cases, a fix procedure guides the user through replacing the disks. See Figure 2-16.



Fix Procedure: Replace Disks

Prerequisite

Before you get started

Do not replace the disk yet! You are about to replace disk named **e1s01**, located at **Rack Rack 1 U07-08, Enclosure 5141-FN1-78E4000 Drive 1**. Make sure you have a replacement disk with the FRU **3.84TB NVMe G3** before proceeding.

☐ Use a replacement disk with a different FRU than **3.84TB NVMe G3**

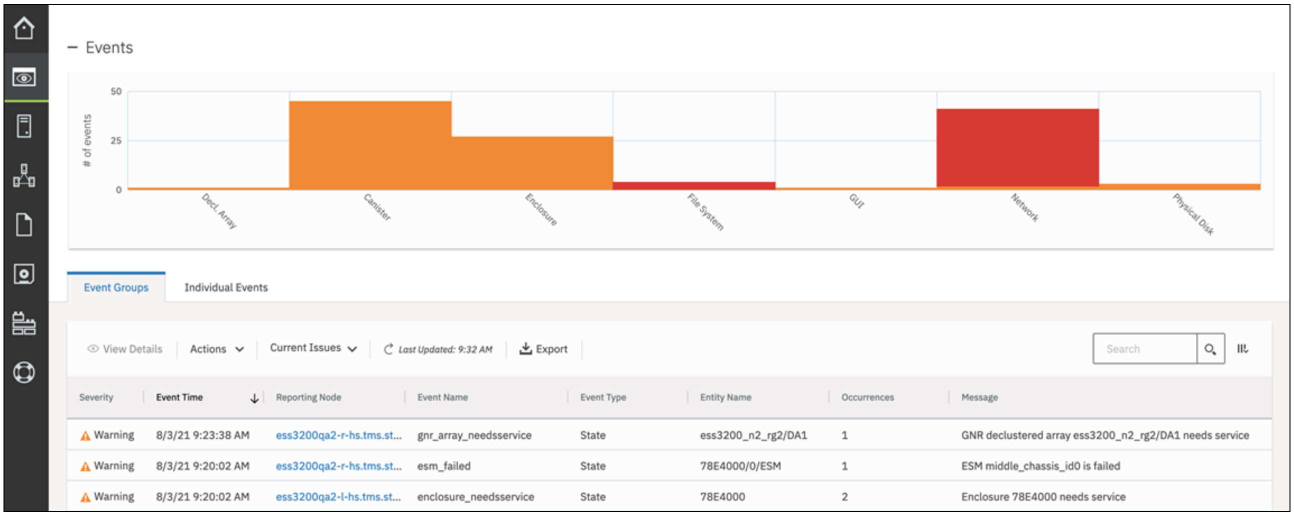
If you have a replacement disk, click Next to prepare the disk for removal. This suspends the I/O to the disk and turns off its power.

Status: No task running

Figure 2-16 Fix procedure for replacing disks

2.2.7 Health events

Use the **Monitoring** → **Events** page to review the entire set of events that are reported in the ESS system. Under the **Event Groups** tab, all individual events with the same name are grouped into single rows, which is especially useful for a large volume of events. The **Individual Events** tab lists all the events, irrespective of the multiple occurrences. Events are assigned to a component, such as canister, enclosure, or file system. The user can click any of the components in the bar chart above the grid to filter for events of that selected component. See Figure 2-17.



Severity	Event Time	Reporting Node	Event Name	Event Type	Entity Name	Occurrences	Message
⚠ Warning	8/3/21 9:23:38 AM	ess3200qa2-r-hs.tms.st...	gnr_array_needsservice	State	ess3200_n2_rg2/DA1	1	GNR declustered array ess3200_n2_rg2/DA1 needs service
⚠ Warning	8/3/21 9:20:02 AM	ess3200qa2-r-hs.tms.st...	esm_failed	State	78E4000/0/ESM	1	ESM middle_chassis_id0 is failed
⚠ Warning	8/3/21 9:20:02 AM	ess3200qa2-l-hs.tms.st...	enclosure_needsservice	State	78E4000	2	Enclosure 78E4000 needs service

Figure 2-17 The Events page

The following filter options by event type are available as a drop-down list in the **Events** page shown in Figure 2-17:

- **Current Issues** displays all unfixed errors and warnings.
- **Notices** displays all transient messages of type “notice” that were not marked as read. While active state events disappear when the related problem is solved, the notices stay forever until they are marked as read.
- **Current State** displays all events that define the current state of the entities, and excludes notices and historic events.
- **All Events** displays all messages, even historic messages and messages that are marked as read. This filter is not available in the Event Groups view because of performance implications.

The user can mark events of type **Notices** as `read` to change the status of the event in the Events view. The status icons become gray if an error or warning is fixed, or if it is marked as `read`.

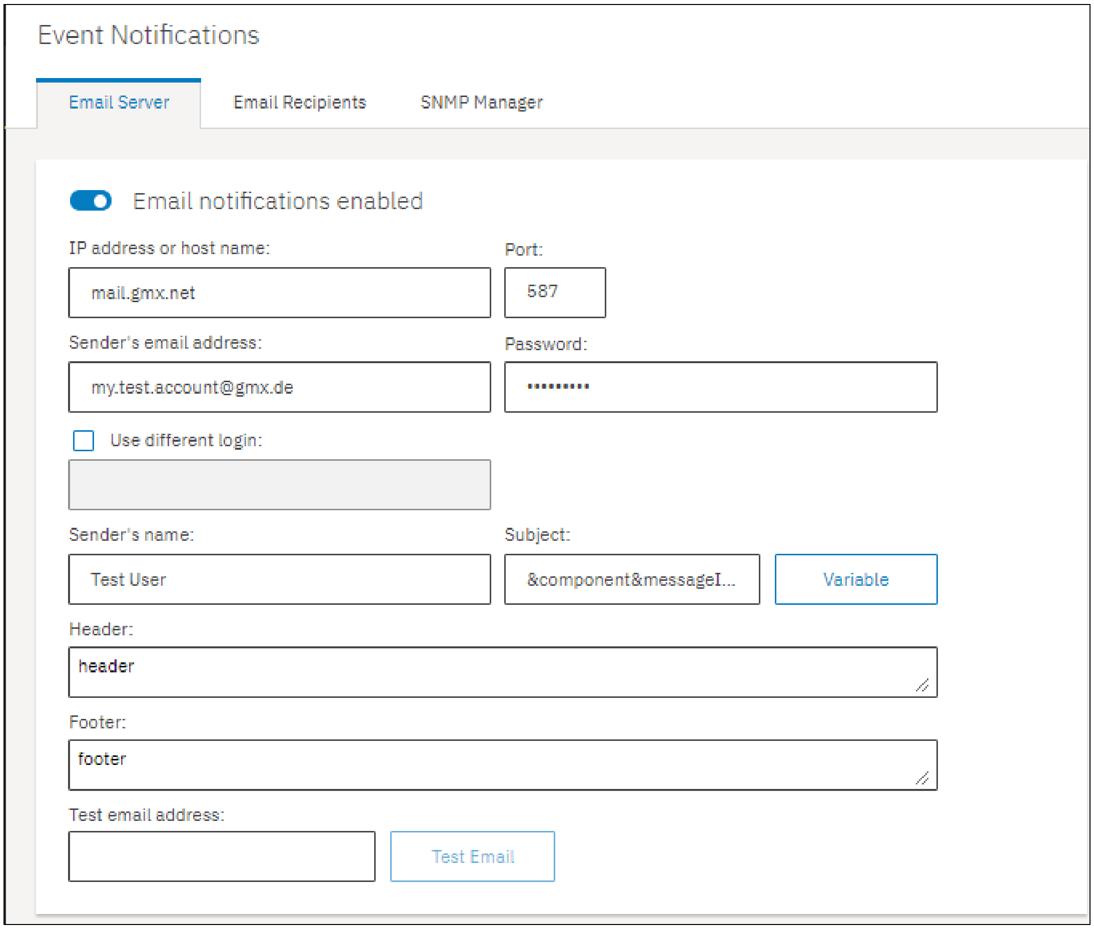
Some issues can be resolved by using the **Run Fix Procedure** action, which is available on select events. Right-click an event in the events table to see this option.

2.2.8 Event notification

The system can send emails and Simple Network Management Protocol (SNMP) notifications when new health events appear. Any combination of these notification methods can be used simultaneously. Use the **Monitoring** → **Event Notifications** page in the GUI to configure event notifications.

Sending emails

Use the **Monitoring** → **Event Notifications** → **Email Server** page to configure the email server where the emails should be sent. In addition to the email server, an email subject and the senders name can also be configured. The **Test Email** action enables the user to send a test-email to an email address. See Figure 2-18.



Email notifications enabled

IP address or host name: mail.gmx.net

Port: 587

Sender's email address: my.test.account@gmx.de

Password: ••••••••••

□ Use different login:

Sender's name: Test User

Subject: &component&messageI...

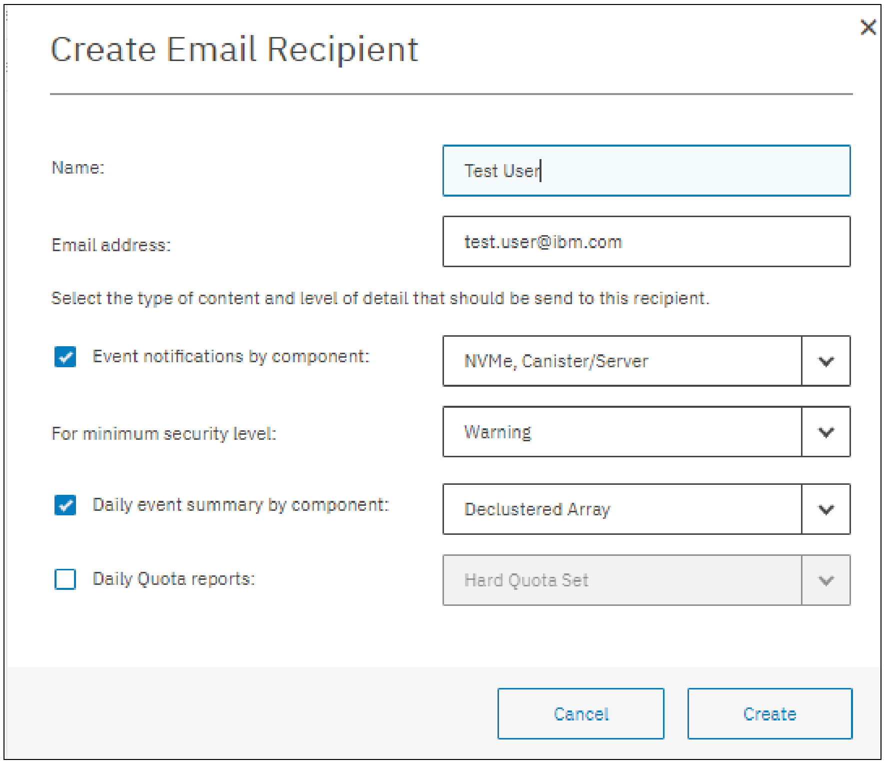
Header: header

Footer: footer

Test email address:

Figure 2-18 Configure the email server

The emails can be sent to multiple email recipients, which are defined in the **Monitoring** → **Event Notifications** → **Email Recipients** page. For each recipient, the user can select the components for which to receive emails, and the **For minimum severity level** (Tip, Info, Warning, or Error). Instead of receiving a separate email per event, optionally a daily summary email can be sent. Another option is to receive a **Daily Quota report**. See Figure 2-19.



Create Email Recipient

Name: Test User

Email address: test.user@ibm.com

Select the type of content and level of detail that should be send to this recipient.

☑ Event notifications by component: NVMe, Canister/Server

For minimum security level: Warning

☑ Daily event summary by component: Declustered Array

☐ Daily Quota reports: Hard Quota Set

Figure 2-19 Create email recipient

Sending SNMP notifications

Use the **Monitoring** → **Event Notifications** → **SNMP Manager** page to define one or more SNMP managers that receive an SNMP notification for each new event. Unlike with Email notification, filters cannot be applied to SNMP notification and an SNMP notification is sent for any health event that occurs in the system. For more information on configuring SNMP, see Configuring SNMP manager or use the GUI help topic for event notifications.

2.2.9 Dashboards

The **Monitoring** → **Dashboard** page provides an easy-to-read, single-page, real-time user interface that provides a quick overview of the system performance.

Some default dashboards are included with the product. Users can further modify or delete the default dashboards to suit their requirements and can create additional new dashboards. The same dashboards are available to all GUI users, so modifications are visible to all users.

A dashboard consists of several dashboard widgets that can be displayed within a chosen layout.

Widgets are available to display the following items, as shown in Figure 2-20:

- Performance metrics
- System health events
- File system capacity by file set
- File sets with the largest growth rate in the last week
- Time lines that correlate performance charts with health events

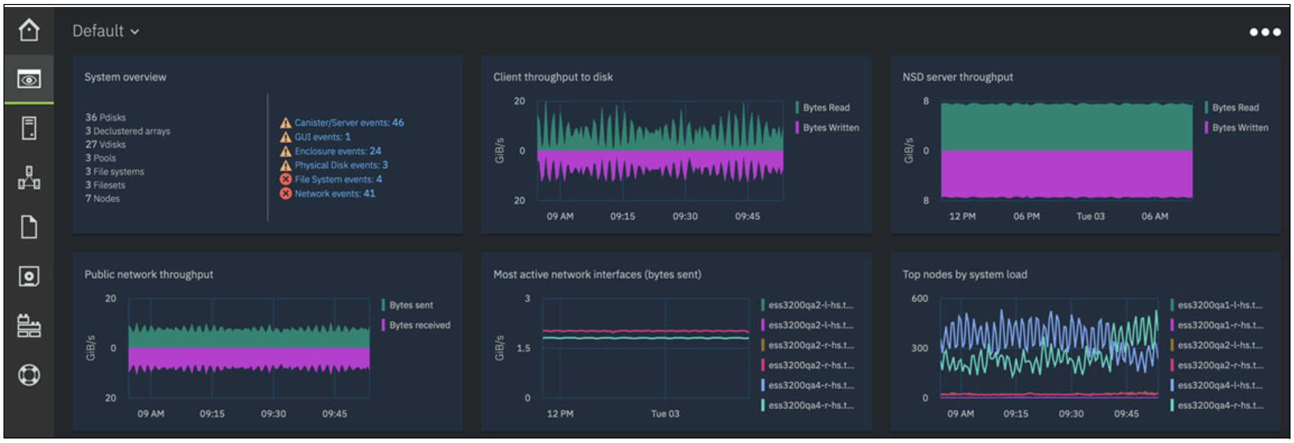


Figure 2-20 The dashboard

2.2.10 More information

The previous sections provided a rough overview of the GUI. For more detailed information on the GUI, read the *Monitoring and Managing the IBM Elastic Storage Server Using the GUI*, REDP-5471 IBM Redpaper publication and use the online help pages that are included within the GUI. Additional information is also available in IBM Documentation articles for IBM Spectrum Scale version 5.1.1.

2.3 Software enhancements

In this section, the software enhancements in ESS 3200 are described.

2.3.1 Containerized deployment

ESS solution installation and management software includes, but not limited to:

- ESS-specific documentation for installation and upgrade scripts
- A container-based deployment model that focuses on ease of use.
- Other tools for the IBM SSR to use for installing ESS, such as essutils

Third-generation ESS systems deploy a newer container-oriented management software stack in the ESS Management Server that includes Ansible playbooks for installation and orchestration.

IBM preinstalls this complete integrated, tested ESS solution stack on the ESS servers in IBM Manufacturing.

The ESS solution-stack levels are released as a version, release, modification, and fixpack level.

For more information about the release levels of the ESS software solution and the levels of the software components for that ESS release level, see IBM Documentation at:

`https://www.ibm.com/docs/en/ess-p8?topic=SSYSP8/gnrfaq.html#GNRfaqSept2016-gen4supportmatrixq`.

For more information about Containerized deployment, see IBM Documentation:

https://www.ibm.com/docs/en/ess/6.1.1_cd?topic=quick-deployment-guide

The ESS solution-stack components are periodically up-leveled, tested, and released as a new level of ESS solution software. IBM recommends that clients plan to upgrade their ESS to the current level of ESS solution software stack at least once a year.

2.3.2 Red Hat Ansible

In the ESS 3200 container, the Ansible library is included, which help to orchestrate a set of commands (also named tasks). With this capability, you can automate the deployment process into a few commands.

Figure 2-21 shows the tree of the `ansible` directory included in the container.

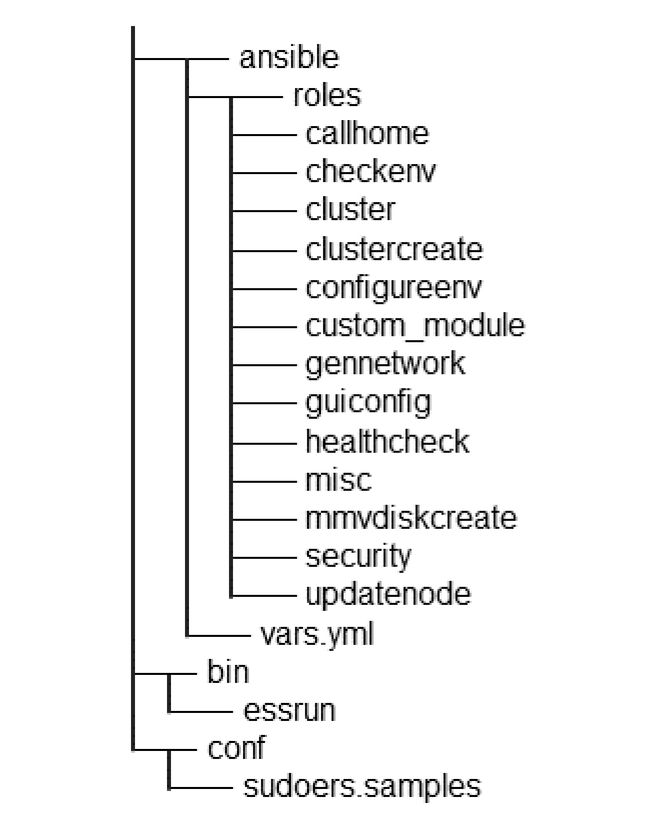


Figure 2-21 Ansible directory tree included in the container

The **`roles`** directory contains a set of folders intended to contain a set of tasks for various purposes, for example: **`configureenv`**, contains the “**`essrun config load`**” set of tasks.

If you want to import or use the roles within your own Ansible Playbook, you can import the roles and the **`vars.yml`** file, since it contains several variables used within every role.

Example 2-1 shows how to import an ESS role into your own Ansible Playbook:

Example 2-1 How to import an ESS role into an Ansible Playbook

```
---
- name: ESS config load
```

```
  hosts: all
  remote_user: root

  vars_files:
    - /opt/ibm/ess/deploy/ansible/vars.yml

  # importing roles
  tasks:
    - include_role:
        name:
/opt/ibm/ess/deploy/ansible/roles/configureenv
```

2.3.3 The mmvdisk command

Although on previous versions of ESS it was possible to use other commands to manage IBM Spectrum Scale RAID, on ESS 3200 the only way to manage IBM Spectrum Scale RAID is with the **mmvdisk** command.

The **mmvdisk** command is an integrated command suite for IBM Spectrum Scale RAID. It greatly simplifies IBM Spectrum Scale RAID administration, and encourages and enforces consistent best practices regarding server, recovery group, VDisk NSD, and file system configuration.

The **mmvdisk** command can be used to manage new IBM Spectrum Scale RAID installations. If you are integrating ESS 3200 with a setup that already has other ESS systems that are non-**mmvdisk** recovery groups, those systems must be online-converted into **mmvdisk** recovery groups before adding the ESS 3200 into the same cluster.

For more information about the **mmvdisk** command, see the following IBM Documentation:

https://www.ibm.com/docs/en/ess-p8/6.1.1?topic=commands-mmvdisk-command

2.3.4 The mmhealth command

The **mmhealth** command is aimed to be the “one stop” for all things health-related on an IBM Spectrum Scale cluster. Although a cluster might be made up of many different types of components, **mmhealth** gives a holistic status of the health of the important cluster health issues.

For any type of cluster, the holistic status includes the status of the general parallel file system (GPFS) daemons, the NODE software status, tracking of EVENTS that happened to the cluster, and the FILESYSTEM health status.

The depth of the details for one NODE depends on a few factors:

- If the node is a software-only node, where IBM Spectrum Scale formats only external block devices to the cluster.
- If the system is one that uses IBM Spectrum Scale RAID, such as the ESS 3200. In the ESS 3200 case, **mmhealth** monitors and reports the following non-exhaustive list:
  - Hardware specific, the same as other ESS hardware solutions:
    - Temperature of different sensors of the enclosure
    - Power supply hardware status
    - Fan speeds and status

    - Voltage sensors data
    - Firmware levels reporting and monitoring
    - Boot drive status and monitoring
  - IBM Spectrum Scale RAID specific, the same as other IBM Spectrum Scale RAID solutions:
    - Recovery Groups status and monitoring
    - Declustered Array status and monitoring
    - Physical drives status and monitoring
    - VDisks status and monitoring
  - IBM Spectrum Scale Software related, the same as other IBM Spectrum Scale software related:
    - NSD status and monitoring
    - Network communication status and monitoring
    - GUI status and monitoring (of the GUI nodes)
    - CES status and monitoring (off the CES nodes)
    - File system status and monitoring
    - Pool status and monitoring
    - NSD protocol and statistics, and other protocols' statistics when applicable

The **mmhealth** command provides IBM Spectrum Scale software-related checks across all node and device types present in the cluster. Software RAID checks are present across all *GPFS Native RAID* GNR offerings (such as ESS 5000, ESS 3000, ESS 3200, and ECE). For devices (such as ESS 3200) that are integrated with IBM Spectrum Scale hardware, you also get the hardware checks and monitoring.

For details about how to operate with the **mmhealth** command, see IBM Documentation at:

https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=reference-mmhealth-command

The mmhealth command changes to support ESS 3000 and ESS 3200

The **mmhealth** command, as described in "The mmhealth command" on page 27, has a specific component monitoring when an IBM Spectrum Scale Native RAID (GNR) environment is in-use. As of the June 202119 release of ESS 3200 (version 611x), **mmhealth** now supports GNR health monitoring on the 5141-FN1 solution (x86 based NVME platform).

The **mmhealth** command was extended to support the additional hardware components of the ESS 3000 and ESS 3200, and to address the needs of users who want to monitor the environment. This section initially references much of the current **mmhealth** information available through IBM Redbooks, command references, administration documents, and other publicly available resources. The second section describes the specific additional changes that are made in **mmhealth** to support ESS 3000 and ESS 3200.

Support in the mmhealth command for GNR

This section provides pointers to much of the currently available documentation regarding **mmhealth** and GNR-specific support.

The following link shows all of the current Reliability, Availability and Serviceability (RAS) events supported by **mmhealth**. These include all events supported by IBM Spectrum Scale, with a subset specific to GNR:

https://www.ibm.com/docs/en/ess-p8/6.1.1?topic=references-events

Canister events are new to ESS 3000 and ESS 3200, and are described in "Canister events" on page 29. The rest of the events are applicable to ESS (legacy) and ESS 3000 and ESS

3200. See *IBM Spectrum Scale Erasure Code Edition: Planning and Implementation Guide*, REDP-5557.

This guide describes the **mmhealth** command in the following sections:

- Section 7.7 shows general command usage
- Section 7.8 shows example use case scenarios

https://www.redbooks.ibm.com/abstracts/redp5557.html?Open

The following link is the main page for **mmhealth**. It shows the complete command usage, features, and examples:

https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=reference-mmhealth-command

Updates to mmhealth command to support ESS 3000 and ESS 3200

Several major changes were made between legacy ESS and ESS 3200. The **mmhealth** command was updated to support monitoring these new features. The following list includes some of these differences:

- Architecture change to x86_64 from Power
- Support for NVME drives
- No external storage enclosures
- Dual canister design within single building block

The **mmhealth** command includes several changes to support ESS 3000 and ESS 3200.

- The Canister events category was included to support many of the differences between legacy ESS and ESS 3200.
- The Server category was also adjusted.

Both of these adjustments and other changes to **mmhealth** are included in the following sections.

Canister events

These events are new and specifically added to support the new Canister-based building-block configuration of the ESS 3000 and ESS 3200. For information such as events related to the boot drive, temperature, CPU, and memory, see:

https://www.ibm.com/docs/en/ess-p8/6.1.1?topic=events-canister

A new command (**ess3kplt**) was created by GNR to provide CPU and memory health information to **mmhealth**:

```
/opt/ibm/gss/tools/bin/ess3kplt
```

Command usage:

```
ess3kplt -h
usage: ess3kplt [-h] [-t SELECTION] [-Y] [-v] [--local]
```

Optional arguments:

```
  -h, --help    show this help message and exit
  -t SELECTION  Provide selection keyword: [memory|cpu|all]
  -Y            Select report listing
  -v            Enable additional output
  --local       Select localhost option
```

This program can be used to inspect memory or CPU resources. Example 2-2 shows a sample output.

Example 2-2 Sample output

```
ESS3K Mem Inspection:
  InspectionPassed:          True
  Total Available Slots:     8   (expected 8)
  Total Installed Slots:     8   (expected 0 or 8)
  DIMM Capacity Errors:      0   (Number of DIMMs with a size different from ['64 GB'])
  DIMM Speed Errors:         0   (Number of DIMMs with a speed of neither 3200 MT/s nor 3200
MT/s MT/s)
  Inspection DateTime:       2021-08-31 14:02:28.338486

ESS3K Cpu Inspection:
  InspectionPassed:          True
  Total CPU Sockets:         1   (expected 1)
  Total Populated Sockets:   1   (expected 1)
  Total Enabled CPU Sockets: 1   (expected 1)
  Total Cores:               48  (expected [48])
  Total Enabled Cores:       48  (expected [48])
  Online CPUs:               ---
  Total Threads:             96  (expected 96)
  CPU Speed Errors  :        0   (Number of CPUs with a speed different from ['3300 MHz'] MHz)
  Inspection DateTime:       2021-08-31 14:02:28.562756
```

Example 2-3 shows a sample verbose output.

Example 2-3 Sample verbose output:

```
ess3kplt:memory:HEADER:version:reserved:reserved:location:size:speedMTs:
ess3kplt:memorySummary:HEADER:version:reserved:reserved:availableSlots:installedSlots:capacityEr
ror:speedError:inspectionPassed:
ess3kplt:memory:0:1:::P0_CHANNEL_D:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_C:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_B:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_A:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_E:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_F:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_G:passed:passed:
ess3kplt:memory:0:1:::P0_CHANNEL_H:passed:passed:
ess3kplt:memorySummary:0:1:::8:8:0:0:true:
ess3kplt:cpu:HEADER:version:reserved:reserved:location:speedMHz:status:status2:numCores:numCores
Enabled:numThreads:
ess3kplt:cpuSummary:HEADER:version:reserved:reserved:totalSockets:populatedSockets:enabledSocket
s:totalCores:enabledCores:totalThreads:speedErrorrs:inspectionPassed:
ess3kplt:cpu:0:1:::CPU0:passed:ok:ok:48:48:96:
ess3kplt:cpuSummary:0:1:::1:1:1:48:48:96:0:true:
```

CPU and DIMM-related events that **`mmhealth`** reports rely on the **`ess3kplt`** command in the ESS 3000 and ESS 3200 environments.

2.4 RAS enhancements

ESS 3200 is the next generation of the ESS product family that is built on a high availability and performance storage server platform in a 2U form factor.

ESS 3200 is targeted at delivering the following key traits in *Appliance customer experience*:

- Easy to order
- Easy to install
- Easy to upgrade
- Easy to use
- Easy to service

The following list includes the key components:

- IBM storage enclosure with commercial NVMe drives
- Red Hat Enterprise Linux (RHEL) 8.2 with NVMe support
- IBM Spectrum Scale 5.1.1.x software features and functions
- IBM Spectrum Scale Software RAID, also known as *GPFS Native RAID* (GNR)

ESS 3200 is a customer setup (CSU) product with a combination of customer-replaceable units (CRUs) and FRUs.

2.4.1 RAS features

ESS 3200 aims to reduce the frequency of failures, minimize workload interruptions, and easily detect, identify, and report problems to decrease service-repair time. It incorporates redundancy in its design so that component replacements do not interfere or affect system operations. To maintain availability, multiple methods and services are used to monitor various components of the system. To report this status, notifications provide detailed event information including user-action for necessary next-steps. If enabled, ESS 3200 generates call-homes for various failures that can occur on the system and provides inventory updates. This automated service processes and sends the proper diagnostic information to IBM Support servers to assist with debug and troubleshooting.

ESS 3200 RAS features consists of the following items:

- Monitoring
  - Hardware components
  - Firmware levels
  - GNR and IBM Spectrum Scale components
- Event notification
- Call home
- IBM Spectrum Scale Healthchecker
- First time data capture (FTDC)

2.4.2 Enclosure overview

ESS 3200 includes node-to-node communication through internal Ethernet private network and Non-Transparent Bridge (NTB) for peer node diagnostic and control. Remote console through Serial Over LAN (SOL) using baseboard management controller (BMC) Intelligent Platform Management Interface (IPMI) is available for monitoring and controlling of the ESS 3200 enclosure and to assist with deployment and installation.

Several methods of power control are available on the system for the canister and drive slots to assist with system recovery and troubleshooting, which helps to decrease component down-time. LED power indicators are in place on the front of the enclosure, drive carrier, fan, power module, and the canister as a visible sign that the hardware is receiving power. LED status indicators are used for the drive, fan, canister, power module, and enclosure that help point out if any components might be experiencing issues.

ESS 3200 offers a Samsung-only NVMe drive with the following capacity options at its initial GA in 2Q 2021 with either 12-drive or 24-drive installation options:

- 3.8 TB
- 7.6 TB
- 15.3 TB

ESS 3200 uses a mirrored set of 960 GB M.2 NVMe SSD drives as the boot disks. The M.2 SSD includes the *Power Loss Protection* (PLP) feature. It offers only one memory-configuration per canister, as shown in Table 2-3.

Table 2-3 Per-Canister Memory configuration

Memory	Configuration details
512 GB	Fully populated with 8 x 64 GB (8 slots for single CPU) DIMMs

Given the IBM Spectrum Scale GNR design and the M.2 PLP feature to ensure the data persistency for GNR log files maintained in the boot disks, ESS 3200 does not require a Battery Backup Unit (BBU).

When you plan to install a 100G adapter, the following adapters are available:

- AJZL: CX-6 InfiniBand/VPI in PCIe form factor
  - InfiniBand - HDR200 200 Gb / HDR100 100 Gb / EDR 100 Gb
  - Ethernet - 100 GbE
- AJZN: CX-6 DX in PCIe form factor (Ethernet - 100 GbE)

Figure 2-22 shows the front view of the ESS 3200 enclosure.

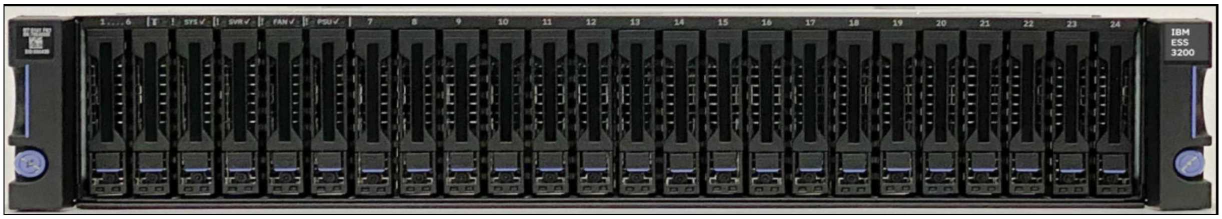


Figure 2-22 Front view of the ESS 3200 enclosure

Figure 2-23 shows the rear view of the ESS 3200 enclosure.

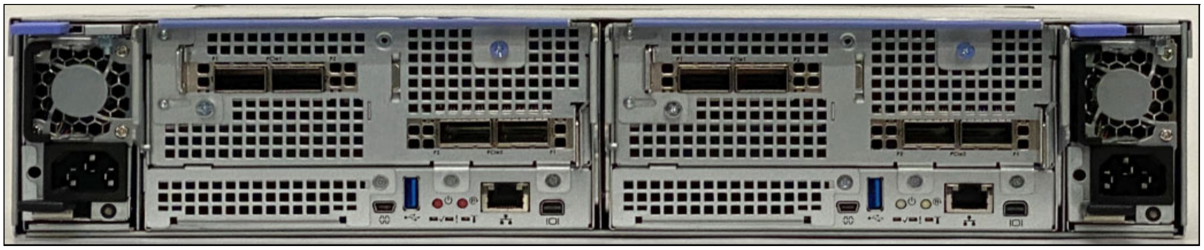


Figure 2-23 Rear view of the ESS 3200 enclosure

Figure 2-24 shows the rear view of the two canisters. The key connectors for I/O and system management are highlighted. In a typical configuration, two high-speed, dual-port adapters are included. Up to two adapters can be installed per canister. Both canisters must be identically populated with matching adapter type and slot location.

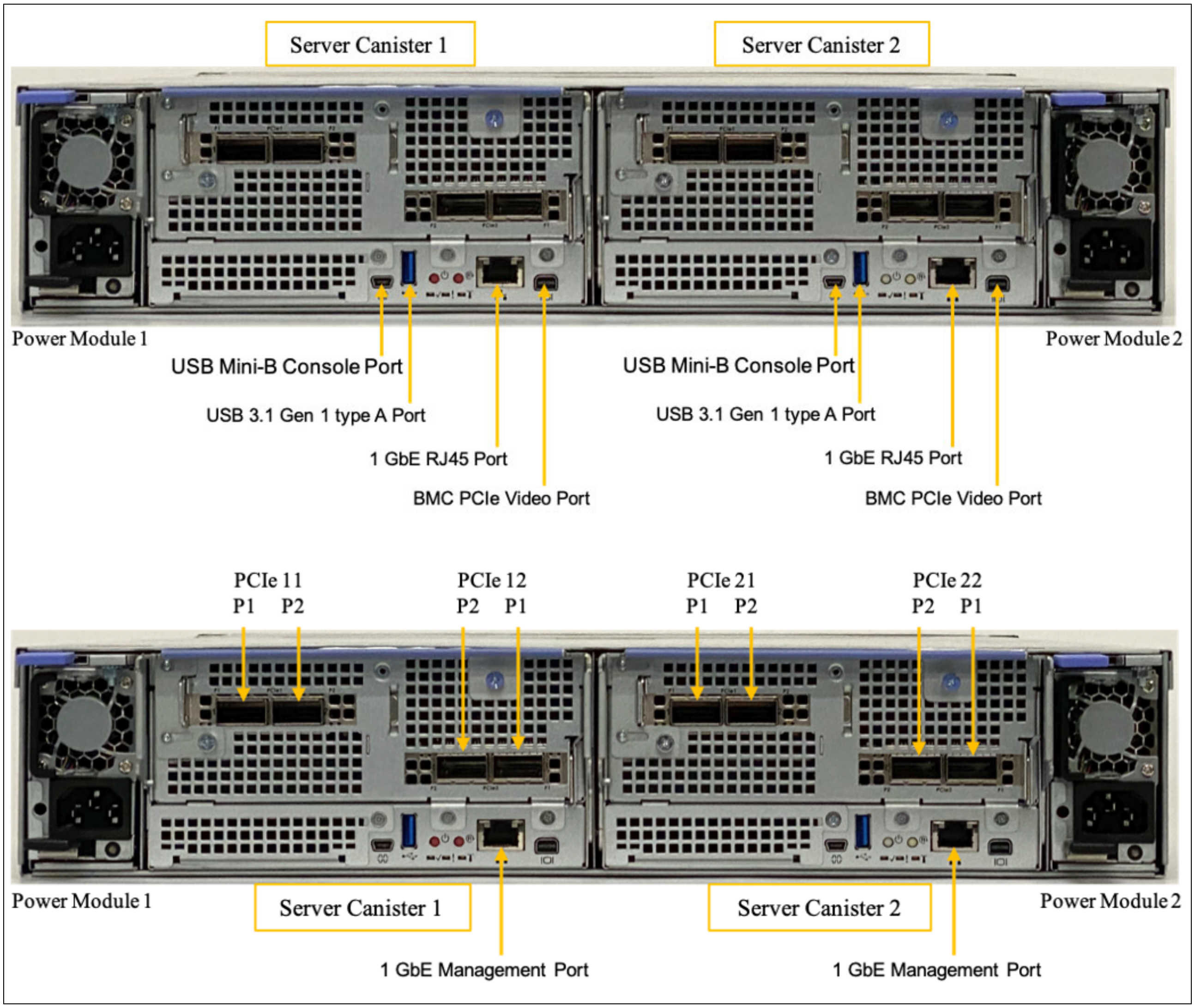


Figure 2-24 Rear view of the two canisters

2.4.3 Machine type model and warranty

ESS 3200 has a single MTM value: 5141-FN1. It includes a 3-year warranty (flex: 1-year warranty + 2 years maintenance).

ESS 3200 also offers same-day service upgrade options, and optional priced services that include lab-based services (LBS) installation.

2.4.4 Components: FRU versus CRU

Compared to previous ESS models, ESS 3200 includes fewer replaceable parts based on its dual-canister architecture. With two hot swappable I/O canister nodes, the new canister design is easier to access from the rear of the system using a simple pull-down lever mechanism. Six Fan units (5+1 redundant) located at the front of the enclosure from the top, allow for easy removal and replacement without interrupting system functionality.

In general, ESS 3200 service strategy takes the following approach:

- Any hardware component that can be accessed only by opening the cover of the canister is considered to be a FRU and therefore requires IBM service personnel to perform any relevant service action.
- Any hardware component that can be accessed for maintenance without removing the canister is recommended as a CRU where the user can perform the repair using the assistance of the ESS GUI or IBM Documentation.

The following are key hardware components that fall into the FRU category:

- Canister
- Dual in-line memory module (DIMM)
- Adapter
- M.2 boot drive

The following are key hardware components that fall into the CRU category:

- NVMe drive
- Drive filler
- Power module

2.4.5 Maintenance and service procedures

Guided procedures are available to perform service actions after a failed component is identified for repair based on the monitoring and failure isolation capability in the ESS 3200 platform and RAS software. Maintenance and service procedures are maintained in IBM Documentation for ESS 3200 at:

https://www.ibm.com/docs/en/ess/6.1.1_cd?topic=service-guide

Since IBM Documentation is a single point of reference that provides information about IBM systems hardware, operating systems, and server software, it is recommended that the user conduct a search for ESS 3200 to get the latest updates, as the online source is actively maintained. A Service Guide option is listed that brings up the appropriate service procedure. If preferred, ESS Service Guide can be downloaded using the link in the navigation.

Concurrent maintenance repair and update are available on the ESS 3200 for servicing and replacing of FRUs and CRUs in the system. This allows for maintenance to be performed on the system while it is used in normal operations. In addition, CRUs are hot-swappable components in the ESS 3200 that allow for replacement without powering down the server. Components such as fans and power modules are supported as concurrently removable and replaceable parts of the ESS 3200.

2.4.6 Software-related RAS enhancements

Managing and monitoring software components of the ESS 3200 is critical to increase the user's ability to attain the necessary detail from the command line interface to properly debug a failing component. To increase availability to physical disk, there is physical disk redundancy in the ESS 3200, where each server has one path to the data physical disks, but the paths from both servers are always active. In addition, to achieve optimal performance on the ESS, a shared recovery group layout is available that allows both servers in an ESS 3200 building-block to concurrently access all the available drives and their bandwidth. Additional information on recovery groups is provided in the next section.

FRU and Location

The `mmlsenclosure` command displays the environmental status of IBM Spectrum Scale RAID disk enclosures. The enclosure status reported by the `mmlsenclosure` command (and consequently the GUI) displays the FRU number and the location of that FRU within the enclosure to easily determine how to triage failures. Information about multiple components including fans and power modules is reported with an indication for service, if needed. For more detailed information on the `mmlsenclosure` command, see IBM Documentation:

https://www.ibm.com/docs/en/spectrum-scale-ece/5.1.1?topic=commands-mmlsenclosure-command

Enclosure components

Several component statuses are reported by the enclosure:

canister Failure of the left or right canister (for example server nodes)

cpu The failed CPUs that are associated with a canister

dimm The memory modules that are associated with a canister

fan The status of fans in the enclosure

2.4.7 Call home

The ESS 3200 calls home disk-related events. If one (or more) of the NVMe drives reports a problem from GNR and needs replacement, an event is caught by the monitoring service on the ESS Management Server (EMS) and sent to the Electronic Service Agent (ESA) for processing. ESA screens the events (prevents duplicates, for example) and automatically creates a Problem Management Report (PMR) if action is needed.

Call-home events are also generated for other hardware-related events including boot drives, fan modules, and power modules. Software call home is a supported call-home configuration in the ESS 3200. In this case, the EMS node acts as the software call-home server. The software call-home feature collects files, logs, traces, and details of certain system health events from different nodes and services in an IBM Spectrum Scale cluster.

For more detailed information about how call home works, see the Elastic Storage Server Version 6.1.1 Quick Deployment Guide:

https://www.ibm.com/docs/en/SSZL24_5K_6.1.1/pdf/ess_sg.pdf

For a complete background and overview of ESA, see:

https://www.ibm.com/docs/en/linux-on-systems?topic=tools-electronic-service-agent

2.5 Performance

Measurements in the IBM lab on a freshly-installed and fully-populated ESS 3200 achieved a sequential-read performance of over 80 GBps and a sequential write-performance of over 55 GBps when using an InfiniBand network with Remote Direct Memory Access (RDMA) enabled.

> **Note:** The performance measurements referenced here were made using standard benchmarks in a controlled environment. The actual performance might vary depending on several factors such as the interconnection network, the configuration of client nodes, and the workload characteristics.

Some factors related to ESS 3200 performance are listed in the following sections.

2.5.1 Network

Network components play a key role in the overall performance of IO operations. This section provides details of types of network hardware, associated configuration, and utilities in assessing the network device throughput.

High Speed Network Type

The first choice that must be made is to decide between configuring IBM Spectrum Scale to use Ethernet or InfiniBand. One of the desirable features of InfiniBand is its RDMA capability. With RDMA enabled, the communication between servers can bypass the operating system kernel, so the applications have lower latency and CPU utilization. With an Ethernet network that uses the TCP/IP protocol, the communications must go through the kernel stack, resulting in higher latencies than RDMA and reduced read and write bandwidths.

RDMA is available on standard Ethernet-based networks by using the RDMA over Converged Ethernet (RoCE) interface. For more information on how to set up RoCE, see *Highly Efficient Data Access with RoCE on IBM Elastic Storage Systems and IBM Spectrum Scale*, `REDP-5658`.

Post ESS 3200 installation, if required, the network type can be changed using the **`mmchnode`** command. For more information, see the mmchnode command.

For using RDMA, verify the following settings using the **`mmlsconfig`** command. These settings can be modified to the desired values by using the **`mmchconfig`** command.

- The **`verbsRdma`** option controls whether RDMA (instead of TCP) is used for NSD data transfers. Valid values are enable and disable.
- The **`verbsRdmaSend`** option controls whether RDMA (instead of TCP) and is also used for most non-data IBM Spectrum Scale daemon-to-daemon communication. Valid values are yes and no.
- The **`verbsPorts`** option specifies the device names and port numbers that are used for RDMA transfers between IBM Spectrum Scale client and server nodes. You must enable **`verbsRdma`** to enable **`verbsPorts`**.

Bandwidth Optimization when using TCP/IP

IBM Spectrum Scale achieves high IO throughput by establishing multiple connections between source and destination. This section describes the related configuration and link-aggregation algorithm that is required for optimal ESS 3200 network device bandwidth.

Multiple Connections Over TCP (MCOT)

Starting with IBM Spectrum Scale 5.1.1, you can establish multiple TCP connections between nodes (with the same daemon IP address that is used on each end) to optimize the use of network bandwidth. The number of connections is controlled through the **maxTcpConnsPerNodeConn** parameter, which can be changed by using the **`mmchconfig`** command. Valid values are 1-8, with the default of 2.

The value that is assigned to **maxTcpConnsPerNodeConn** must be defined after you consider the following factors:

- The overall bandwidth of the cluster network
- The number of nodes in the cluster
- The value that is configured for the **maxReceiverThreads** parameter
- Memory resource implications of setting a higher value for the **maxTcpConnsPerNodeConn** parameter

For more information, see Configuring MCOT.

Link Aggregation

The bonding or link aggregation can be an important factor for Ethernet TCP/IP performance. Most deployments use the Link Aggregation Control Protocol (LACP) or 802.3ad as the aggregation mode. The LACP aggregation determines the interface to use based on the hash of packet's source and destination information.

With multiple connections over TCP (MCOT), multiple TCP port numbers are used. By using LACP xmit_hash_policy=1 or layer3+4 (hash is generated using the IP and Port information of source and destination), a better chance exists of using multiple interfaces between a given pair of nodes. The load-balancing algorithm on the switch is also important to ensure better balancing across links from switch to destination.

Assessing network bandwidth

The network bandwidth can be assessed by using a tool like ***nsdperf***. For an overview and usage instructions about this tool, see IBM Spectrum Scale and IBM Elastic Storage System Network Guide, redp-5484.

2.5.2 Non-volatile memory express drives

Non-volatile memory express (NVMe) is an interface by which non-volatile storage media can be accessed through a PCIe bus. As a result of the efficiency of the protocol, NVMe generally provides better performance over alternatives, such as Serial Advanced Technology Attachment (SATA), when comparing devices that share the same underlying technology.

> **Note:** A NAND (short for `NOT AND`) flash NVMe drive has the potential for improved performance over a NAND flash SATA drive because of its more efficient bus connection and protocol improvements. For example, NVMe allows for longer command queues.

NVMe Drives I/O Completion

The time taken by NVMe drives to complete an I/O request accounts for only a portion of the overall time that it takes for IBM Spectrum Scale to complete the I/O request. To get the breakup of time spent, you can run the following command on the ESS 3200 and client nodes that are processing I/O requests.

```
/usr/lpp/mmfs/bin/mmdiag --iohist
```

At the lowest layer is the physical disk (pd) I/O times, obtained by running this command on an ESS 3200 server and looking at the NVMe drive I/O latencies. For example, in Example 2-4, 192 (512 byte) sectors were read in 125 microseconds.

Example 2-4   192 (512 byte) sectors were read in 125 microseconds

```
I/O start time RW    Buf type disk:sectorNum     nSec  time ms      tag1        tag2           Disk UID typ        NSD node   context thread […]
--------------- -- ----------- ----------------- -----  ------- --------- ------------ ------------------ --- --------------- --------- ---------[…]
19:07:22.558042  R        data   19:7219193624     192    0.125  83733675         103  C0A85216:61002B87 pd                   Pdisk     NSDThread […]
```

To look at the I/O latencies of requests at the NSD layer on the ESS 3200 server, look for srv layer I/O times. These times show I/O latencies that account for disk I/O times and NSD processing on the server.

On the ESS 3200, disk I/Os are generally faster than on the non-NVMe based ESS models. This means that the ratio of time spent in remote procedure call (RPC) overhead tends to be higher relative to the actual disk I/O times. For this reason, systems that support RDMA should enable the **verbsRdmaSend** option, so that RPCs can be handled through low latency RDMA operations.

Example 2-5 shows a 195-microsecond network shared disk (NSD) server (srv) layer I/O on an ESS 3200, which corresponds to the previously shown 125-microsecond pDisk I/O. Example 2-5 also shows 128 sectors that are shown in Example 2-4, as pDisk layer I/O accounts for additional sectors read for checksum validation.

Example 2-5   195-microsecond NSD server (srv) layer I/O on an Elastic Storage System 3200

```
I/O start time RW    Buf type disk:sectorNum     nSec  time ms      tag1        tag2           Disk UID typ        NSD node   context thread […]
--------------- -- ----------- ----------------- -----  ------- --------- ------------ ------------------ --- --------------- --------- --------- […]
19:07:22.558003  R        data    2:207962112      128    0.195  83733675         103  C0A85216:61002E71 srv  100.168.85.111 NSDWorker NSDThread […]
```

If an I/O is satisfied from the GNR cache, there will not be a corresponding pdisk level I/O as seen in the Example 2-4. Even though the drive accesses on the ESS 3200 are more efficient than comparable drive accesses on non-NVMe devices, the relative benefit of data residing in the GNR disk cache will be lower. However, an improved performance is expected as the elements can be efficiently swapped in and out of the GNR cache.

To see the latency of I/O requests from the client's perspective, look for 'cli' I/O times on the client-node in the output of ***mmdiag --iohist.*** (These times include network processing time and the time that requests wait for processing on the server.) Example 2-6 shows that for the previously shown 128 sector I/O, it took about 575 microseconds from the client's perspective.

Example 2-6   mmdiag --iohist output on the client node

```
I/O start time RW    Buf type disk:sectorNum     nSec  time ms      tag1        tag2           Disk UID typ        NSD node   context thread          […]
--------------- -- ----------- ----------------- -----  ------- --------- ------------ ------------------ --- --------------- --------- ---------------- […]
19:07:22.557911  R        data    2:207964928      128    0.575  83733675         103  C0A85216:61031045 cli   100.168.82.21 MBHandler DioHandlerThread […]
```

IBM Spectrum Scale RAID TRIM Support

Optimal *write* performance is achieved when the ESS 3200 NVMe drives are new, or after a purge (NVMe format), and this performance might degrade over time, depending on the usage, particularly the amount of free space that is left for internal-drive garbage collection. TRIM commands must be issued to a NVMe drive for the drive to be aware that the deleted space is available for garbage collection. IBM Spectrum Scale's use of the TRIM command enables optimal performance by enabling storage controllers to reclaim the free space. The IBM Spectrum Scale **mmreclaimspace** command can be used to send TRIM commands to the storage media. The TRIM feature must be enabled at the physical disk level, within a declustered array, and at the file system NSD level.

For more information, see Managing TRIM support for storage space reclamation.

2.5.3 Shared Recovery Group

A shared recovery group layout allows both canisters in an ESS 3200 building block to concurrently access all the available drives and their bandwidth to achieve optimal performance. Instead of a paired recovery group as in non-NVMe based ESS models, both servers access a single shared recovery group. This allows the servers to drive the full bandwidth that the disks are capable of in both the configurations: fully-populated (24 disk) and half-populated (12 disk). In this shared recovery-group model, two user-log groups are created per server-node, allowing I/O to be evenly distributed across both nodes.

The optimal performance is achieved when both the servers access all the drives in parallel, because a single server does not have sufficient PCIe bandwidth to drive all 24 disks at full bandwidth.

For more information, see Recovery Group Issues.

2.5.4 Tuning

ESS 3200 configuration parameters are set automatically for optimal performance at the time of install procedures. This section describes the key configuration parameters on the I/O servers and client nodes that are required to be further modified to achieve optimal performance. Modifications are based on the nature of the application I/O activities.

I/O Server tuning

Choosing the right file system block size influences the subblock sizes that would be set for both data and metadata blocks. This section explains block size settings, and procedures to verify the ESS 3200 configuration parameters.

File-system block size

Larger data block sizes traditionally help large sequential streaming I/O workloads. However, storage systems using erasure encoding might experience a write-amplification effect, when the applications do writes that are smaller than the file system block size, and when those writes are not coalesced. This can have a negative performance impact. On such storage systems, workloads for which small write-performance is an important component might see a performance improvement if the file system block-size is optimized to minimize write-amplification.

IBM Spectrum Scale version 5 introduced variable subblock sizes, making space allocations for smaller files more efficient with larger block sizes, and improving file creation and block allocation times. With variable subblock sizes, it is advised to avoid using different block sizes for data and metadata within the same file system. Setting metadata block size that is smaller than the data block size results in a larger subblock for user storage pools. This causes block-allocation time to become longer than it would as compared to the case where the block size for both metadata and data block is the same.

See the descriptions of the `-B BlockSize` parameter and the `-metadata-block-size` parameter in the help topic `mmcrfs` command. For more information, see Block Size.

Verification of server configuration

IBM ESS performance also depends on the correct IBM Spectrum Scale RAID configuration, operating system, and network tuning. The IBM ESS tuning parameters are automatically configured during the file system creation using `essrun` (which is also used for deployment and cluster creation) or by directly running the `mmvdisk server configure` command before creating IBM Spectrum Scale RAID recovery group. When debugging performance issues, it is a good idea to validate that the correct and intended configuration parameters are in place.

The tuning can be verified by using the following methods.

- The **essinstallcheck** command checks various aspects of the installation along with the IBM Spectrum Scale RAID configuration settings and tuned profile. For more information about how to run this command, see the essinstallcheck command. Review the output carefully to address any issues.
- The IBM Spectrum Scale RAID configuration values can also be checked by using **mmvdisk server configure --verify** option.

  The **--verify** option checks whether the IBM Spectrum Scale RAID configuration attributes for the node class are set to the expected values by checking the real memory and server disk topology for each of the nodes in the node class. The **`--verify`** option can also be used to check whether the IBM Spectrum Scale RAID has newer best practice configuration values applied.

  The **`mmvdisk server configure --update`** command can be used to apply newer best practice configuration values or reset the node class to the intended default configuration values.

  Example 2-7 shows an example of the **`mmvdisk server configure`** command with the **`--verify`** option and its output.

Example 2-7 The mmvdisk server configure command with the --verify option

```
# mmvdisk server configure --nc ess3200_x86_64_mmvdisk_78E4005 -verify
mmvdisk: Checking resources for specified nodes.
mmvdisk: Node class 'ess3200_x86_64_mmvdisk_78E4005' has 503 GiB total real memory
per server.
mmvdisk: Node class 'ess3200_x86_64_mmvdisk_78E4005' has a shared recovery group
disk topology.
mmvdisk: Node class 'ess3200_x86_64_mmvdisk_78E4005' has server disk topology 'ESS
3200 FN1 12 NVMe'.
mmvdisk: Node class 'ess3200_x86_64_mmvdisk_78E4005' uses 'ess3200.shared'
recovery group configuration.

daemon configuration attribute        expected value  configured value
------------------------------------  --------------  ----------------
pagepool                              324389761843    as expected
nsdRAIDTracks                         128K            as expected
nsdRAIDBufferPoolSizePct              80              as expected
nsdRAIDNonStealableBufPct             50              as expected
pagepoolMaxPhysMemPct                 90              as expected
nspdBufferMemPerQueue                 24m             as expected
nspdQueues                            64              as expected
nspdThreadsPerQueue                   2               as expected
nsdRAIDBlockDeviceMaxSectorsKB        0               as expected
nsdRAIDBlockDeviceNrRequests          0               as expected
nsdRAIDBlockDeviceQueueDepth          0               as expected
nsdRAIDBlockDeviceScheduler           off             as expected
nsdRAIDDefaultGeneratedFD             no              as expected
nsdRAIDEventLogToConsole              all             as expected
nsdRAIDMasterBufferPoolSize           2G              as expected
nsdRAIDReconstructAggressiveness      0               as expected
nsdRAIDSmallThreadRatio               2               as expected
nsdRAIDSSDPerformanceShortTimeConstant 2500000        as expected
```

```
nsdRAIDThreadsPerQueue                  16              as expected
ignorePrefetchLUNCount                  yes             as expected
maxFilesToCache                         128k            as expected
maxMBpS                                 50000           as expected
maxStatCache                            128k            as expected
nsdMaxWorkerThreads                     3842            as expected
nsdMinWorkerThreads                     3842            as expected
nsdSmallThreadRatio                     1               as expected
numaMemoryInterleave                    yes             as expected
panicOnIOHang                           yes             as expected
pitWorkerThreadsPerNode                 32              as expected
prefetchPct                             50              as expected
workerThreads                           1024            as expected
```

mmvdisk: All configuration attribute values are as expected or customized.

- The tuned profile should be set to “scale” automatically after ESS deployment. Check the current active profile using **`tuned-adm active`** command, as shown in Example 2-8.

Example 2-8 The tuned-adm active command

```
# tuned-adm active
Current active profile: scale
```

- If tuned profile is not set to “scale”, modify using **tuned-adm profile** as shown in Example 2-9.

Example 2-9 The tuned-adm profile command

```
# tuned-adm profile scale
```

- The system settings can be verified against current profile by using the **`tuned-adm verify`** command, as shown in Example 2-10.

Example 2-10 The tuned-adm verify command

```
# tuned-adm verify
Verification succeeded, current system settings match the preset profile. See
tuned log file ('/var/log/tuned/tuned.log') for details.
```

Client Tuning

After the client cluster is created in the installation phase, extra networking and performance settings can be applied. The `gssClientConfig.sh` script can be used to apply basic best practice settings for your client NSD cluster. Running this script with the **`-D`** option shows the configuration settings that it intends to set without setting them.

> **Note:** The `/usr/lpp/mmfs/samples/gss/gssClientConfig.sh` script is part of gpfs.gnr install package. Typically, the `gpfs.gnr` package is not installed on clients. Hence, you must manually copy the script to the client cluster.

Additionally, this script attempts to configure the client nodes for RDMA access by setting the following **mmchconfig** parameter values if applicable:

- verbsRdma enable
- verbsRdmaSend yes

- verbsPorts <Infiniband ports>

Client-side **pagepool** configuration influences the end-to-end performance of applications.

pagepool

Defines the amount of memory to be used for caching file-system data and metadata. Additionally, used in few non-caching operations, that is, buffers allocated for encryption buffers and DMA transfers for DIO data.

pagepool is a pinned memory region that cannot be swapped out to disk, that is, IBM Spectrum Scale will always consume at least the value of the **pagepool** attribute in the system memory. Users need to consider the memory requirements of other applications that are running on the node while determining a value for the **pagepool** attribute.

For best sequential performance, it is required that you tune the **pagepool** attribute. Increasing **pagepool** beyond this value is most beneficial for workloads (non-direct IO) that re-read the same data because more data can be cached in the pagepool.

Use the **-P** option of the `gssClientConfig.sh` script to set the **pagepool** value as follows:

```
#gssClientConfig.h -P <size in MiB> <node names>
```

3

Planning considerations

This chapter provides planning considerations for installing Elastic Spectrum Scale (ESS) 3200. It includes the following sections:

- 3.1, “Planning” on page 44
- 3.2, “Standalone environment” on page 49
- 3.3, “Mixed environment” on page 50

ESS 3200 storage building blocks are part of an overall IBM Spectrum Scale solution. Other elements of this solution might include (but are not limited to) other ESS models, Spectrum Scale client or server or protocol nodes, integration with the customer network, physical installation, racking, and cabling.

Planning for these broader factors is described in the IBM Redbooks *Introduction Guide to Elastic Storage System*, REDP-5253, Chapter 3, “IBM Elastic Storage System planning and integration”.

3.1 Planning

This chapter provides planning information specific to deploying the ESS 3200 hardware, software, networking, and ESS Management Server. Guidance is also provided on required skills and recommendations for services that you might want to consider.

3.1.1 Technical and Delivery Assessment (TDA)

When you order an ESS 3200, certain functional and non-functional requirements need to be fulfilled before the configuration can be created and the order can be entered.

A TDA is an internal IBM process that includes a technical inspection of a completed solution design. This process assures customer satisfaction and insures a smooth and timely installation. Technical Subject Matter Experts (SMEs) who were not involved in the solution design participate to determine:

- Will the ESS 3200 solution work?
- Is the implementation and plan sound?
- Will it meet customer requirements and expectations?

The two TDA processes are as follows:

- The pre-sales TDA. This is done by IBMers or IBM Business Partners using the File and Object Solution Design Engine (FOSDE) tool that can be found on the following link: `http://www.ibm.biz/FOSDesignEngine`.
- The pre-install TDA. SMEs also evaluate the customer's readiness to install, implement, and support the proposed solution. This can be done with the IMPACT tool that can be found on the following link: `https://www.ibm.com/tools/impact/`.

The two TDA processes have assessment questions, but also baseline benchmarks, that need to be performed before the order can be fulfilled. Those tools are driven by IBM sales or resellers, so they can help and direct you regarding this process.

3.1.2 Hardware planning

These requirements include the hardware solution components that are mandatory to have but are not included inside of the ESS 3200 building block (2U). Two types of these requirements are as follows:

- The requirements that must be IBM-provided, such as the management switch
- The requirements that can be either customer-provided or IBM-provided, such as the HS network or rack

Rack solution

The ESS 3200 comes with at least one rack from IBM; it can come with more if multiple building blocks (BB) are ordered. The rack also holds the ESS Management Server and the management switch. If the HS switches are ordered from IBM, those are also included in the rack.

Although the preferred option is the rack version of the ESS 3200 solution, it is possible to order ESS 3200 without the rack. If you choose to follow this path, you must verify that the rack can hold the weight of the solution, and that the power distribution units (PDUs) on the rack are the right ones for the ESS 3200 solution. In addition, you must contact IBM to configure the management switch that comes with the solution.

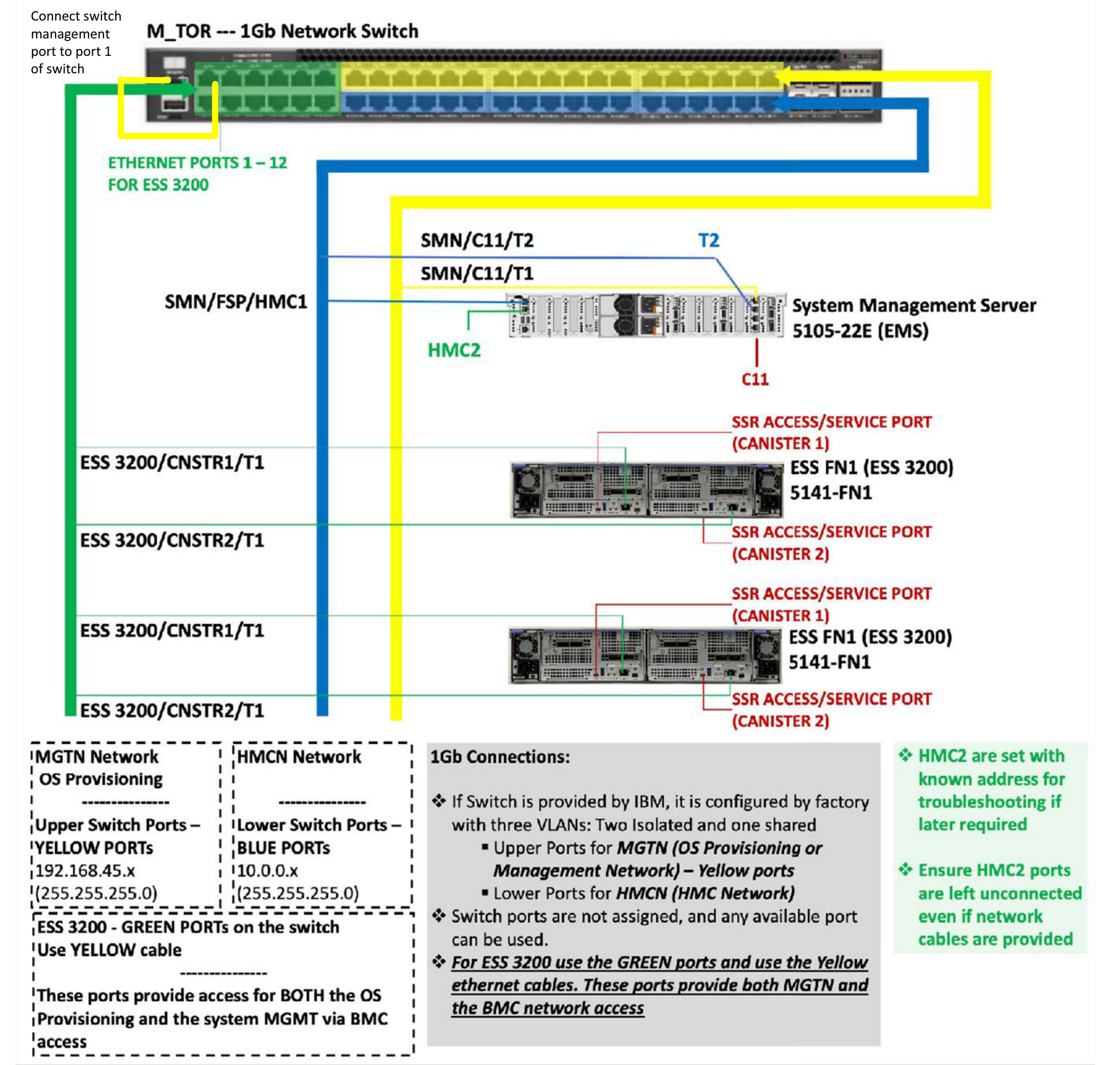
Management switch

The ESS 3200 first building block on a site includes a 1GbE management switch from IBM (8831-S52). This switch is part of the ESS 3200 solution, and it is not an independent part that can be replaced with an equivalent hardware by the customer.

Introduced with ESS 3200, is a new deployment configuration where the 1GbE management switches have a specific configuration where ports 1 - 12 are “ESS 3200" ports as shown in Figure 3-1.

Figure 3-1 is an example of the 1GbE Management Switch deployment. Deployment specifics can change from release to release. Always check the ESS Quick Deployment Guide for the information for your ESS implementation. At the time of this writing, the most recent ESS Quick Deployment Guide can be found at:

https://www.ibm.com/docs/en/ess/6.1.1?topic=quick-deployment-guide


Connect switch management port to port 1 of switch
M_TOR --- 1Gb Network Switch
ETHERNET PORTS 1 – 12 FOR ESS 3200
SMN/C11/T2
T2
SMN/C11/T1
SMN/FSP/HMC1
System Management Server 5105-22E (EMS)
HMC2
C11
SSR ACCESS/SERVICE PORT (CANISTER 1)
ESS 3200/CNSTR1/T1
ESS FN1 (ESS 3200) 5141-FN1
ESS 3200/CNSTR2/T1
SSR ACCESS/SERVICE PORT (CANISTER 2)
SSR ACCESS/SERVICE PORT (CANISTER 1)
ESS 3200/CNSTR1/T1
ESS FN1 (ESS 3200) 5141-FN1
ESS 3200/CNSTR2/T1
SSR ACCESS/SERVICE PORT (CANISTER 2)

MGTN Network
OS Provisioning

Upper Switch Ports –
YELLOW PORTs
192.168.45.x
(255.255.255.0)

HMCN Network

Lower Switch Ports –
BLUE PORTs
10.0.0.x
(255.255.255.0)

ESS 3200 - GREEN PORTs on the switch
Use YELLOW cable

These ports provide access for BOTH the OS Provisioning and the system MGMT via BMC access

1Gb Connections:

- ❖ If Switch is provided by IBM, it is configured by factory with three VLANs: Two Isolated and one shared
  - ▪ Upper Ports for ***MGTN (OS Provisioning or Management Network) – Yellow ports***
  - ▪ Lower Ports for ***HMCN (HMC Network)***
- ❖ Switch ports are not assigned, and any available port can be used.
- ❖ ***For ESS 3200 use the GREEN ports and use the Yellow ethernet cables. These ports provide both MGTN and the BMC network access***

- ❖ HMC2 are set with known address for troubleshooting if later required
- ❖ Ensure HMC2 ports are left unconnected even if network cables are provided

Figure 3-1 Overview of management network with version 2 configuration

All the ESS 3200 server canisters must be connected to ports 1 - 12 only. In case you have already a management switch that was ordered before ESS 3200 and does not have the new management switch configuration (v2), you can ask IBM TSS to convert the switch to be usable by ESS 3200 systems. You can also convert the switch yourself with the instructions in Appendix A., “Convert a previous generation switch into a version 2” on page 63.

High-speed network

As with any other IBM Spectrum Scale/ESS configuration, the ESS 3200 requires a high-speed (HS) network to be used as the data-storage cluster network. In some product documentation, this network is referred to as a *Clustering network*. The hardware for the HS network can be provided by IBM or the customer. If the hardware is provided by the customer, it must be compatible with the network interfaces that the ESS 3200 supports. See section 2.1.1, “Canisters and servers” on page 10 to see the available network options on ESS 3200.

Enterprise Management Server (EMS)

The ESS 3200 requires an POWER9 ESS Management Server (EMS), IBM machine type 5105-22E. The EMS is required on standalone installs. If the ESS 3200 is added to an existing IBM Spectrum Scale / ESS cluster that already has a POWER9 EMS, the existing EMS can support the ESS 3200.

> **Notes:** If your previously-installed IBM Spectrum Scale or ESS configuration uses a previous-generation POWER8 EMS, you must add a POWER9 EMS to support the ESS 3200.
>
> IBM does not support re-purposing an existing server or a VM or LPAR to be used as the EMS.

For more details about the EMS server, see the following IBM Documentation link:

https://www.ibm.com/docs/POWER9/p9hdx/5105_22e_landing.htm

The IBM or IBM Business Partner team uses the FOSDE tool and IBM eConfig for Cloud to configure the EMS. eConfig configures the EMS with the appropriate network cards such that the EMS can participate in the same HS networks that are configured on the ESS 3200.

The default IBM ESS Management Server memory size is enough for most IBM ESS installations. If many IBM ESSs are used in your Spectrum Scale IBM ESS configuration, check with your IBM representative to see whether larger IBM ESS Management Server memory sizes might be required for your installation. More EMS memory can be specified at order time or added later as a field Miscellaneous Equipment Specification (MES).

3.1.3 Software planning

ESS 3200 provides an integrated, tested ESS solution software stack that includes:

- Embedded Red Hat Enterprise Linux license
- Firmware drivers for the Mellanox network cards
- IBM Spectrum Scale
- All necessary supporting software

IBM supports the ESS software stack as a solution.

When performing ESS software currency, each installation cannot be done in the same way. Each installation has different operational and non-operational requirements that can impact what is possible to achieve and when and how often is possible to do software updates.

While an ideal shop would update their systems at least once a year, there are other shops where that is simply not possible. Either due legal certification reasons, operational, or other reasons. To the point that might end up on updates happening once every three or more years.

IBM strongly recommends the following key points that should be followed when doing software currency on ESS-related environments:

- Never do more than N-3 jump of an ESS-software update. Do intermediate jumps if needed to maintain this rule.
- Always update the EMS first.
- Prefer offline over online updates. If online update is a requirement, explore the **`-serial`** option to limit the risk exposure in case some nodes experience problems during the update.
- If you encounter a problem, involve IBM Service/Support. While you do have root access, changing some things might fix your problem today but create future issues due to the automation expecting the configuration to be a certain way. So, stabilize the environment, involve support, and continue on another day.
- Always keep the ESS cluster in the same level. You can update different systems, but goal-level should be the same. If that is not possible, think about partitioning your backend cluster to achieve this rule.
- Use defaults, unless you have an empirical reason backed up with data to not do so.

3.1.4 Network planning

The ESS system includes certain network names that might or might not be familiar to you. To avoid confusion, those networks are defined in this section.

Management network

The *management network* is a non-routable private network. It connects the EMS PCI card slot 11 (C11) port 4 (T1), which acts as a DHCP server to all I/O nodes on C11 T1.

Through this network, EMS, and containers on the EMS, manage the OS of the I/O nodes. This network cannot use VLAN tagging of any kind, so it must be configured as an access VLAN on the switch. You can choose any netblock that fits your needs, but as a best practice use a /24 block. If you have no preference, use the 192.168.45.0/24 block because it is the one that is used in this paper and most of the documentation examples.

Flexible service processor network

The *flexible service processor (FSP) network* is a non-routable private network. It connects the EMS C11 T2 with each of the I/O nodes out of band management ports that are labeled as “HMC 1". That includes the EMS that has a connection to the HMC1 from this network and any other Power-based ESS node in the cluster managed by the same EMS.

The EMS and the containers running on the EMS use this network to do FSP and baseboard management controller (BMC) operations on the physical servers, which include powering-on and powering-off the servers and many other operations. This network uses VLAN tagging at the ESS 3200 ports and no tagging on the rest of ports. You can choose any netblock that fits your needs, but as a best practice use a /24 block. If you do not have a preference, use 172.16.0.0/24 for the same reasons as described for the management network.

High-speed network

The HS data network is where the IBM Spectrum Scale daemon and admin networks should be configured. It is a customer-provided and managed network.

Network design for parallel file system is not a simple topic. The HS network design and implementation is usually the deciding factor on what the overall performance your system delivers. Unfortunately, there is no silver-bullet design that fits every use case.

Some design ideas that must be considered are as follows:

- The IBM Spectrum Scale admin network has these characteristics:
  - Used for the running of administrative commands.
  - Requires TCP/IP.
  - Can be the same network as the IBM Spectrum Scale daemon network or a different one.
  - Establishes the reliability of IBM Spectrum Scale.
- The IBM Spectrum Scale daemon network has these characteristics:
  - Used for communication between the `mmfsd` daemon of all nodes.
  - Requires TCP/IP.
  - In addition to TCP/IP, IBM Spectrum Scale can be optionally configured to use Remote Direct Memory Access (RDMA) for daemon communication. TCP/IP is still required if RDMA is enabled for daemon communication.
  - Establishes the performance of IBM Spectrum Scale, as determined by its bandwidth, latency, and reliability of the IBM Spectrum Scale daemon network.

In cases where the HS data network is Ethernet based, both the daemon and admin network should be on the HS Ethernet network (unless you have good reasons not to do so).

If you have InfiniBand networks, you can use Ethernet adapters on the HS network if they are available, or you can use an IP over InfiniBand encapsulation.

The Management or FSP network should not be part of the IBM Spectrum Scale cluster as a management or daemon network.

With the networking information described in this section, perform the following sizing exercise:

> Expected client-required throughput performance and number of client-ports with their aggregated performance versus the number of ESS 3200 or other ESS I/O nodes or canister aggregated performance.

Include any inter switch links (ISL) that are in place as well as PCIe speeds and feeds for each system. As an example, consider a simple two HS IB 200 Gbit ports scenario, where each ESS 3200 includes eight ports connected. Assuming there are PCI3 Gen 4 x16 lines on the clients and 200 Gbit IB ports connected (high dynamic range (HDR)), it is an ideal scenario to have up to eight of those clients and four ISLs between switches. From there, some network oversubscription occurs that might affect your workload.

IBM SSR network port

The IBM System Services Representative (SSR) network port on the IBM EMS is on C11 port T4. This port should never be cabled or connected to any switch because it is only for IBM field engineers to use. This port is configured on the 10.111.222.100/30 block.

The ESS canister ports do not have Ethernet connectivity to connect to the canister. Therefore, the IBM SSR accesses the device through a serial cable on each canister.

3.1.5 Skills and Services

Skills required to install and support ESS 3200 include administration skills for:

- Red Hat Enterprise Linux
- TCP/IP and high speed networking
- IBM Spectrum Scale

IBM and IBM Business Partners can provide education courses and services to teach these skills.

Customers and IBM Business Partners can engage IBM Systems Lab Services, which are available and recommended to provide help in integrating ESS 3200 into your client environment.

3.2 Standalone environment

This section describes best practices for deploying and optimizing a standalone ESS 3200.

A standalone ESS 3200 unit, which is known as a *building block*, must minimally consist of the following components:

- One EMS node in a 2U form factor
- One ESS 3200 node in 2U form factor
- 1 GbE Network switch for management network (1U)
- 100/200 Gb high speed IB or Ethernet network for internode communication (1U)

The EMS node acts as the administrative end point for your ESS 3200 environment. It performs the following functions:

- Hosts the Spectrum Scale GUI
- Hosts Call-Home services
- Hosts system health and monitoring tools.
- Manages cluster configuration, file system creation, and software updates
- Acts as a cluster quorum node

The ESS 3200 features a brand-new container-based deployment model that focuses on ease-of-use. The container runs on the EMS node. All of the configurations tasks that were performed by the `gssutils` utility in legacy ESS are now implemented as Ansible Playbooks that are run inside of the container. These playbooks are accessed using the `essrun` command.

The `essrun` tool handles almost the entire deployment process, and is used to install software, apply updates, and deploy the cluster and file system. Only minimum initial user input is required, and most of that is covered by the TDA process before setting up the system. The `essrun` tool automatically configures system tuneables to get the most out of a single ESS

3200 system. File system parameters and IBM Spectrum Scale RAID Erasure code selection can be customized from their defaults before file system creation.

For more information about deployment customization, see the *ESS 3200 Quick Deployment Guide*:

https://www.ibm.com/docs/en/ess/6.1.1_cd?topic=quick-deployment-guide

The following are some of the recommended practices:

- Refrain from running admin commands directly on the ESS 3200 I/O canisters. Use the EMS node instead.
- Do not mount the file system on the ESS 3200 I/O canisters because this consumes additional resources. The file system must be mounted on the EMS node for the GUI to function properly.
- To access the file system managed by the ESS 3200 building block, you must use external GPFS client nodes or protocol nodes.
- On a single building block deployment, the I/O canister nodes are specified as GPFS cluster/file system manager nodes while the EMS node is not. Although the EMS node is considered the building block's primary management server, avoid specifying the EMS node as a manager node. The GPFS-management role is an internal designation that used to be the manager of the cluster and the file system and does not directly affect the function of the EMS node.

3.3 Mixed environment

This section provides information about integrating ESS 3200 into an existing ESS environment. Also described are the considerations to take into account when you integrate into a mixed-vendor environment for migration purposes, or for using ESS 3200 as the HS storage tier.

3.3.1 Adding ESS 3200 to an existing ESS cluster

The following guidance is for adding a Standalone ESS 3200 building block into an existing ESS cluster or into an existing ESS 3000 or ESS 5000.

Prerequisites and assumptions

Prerequisites and assumptions for adding an ESS 3200 to an existing ESS cluster are as follows:

- Existing ESS or ESS 3000 cluster is connected or reachable to the same HS network block.
- Existing ESS or ESS 3000 or ESS 5000 or new ESS 3200 is connected or reachable to the same management low-speed network block.
- ESS 3000 is configured with a POWER8 EMS node and running 6.1.1.1 Podman container.
- ESS 5000 contains a POWER9 EMS node and it is running 6.1.1.1 Podman container.
- ESS 3200 nodes were added to `/etc/hosts` and it is common across POWER8 EMS and POWER9 EMS.
  - Low-speed names: fully qualified domain names (FQDNs), short names, and IP addresses

  - High-speed names: FQDNs, short names, and IP addresses (add suffix of low-speed names)
- Host name and domain is set in POWER9 EMS
- Latest code for ESS 3000 and ESS 5000 stored in `/home/deploy` on POWER8 and POWER9 EMS
- Linux root password is common across all of the nodes (Legacy, ESS 3000, ESS 5000, and ESS3200)

Adding ESS 3200 to ESS Legacy cluster

Run the `config load` command within ESS 3200 container that is running in the POWER9 EMS to fix the SSH keys across all of the nodes. See Example 3-1.

Example 3-1 Run config load with ESS Legacy

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N
ESS3200Node1,ESS3200Node2,GSSNode1,GSSNode2,ESS3200EMSNode,GSSEMSNode config load
-p RootPassword
```

Create bonds in ESS 3200 building block within ESS 3200 container that is running in the POWER9 EMS. See Example 3-2.

Example 3-2 Create network bonds

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N
ESS3200Node1,ESS3200Node2,ESS3200EMSNode network --suffix=Suffix
/
```

Add ESS 3200 I/O nodes to the existing cluster from ESS5000Node1. See Example 3-3.

Example 3-3 Add ESS 3200 nodes to ESS Legacy cluster

```
[root@ESS3200Node1~]# essaddnode -N ESS32000Node1,ESS3200Node2 --cluster-node
GSSEMSNode --nodetype ess3200 --suffix=Suffix --accept-license --no-fw-update
```

Add ESS 3200 EMS node (Example 3-4) to the existing cluster from ESS3200Node1. See Example 3-4.

Example 3-4 Add Power 9 EMS to ESS Legacy cluster

```
[root@ESS32000Node1~]# essaddnode -N <Power 9 EMS> --cluster-node ESSLegacyNode1
--nodetype ems --suffix=Suffix --accept-license --no-fw-update
```

Adding ESS 3200 to ESS 5000 cluster

Run the `config load` command within the ESS 3200 container that is running in the POWER9 EMS to fix the SSH keys across all of the nodes. See Example 3-5.

Example 3-5 Run config load with ESS 3200

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N
ESS3200Node1,ESS3200Node2,ESS5000EMSNode,ESS5000Node1,ESS5000Node2 config load -p
RootPassword
```

Create bonds in ESS 3200 building block within ESS 3200 container that is running in the POWER9 EMS. See Example 3-6.

Example 3-6 Create network bonds

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N ESS3200Node1,ESS3200Node2 network
--suffix=Suffix
```

Add ESS 3200 I/O nodes to the existing ESS 5000 cluster from within ESS 3200 container that is running in the POWER9 EMS. See Example 3-7.

Example 3-7 Add ESS 3200 nodes to ESS 5000 cluster

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N ESS5000Node1 cluster --add-nodes
ESS3200Node1,ESS3200Node2 --suffix=Suffix
```

Adding ESS 3200 to ESS 3000 cluster

Run the `config load` command within ESS 3200 container that is running in the POWER9 EMS to fix the SSH keys across all of the nodes. See Example 3-8.

Example 3-8 Run config load with ESS 3000

```
ESS 3000 CONTAINER [root@cems0 /]# essrun -N
ESS3200Node1,ESS3200Node2,ESS3200EMSNode, ESS3000Node1,ESS3000Node2,ESSP8EMSNode
config load -p RootPassword
```

Create bonds in ESS 3200 building block within ESS 3200 container that is running in the POWER9 EMS. See Example 3-9.

Example 3-9 Create network bonds

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N
ESS3200Node1,ESS3200Node2,ESS3200EMSNode network --suffix=Suffix
```

Add ESS 3200 I/O nodes to existing the ESS 3000 cluster from within ESS 5000 container that is running in the POWER9 EMS. See Example 3-10.

Example 3-10 Add ESS 5000 nodes to ESS 3000 cluster

```
ESS 3000 CONTAINER [root@cems0 /]# essrun -N ESS3000Node1 cluster --add-nodes
ESS3200Node1,ESS3200Node2 --suffix=Suffix
```

Add ESS 3200 EMS node to existing ESS 3000 cluster from within ESS 5000 container that is running in the POWER9 EMS. See Example 3-11.

Example 3-11 Add ESS 3200 EMS to ESS 3000 cluster

```
ESS 3000 CONTAINER [root@cems0 /]# essrun -N ESS3000Node1 cluster --add-ems
ESS3200EMSNode --suffix=Suffix
```

Adding ESS 3200 to mixed ESS Legacy + ESS 3000 cluster + ESS 5000 cluster

Run the `config load` command within ESS 3200 container that is running in the POWER9 EMS to fix the SSH keys across all of the nodes. See Example 3-12.

Example 3-12 Run config load with ESS Legacy + ESS 3000 + ESS 5000

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N
ESS32000Node1,ESS32000Node2,ESS3200EMSNode,
ESS5000Node1,ESS5000Node2,ESS3000Node1,ESS3000Node2,GSSNode1,GSSNode2,GSSEMSNode
config load -p RootPassword
```

Create bonds in ESS 5000 building block within ESS 5000 container that is running in the POWER9 EMS. See Example 3-13.

Example 3-13 Create network bonds

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N
ESS3200Node1,ESS32000Node2,ESS32000EMSNode network --suffix=Suffix
```

Add ESS 3200 I/O nodes to existing ESS Legacy + ESS 3000 cluster + ESS 5000 from within ESS 3200 container that is running in the POWER9 EMS. See Example 3-14.

Example 3-14 Add ESS 3200 nodes to ESS Legacy + ESS 3000 cluster

```
ESS 3200 CONTAINER [root@cems0 /]# essrun -N ESS3000Node1 cluster --add-nodes
ESS3200Node1,ESS3200Node2 --suffix=Suffix
```

3.3.2 Scenario-1: Using ESS 3200 for metadata network shared disks

This section describes how to use ESS 3200 for metadata network shared disks (NSDs) for the existing file system

This scenario starts with an existing ESS 5000 cluster and file system deployed from a POWER9 EMS.

The steps for the high-level plan are as follows:

1. Deploy the ESS 3200 container into the POWER9 EMS.
2. Add the ESS 3200 Building Block to the cluster.
3. Create the ESS 3200 VDisk set as metadataOnly.
4. Add the ESS 3200 VDisk set to the existing ESS 5000 file system.

The following steps provide guidance to set up the ESS 3200 for metadata NSDs for the existing file system:

1. Deploy the ESS 3200 container into the POWER9 EMS.

   Customer logs in to the POWER9 EMS and completes the “Common Installation Instructions” from the Quick Deployment Guide at:

   https://www.ibm.com/docs/en/ess/6.1.1_cd?topic=guide-ess-common-installation-instructions

2. Add the ESS 3200 Building Block to the cluster.

   From step 1, within container run the ansible **essrun** command to add the new ESS 3200 to the ESS 5000 cluster (essio1 is equal to an existing ESS I/O node in the cluster):

   ```
   root@cems0:/ # essrun -N essio1 cluster --add-nodes CommaSeparatedNodesList --suffix=-hs
   ```

3. Create the ESS 3200 VDisk set as metadataOnly.

   Within container run the ansible **essrun** command to create the new ESS VDisk set (using 16M as block size, since ESS 5000 file system is the default one):

   ```
   root@cems0:/ # essrun -N ess32001a,ess32001b vdisk --name newVdisk --bs 16M
   --suffix=-hs --extra-vars "--nsd-usage metadataOnly --storage-pool system"
   ```

4. Add the ESS 3200 VDisk set to the existing ESS 5000 file system.

From the POWER9 EMS node, use **mmvdisk** command to add the new VDisk to the existing file system:

```
[root@p9ems ~]# mmvdisk filesystem add --file-system filesystemName
--vdisk-set vs_newVdisk
```

3.3.3 Scenario-2: Using ESS 3200 to create a new file system

To create a file system with IBM ESS 3200, the IBM ESS 3200 container must be running.

After the container is running and the cluster and recovery groups are created, the user can create the file system by running the **essrun** command:

```
$ essrun -N essio1,essio2 filesystem --suffix=-hs
```

Note: This command creates vdisk sets, NSDs, and file systems by using **mmvdisk**. The defaults are 4M blocksize, 80% set size, and 8+2p RAID code. These values can be customized by using additional flags.

For CES deployment, the IBM ESS 3200 system should have a CES file system. To create the CES file system, run the following command:

```
$ essrun -N essio1,essio2 filesystem --suffix=-hs --name cesSharedRoot --ces
```

Note: A CES and other file systems can coexist on the same IBM ESS cluster.

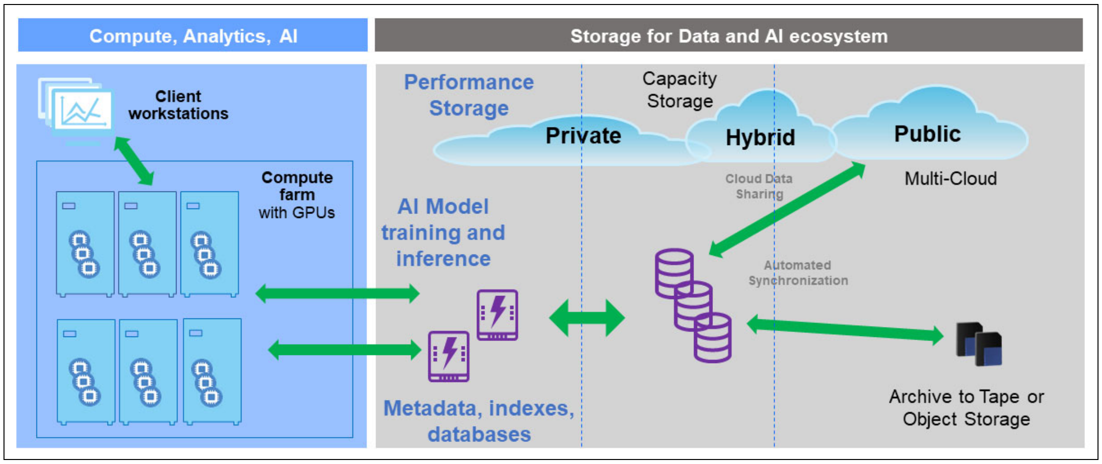
4

Use cases

This chapter discusses Elastic Spectrum Storage (ESS) 3200 use cases. It includes the following topics:

4.1 Introduction to performance storage use cases

Across industries and geographies, high-performance storage use cases can appear to vary significantly. However, upon looking closely at today's Data and AI applications, a generalized view exists of where high-performance storage fits into today's Data and AI storage infrastructure. See Figure 4-1.



Compute, Analytics, AI
Storage for Data and AI ecosystem
Client workstations
Compute farm with GPUs
Performance Storage
Capacity Storage
Private
Hybrid
Public
Multi-Cloud
Cloud Data Sharing
AI Model training and inference
Automated Synchronization
Metadata, indexes, databases
Archive to Tape or Object Storage

Figure 4-1 Performance storage use case positioning

Data and AI use cases require not only high-performance storage, but also:

- An **ecosystem** of dynamic, scalable, reliable, high-performance storage
- A performance-storage tier that delivers GBps to TBps performance to drive GPUs and modern compute
- Performance storage that must also seamlessly integrate as part of an enterprise data fabric that also has capacity tiers for:
  - Enterprise data repositories
  - Scalable flexible Hybrid Cloud tiers
  - Cost-effective deep archive Tape and Object tiers

The following sections describe how ESS 3200 as performance storage solves essential Data and AI application use cases for AI model training, inference, metadata, indexes, and databases. ESS 3200 is also described as a seamless, integrated data component in a larger Storage for Data and AI ecosystem based on IBM Spectrum Scale.

4.1.1 ESS 3200 as part of a larger Storage for Data and AI ecosystem

As we look further at the performance-storage use case storage ecosystem, ESS 3200 is positioned as high-performance storage system within the performance tier.

More importantly, ESS 3200 is part of a larger family of IBM Storage solutions that comprehensively covers all aspects of the Storage for Data and AI ecosystem, as shown in Figure 4-2.

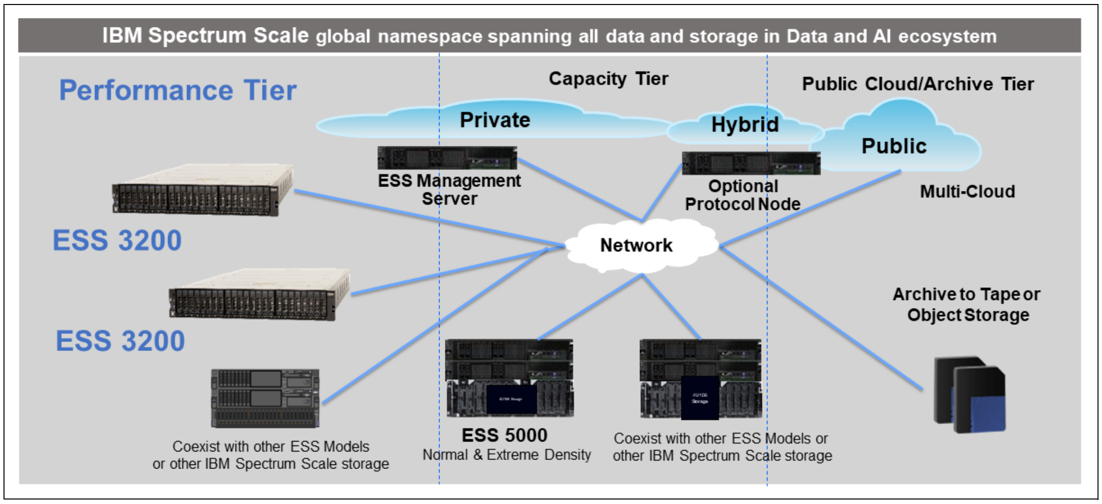

IBM Spectrum Scale global namespace spanning all data and storage in Data and AI ecosystem
Performance Tier
Capacity Tier
Public Cloud/Archive Tier
Private
Hybrid
Public
ESS Management Server
Optional Protocol Node
Multi-Cloud
ESS 3200
Network
ESS 3200
Archive to Tape or Object Storage
Coexist with other ESS Models or other IBM Spectrum Scale storage
ESS 5000
Normal & Extreme Density
Coexist with other ESS Models or other IBM Spectrum Scale storage

Figure 4-2 ESS 3200 positioning within data and AI storage ecosystem

Because ESS 3200 is part of a larger storage ecosystem, you can start small, and then grow your Data and AI ecosystem to enterprise levels, all non-disruptively. With ESS 3200, you can:

- Start small as standalone, single ESS 3200 high-performance system.
- You can add additional ESS 3200s to expand the performance tier.
- You can add one or multiple IBM ESS 5000s as HDD high-capacity tier.
- You can add flexibly by adding other IBM and non-IBM Storage components to the storage ecosystem, including hybrid cloud capacity, and archive capacity to tape or object storage.

All this can be done because ESS 3200 is an IBM Spectrum Scale storage system. IBM Spectrum Scale provides a global namespace across all the physical storage and data under its control. IBM Spectrum Scale provides the ability to non-disruptively, add, expand, grow, and modify the Data and AI storage ecosystem as needed.

ESS 3200 is part of a larger set of Data and AI use cases that provide an end-to-end enterprise data-fabric and data-management storage solution. With ESS 3200 and other IBM Storage for Data and AI solutions, you can start small, seamlessly expand, and grow your Storage for Data and AI ecosystem in many flexible ways, to enterprise levels. This is all powered by the IBM Spectrum Scale single global namespace which spans all storage tiers.

4.1.2 Representative ESS 3200 performance storage use cases

ESS 3200 is designed to provide scalable, reliable, dense, fast storage for the performance storage tier. Representative ESS 3200 use cases are listed in Figure 4-3.

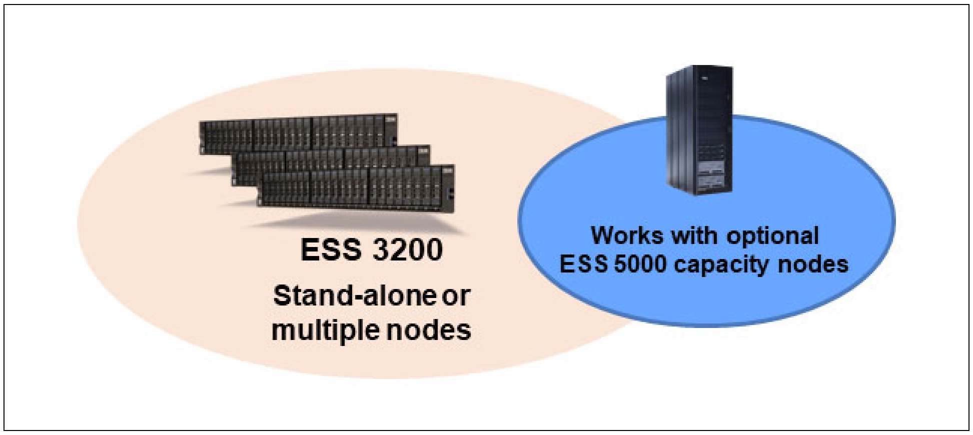

ESS 3200
Stand-alone or multiple nodes
Works with optional ESS 5000 capacity nodes

Figure 4-3 ESS 3200 representative use cases

Typical use cases for ESS 3200 include specific performance tier High-Performance Computing (HPC), AI, analytics, or other high-performance workloads with demanding requirements such as:

- AI applications requiring high-performance data effectively exploiting GPU technology at high resource utilization
- Acceleration of scale-out applications with dense NVMe Flash technology
- Information Lifecyle Management and data-tiering management of data in new or existing IBM Spectrum Scale environments
- Metadata acceleration, indexes, database acceleration
- High-performance storage at the edge

The following sections of this chapter explore some of these use cases.

4.2 Metadata and High Speed Data Tier

In an IBM Spectrum Scale cluster, performance of the entire cluster can be accelerated by placing IBM Spectrum Scale and other metadata on ESS 3200. Thus, in a High Performance computing environment, a predominant use case for ESS 3200 is to provide the high-performance metadata storage for IBM Spectrum Scale, exploiting the extremely high throughput of NVMe flash storage

Metadata generally refers to *data about data*, and in the context of IBM Spectrum Scale *metadata* refers to various on-disk data structures that are necessary to manage user data. Directory entries and inodes are defined as metadata, but at times the distinction between data and metadata might not be obvious.

For example, in the case of a 4 KB inode, although the inode might contain user data, the inode is still classified as IBM Spectrum Scale metadata. The inode is placed in a metadata pool if data and metadata are separated. Another example is the case of directory blocks, which are classified as metadata but also contain user file and directory names.

In many high-performance use cases, performance improvements might be obtained by placing IBM Spectrum Scale metadata to a fast tier, which can be accomplished by placing faster ESS 3200 storage in its own storage pool. For details about how to implement this technique, see 3.3.2, “Scenario-1: Using ESS 3200 for metadata network shared disks” on page 53.

This approach to metadata tiering can be adopted when trying to optimize the performance of metadata operations, such as listing directories and making **`stat()`** calls on files. For more information, see IBM Documentation for IBM Spectrum Scale on User Storage Pools.

Another alternative tiering approach involves, instead of tiering data based on a data or metadata classification, using the IBM Spectrum Scale File Heat function to migrate data between storage pools based on how frequently data is accessed. For more details on this approach, see IBM Documentation for IBM Spectrum Scale File Heat: Tracking File Access Temperature.

4.3 Data feed to GPUs for massive AI data acceleration

Another predominant use case for ESS 3200 is to provide the high-performance storage throughput required to feed modern GPU data acceleration systems for real time AI and Machine Learning. NVIDIA and IBM have created a reference architecture for NVIDIA DGX and ESS 3200 working together on AI and machine learning (ML) workloads. Together, NVIDIA and IBM provide an integrated, individually scalable compute and storage solution with end-to-end parallel throughput from flash to GPU for accelerated DL training, and inference. The reference architecture can be found on the following links:

https://www.ibm.com/downloads/cas/MJLMALGL
https://www.ibm.com/downloads/cas/MNEQGQVP

These reference architectures provide a proven blueprint for enterprise leaders, solution architects, and other readers who are interested in learning how the IBM Spectrum Storage for AI with NVIDIA DGX systems simplifies and accelerates AI. The scalable infrastructure solution integrates the NVIDIA DGX systems with IBM Spectrum Scale GPU Direct file storage software, which powers the IBM ESS family of storage systems that includes the new IBM ESS 3200.

The reference architectures describes the linear growth of the AI or ML system from both of the GPU workloads on the NVIDIA DGX GPU data compute acceleration systems. The IBM Spectrum Scale ability is used to deliver the linear growth in throughput, scaling linearly the maximum of 80 GBps read-throughput for each ESS 3200.

In the market for AI and ML workloads, any GPU or data acceleration, AI, and ML workload can benefit from the outstanding performance capabilities of the ESS 3200 system. For more information on using ESS 3200 with high performance GPU, see:

https://community.ibm.com/community/user/storage/blogs/douglas-oflaherty1/2021/06/22/ibm-nvidia-team-on-supercomputing-scalability

For more information on the IBM Spectrum Scale GPUDirect Storage (GDS) Technical Preview, see:

https://www.ibm.com/support/pages/node/6444075

4.4 Other use cases

ESS 3200 runs IBM Spectrum Scale as its file system, so some use cases and planning that apply to other members of the IBM Spectrum Scale family also apply for ESS 3200.

4.4.1 IBM Spectrum Scale with big data and analytics solutions

IBM Spectrum Scale is flexible and scalable software-defined file storage for analytics workloads. Enterprises around the globe deploy IBM Spectrum Scale to form large data lakes and content repositories to perform high-performance computing (HPC) and analytics workloads. IBM Spectrum Scale is known to scale performance and capacity without bottlenecks.

Cloudera is a leader in Hadoop and Spark distributions. Cloudera addresses the needs of data-at-rest, powers real-time customer applications, and delivers robust analytics that accelerate decision-making and innovation. IBM Spectrum Scale solves the challenge of explosive growth of unstructured data against a flat IT budget. IBM Spectrum Scale provides unified file and object software-defined storage for high-performance, large-scale workloads, and it can be deployed on-premises or in the cloud. Refer to *Cloudera Data Platform Private Cloud Base with IBM Spectrum Scale*, REDP-5608.

IBM Spectrum Scale is Portable Operating System Interface (POSIX) compatible, so it supports various applications and workloads. By using IBM Spectrum Scale Hadoop Distributed File System (HDFS) Transparency Hadoop connector, you can analyze file and object data in place, without data transfer or data movement. Traditional systems and analytics systems use and share data that is hosted on IBM Spectrum Scale file systems.

Hadoop and Spark services can use a storage system to save IT costs because special-purpose storage is not required to perform the analytics. IBM Spectrum Scale features a rich set of enterprise-level data management and protection features. These features include snapshots, information lifecycle management (ILM), compression, and encryption, all of which provide more value than traditional analytic systems do. For more information, see *IBM Spectrum Scale: Big Data and Analytics Solution Brief*, REDP-5397.

4.4.2 Genomics Medicine workloads in IBM Spectrum Scale

IT administrators, physicians, data scientists, researchers, bioinformaticians, and other professionals who are involved in the genomics workflow need the right foundation to achieve their research objectives efficiently. At the same time, they want to improve patient care and outcomes. Thus, it is important to understand the different stages of the genomics workload and the key characteristics of it.

Advanced genomics medicine customers are outgrowing network-attached storage (NAS). The move from a traditional NAS system or a modern scale-out NAS system to a parallel file system like IBM Spectrum Scale requires a new set of skills. Thus, the IBM Spectrum Scale Blueprint for Genomics Medicine Workloads must provide basic background information. It must also offer optional professional services to help customers successfully transition to the new infrastructure.

For more information, see the following document:
`http://www.redbooks.ibm.com/abstracts/redp5479.html`

4.4.3 IBM Spectrum Scale-to-Cloud Object Storage and NFS data sources

You can combine ESS 3200 in an IBM Spectrum Scale cluster, with Cloud Object Storage (COS) through the IBM Spectrum Scale Active File Management (AFM)-to-COS feature. This function enables copies of files or objects in an ESS 3200 or IBM Spectrum Scale cluster to be written to, or retrieved from, an external COS. The same functionality can also be used to read or write data to or from other external NFS data sources.

– This integration provides ESS 3200 with the ability to seamlessly integrate and accelerate IBM Spectrum Scale data access as follows:

  - To and from external NFS storage
  - To object storage such as Amazon S3 and IBM Cloud® Object Storage

  ESS 3200 integrates external NFS storage and object storage into a common data repository with enterprise-high scalability, data availability, security, and performance. The AFM-to-cloud object storage associates an ESS 3200 / IBM Spectrum Scale fileset with a COS bucket. Figure 4-4 shows an example of this configuration.

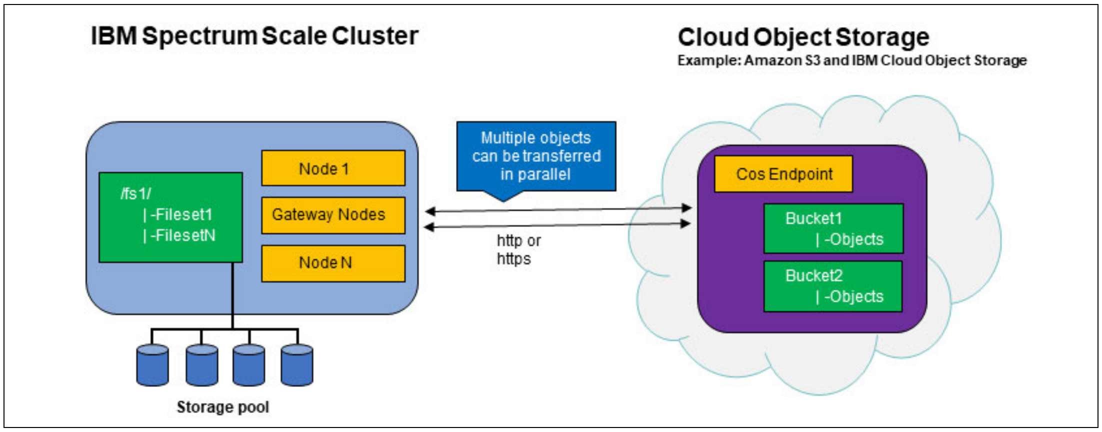


Figure 4-4 IBM Spectrum Scale Active File Management connected to Cloud Object Storage

Using this function, IBM Spectrum Scale filesets and COS buckets become extensions of each other. Files and objects required for applications such as AI and big data analytics can be shared, downloaded, worked upon, and uploaded between ESS 3200, IBM Spectrum Scale, and the COS. These use cases are shown in Figure 4-5.

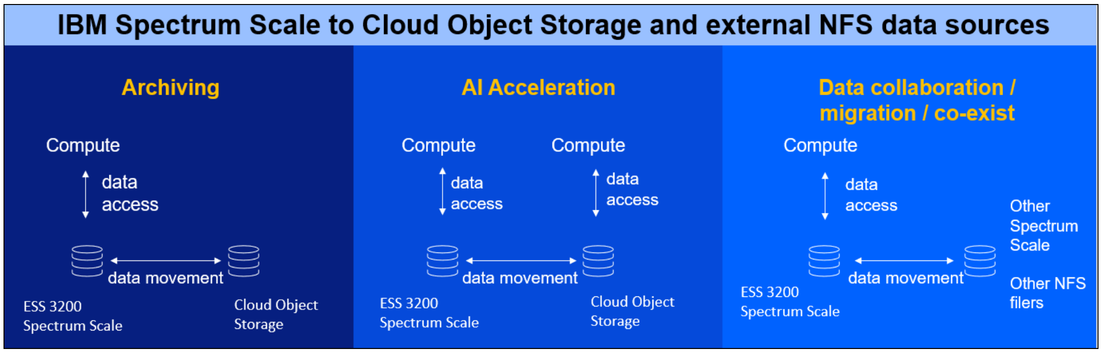


Figure 4-5 IBM Spectrum Scale to cloud object storage and external NFS data sources

The use cases include (but are not limited to):

- Archive data to and from external object storage data sources.
- Optimize data movement and data access, speeding time to data value.

- The same functionality can also connect and migrate data to consolidate, or co-exist with, external network file system (NFS) network-attached storage (NAS) data sources.

The workloads and workflows that might benefit from these use cases include (but are not limited to) mobile applications, backup and restore, enterprise applications, big data analytics, and file servers.

The AFM-to-COS feature also allows data center administrators to free ESS 3200 and IBM Spectrum Scale storage capacity by policy management. Data is moved to lower-cost or off-premise cloud storage, which reduces capital and operational expenditures. The data movement can be done automatically through the AFM-based cache eviction feature or through policy. The data movement can be used to automate and optimize data placement between ESS 3200 and other storage within the IBM Spectrum Scale storage ecosystem.

A

Convert a previous generation switch into a version 2

This appendix describes the management switch version 2 configuration and how to convert a previous generation switch into a version 2.

This appendix also provides you with the data points to consider if you are not using the IBM-provided and supported switches for the management network. It describes what to do in your network switch to achieve the same functionality with your own network devices.

As shown on 3.1.2, “Hardware planning” on page 44, the management switch includes ports 1 to 12 as “ESS 3200" ports. Those ports are different from version 1 because both management FSP networks are configured in the same port.

The process to platform the switch has not changed from version 1. The configuration-content of the file is used to platform the switch. The same two VLANs that were used on version 1 are used in version 2. New VLANs are not added from version 1.

Follow these steps to configure a version 2 switch:

1. Connect through SERIAL. You will lose access to the switch if you apply it through IP access because the default configuration does not have any IP configured.

   You must know the cumulus user password combo. It likely is `cumulus/CumulusLinux!`, but it might be something like the serial number. The serial configuration is:

   – Baud rate: 115200
   – Parity: None
   – Stop bits: 1
   – Data bits: 8
   – Flow control: None

   You need the configuration file that is detailed at the end of this appendix in Example A-1 on page 64.

2. Login to the switch as cumulus user and `sudo su -.`

3. As `root`, put the contents of the configuration file into `/etc/network/interfaces`. (You can copy and paste.) Previous contents of the configuration file must be discarded.

LAST WARNING: You must be connected through SERIAL or you will lose access in step 4.

4. Apply the configuration with `ifreload -a`.

   At this point the new configuration is applied.

   You can check that the configuration is applied with `ifquery -a`.

5. RECOMMENDED: Set a static IP to log remotely on the switch (Example 192.168.44.0/24 network IP switch 192.168.44.20, gateway 192.168.44.1)

   – net add interface eth0 IP address 192.168.44.20/24
   – net add interface eth0 IP gateway 192.168.44.1
   – net pending
   – net commit

Note: If you are converting a switch that has already non ESS 3200 using the switch on any port from 1 to 12, you need to evacuate one by one those ports. If you are not using ports in the range 1-12 you just need to apply the process above.

That means to move the cables on the upper ports from 1 to 12 to any free upper port that is not in the range ports 1-12. Equally any lower cable plugged to any port in the range 1-12 needs to be moved to any lower port not in the range of ports 1-12.

You should do the move one cable at the time and wait until the link LED on the destination port comes up. Once all ports in the range 1-12 are no longer cabled, you can apply the procedure explained here.

The file with the configuration must contain the data shown in Example A-1.

Example A-1 Data required for the configuration file

```
# This file describes the network interfaces available on your system
# and how to activate them. For more information, see interfaces(5).
source /etc/network/interfaces.d/*.intf
# The loopback network interface
auto lo
iface lo inet loopback
# The primary network interface
auto eth0
iface eth0 inet dhcp
# EVEN Ports/Lower ports PVID 101 for FSP network
auto swp14
iface swp14
bridge-access 101
auto swp16
iface swp16
bridge-access 101
auto swp18
iface swp18
bridge-access 101
auto swp20
iface swp20
bridge-access 101
auto swp22
iface swp22
bridge-access 101
auto swp24
```

```
iface swp24
bridge-access 101
auto swp26
iface swp26
bridge-access 101
auto swp28
iface swp28
bridge-access 101
auto swp30
iface swp30
bridge-access 101
auto swp32
iface swp32
bridge-access 101
auto swp34
iface swp34
bridge-access 101
auto swp36
iface swp36
bridge-access 101
auto swp38
iface swp38
bridge-access 101
auto swp40
iface swp40
bridge-access 101
auto swp42
iface swp42
bridge-access 101
auto swp44
iface swp44
bridge-access 101
auto swp46
iface swp46
bridge-access 101
auto swp48
iface swp48
bridge-access 101


# ODD Ports/Upper ports PVID 102 for xCAT network
auto swp13
iface swp13
bridge-access 102
auto swp15
iface swp15
bridge-access 102
auto swp17
iface swp17
bridge-access 102
auto swp19
iface swp19
bridge-access 102
auto swp21
iface swp21
```

```
bridge-access 102
auto swp23
iface swp23
bridge-access 102
auto swp25
iface swp25
bridge-access 102
auto swp27
iface swp27
bridge-access 102
auto swp29
iface swp29
bridge-access 102
auto swp31
iface swp31
bridge-access 102
auto swp33
iface swp33
bridge-access 102
auto swp35
iface swp35
bridge-access 102
auto swp37
iface swp37
bridge-access 102
auto swp39
iface swp39
bridge-access 102
auto swp41
iface swp41
bridge-access 102
auto swp43
iface swp43
bridge-access 102
auto swp45
iface swp45
bridge-access 102
auto swp47
iface swp47
bridge-access 102


# ESS 3200 ports (1 to 12) FSP + OS on single physical port
auto swp1
iface swp1
bridge-pvid 102
bridge-vids 101
auto swp2
iface swp2
bridge-pvid 102
bridge-vids 101
auto swp3
iface swp3
bridge-pvid 102
bridge-vids 101
```

```
auto swp4
iface swp4
bridge-pvid 102
bridge-vids 101
auto swp5
iface swp5
bridge-pvid 102
bridge-vids 101
auto swp6
iface swp6
bridge-pvid 102
bridge-vids 101
auto swp7
iface swp7
bridge-pvid 102
bridge-vids 101
auto swp8
iface swp8
bridge-pvid 102
bridge-vids 101
auto swp9
iface swp9
bridge-pvid 102
bridge-vids 101
auto swp10
iface swp10
bridge-pvid 102
bridge-vids 101
auto swp11
iface swp11
bridge-pvid 102
bridge-vids 101
auto swp12
iface swp12
bridge-pvid 102
bridge-vids 101


# Bridge setup
auto bridge
iface bridge
bridge-vlan-aware yes
bridge-ports glob swp1-48
bridge-pvid 101
bridge-pvid 102
bridge-stp
off
```

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Monitoring and Managing the IBM Elastic Storage Server Using the GUI*, REDP-5471
- *Introduction Guide to the IBM Elastic Storage Server*, REDP-5253
- *Implementation Guide for IBM Elastic Storage System 5000*, SG24-8498
- *Implementation Guide for IBM Elastic Storage System 3000*, SG24-8443
- *Highly Efficient Data Access with RoCE on IBM Elastic Storage Systems and IBM Spectrum Scale,* REDP-5658

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM Documentation - IBM Elastic Storage System 3200:

  https://www.ibm.com/docs/en/ess/6.1.1_cd
- IBM Spectrum Scale V 5.1.1 Planning:

  https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=planning
- Licensing on IBM Spectrum Scale

  https://www.ibm.com/docs/en/spectrum-scale/5.1.1?topic=overview-capacity-based-licensing
- Using IBM Cloud Object Storage with IBM Spectrum Scale:

  https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WUS12361USEN
- mmvdisk Command Reference:

  https://www.ibm.com/docs/en/spectrum-scale-ece/5.1.1?topic=commands-mmvdisk-command

Help from IBM

- IBM Support and Downloads

  `ibm.com/support`

- IBM Global Services

  `ibm.com/support`

Redbooks

Implementation Guide for
IBM Elastic Storage System

SG24-8516-00

ISBN 0738460176

IBM

(1.5” spine)
1.5”<-> 1.998”
789 <->1051 pages

Redbooks

Implementation Guide for
IBM Elastic Storage System 3200

SG24-8516-00

ISBN 0738460176

IBM

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

Redbooks

Implementation Guide for IBM Elastic Storage System 3200

SG24-8516-00

ISBN 0738460176

IBM

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

Redbooks

Implementation Guide for IBM Elastic Storage System 3200

(0.2”spine)
0.17”<->0.473”
90<->249 pages

(0.1”spine)
0.1”<->0.169”
53<->89 pages

Redbooks

Implementation Guide for IBM Elastic Storage System

SG24-8516-00

ISBN 0738460176

IBM

(2.5” spine)
2.5”<->nnn.n”
1315<-> nnnn pages

Redbooks

Implementation Guide for IBM Elastic Storage System 3200

SG24-8516-00

ISBN 0738460176

IBM

(2.0” spine)
2.0” <-> 2.498”
1052 <-> 1314 pages

SG24-8516-00

ISBN 0738460176

Printed in U.S.A.